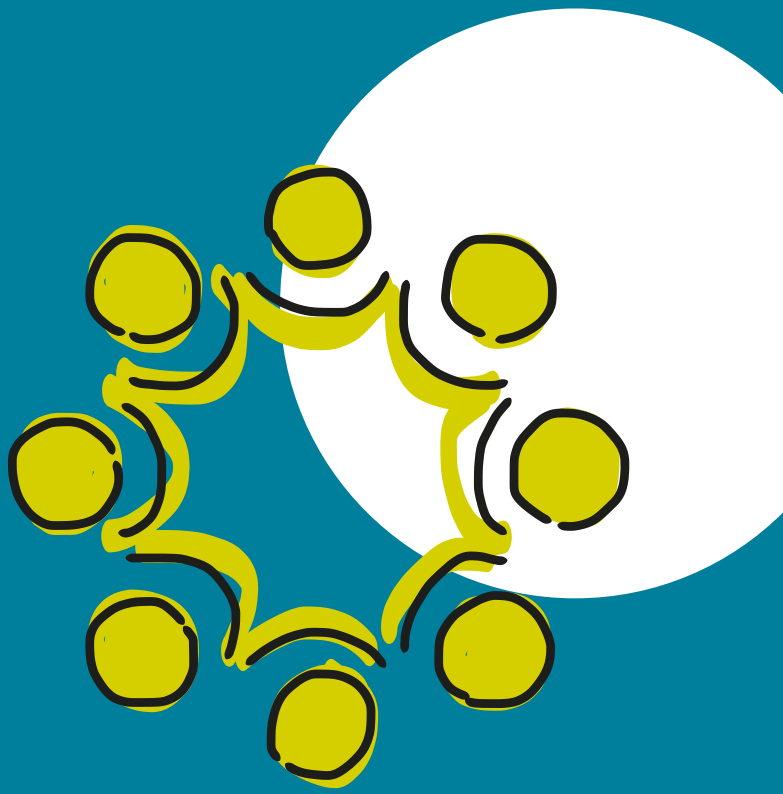
Editors
Andreas Wagner
Ursula Gmelch
Michael Granitzer
Christian Guetl
Megi Sharikadze
Stefan Voigt

Proceedings

8th International Open Search Symposium

#ossym2026

7-9 October 2026
CODE University of Applied Sciences
Berlin, Germany



open search foundation

CODE UNIVERSITY OF APPLIED SCIENCES

ISSN: 2957-4935
ISBN: 978-92-9083-744-2

Impressum

Editors:

Andreas Wagner, CERN, Geneva, Switzerland
Ursula Gmelch, Open Search Foundation, Germany
Michael Granitzer, University Passau, Germany
Christian Gütl , Graz University of Technology, Austria
Megi Sharikadze, Leibniz Supercomputing Centre, Munich, Germany
Stefan Voigt, Open Search Foundation, Germany

ISSN: 2957-4935
ISBN: 978-92-9083-744-2
DOI: 10.5281/zenodo.23171728

Copyright © CERN, 2026

This work is published under the Creative Commons Attribution-NoDerivatives International License (CC BY-ND 4.0)

The terms are defined at https://creativecommons.org/licenses/by-nd/4.0/

This report should be cited as:

Proceedings of 8h International Open Search Symposium #ossym2026, CODE University of Applied Sciences, Berlin, Germany and Online, 7-9 October 2026, A. Wagner, U. Gmelch, M. Granitzer, C. Gütl, S. Voigt (eds). http://doi.org/10.5281/zenodo.23171728

More Information

- Conference Website on CERN Indico
  https://indico.cern.ch/e/OSSYM-2026
- Open Search Community at Zenodo
  https://zenodo.org/communities/opensearch
- Event information at the Open Search Foundation
  https://opensearchfoundation.org/events-osf/ossym2026

Foreword

Dear readers,

Already in the 8th year of our #ossym conference series, it is our great pleasure to present the proceedings of the 8th International Open Search Symposium - #ossym2026 - which takes place from 7 to 9 October 2026 in Berlin, Germany, hosted by CODE University of Applied Sciences.

In this year's symposium there are 28 accepted papers from 84 authors. A growing community around the topic and the constant strong interest in open search and artificial intelligence is reflected in many inspiring contributions in the topic clusters "Crawling", "User behaviour and preferences in search and AI", "Legal aspects of Open Web Search and AI training", "Ethics and Society", and "AI in Science".

All in all, the #ossym symposium offers a variety of formats to exchange and to collect 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 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:

- "Sovereign AI Services Leading to a European Future of Search" Julian Kunkel, Deputy Head of High-Performance Computing at the Göttingen Supercomputing Center (GWDG) and Professor for High-Performance Computing at University of Göttingen
- "The Sovereign AI Stack: A Builder's Perspective" Stefan Andresen, Chief Scientist, DeepL
- "The Pulse of the Planet: Remote Sensing and geoweb data redrawing the map of human behaviour" Hannes Taubenböck, Head of department „Georisks and Civil Security" at the Earth Observation Center (EOC) of the German Aerospace Center (DLR) and Professor at the University Würzburg holding the chair "Global Urbanization and Remote Sensing"
- "Sovereignty for whom? The politics of breaking with dependencies in search" Aline Blankertz, Structural Integrity

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, CODE University of Applied Sciences, for all the organizational efforts. Without all these great contributions and helpful support, it would not be possible to successfully run the 8th 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. Looking back at our previous events, we are now even more convinced that #ossym provides this place of exchange in a very fruitful and interesting manner. 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 announce and look forward to next year's gathering – #ossym2027 - to be hosted by the University of Applied Sciences for Business & Society BFI Vienna, Austria.

On behalf of the #ossym2026 conference chairs,

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

Symposium Organisation

Programme Committee

Prof. Emmanuel Cartier | European Commission JRC
Prof. Dr. Alexander Decker | Technische Hochschule Ingolstadt, Germany
Prof. Dr. Arjen P. de Vries | Radboud University, Netherlands
Prof. Dr. Stefan Dietze | Heinrich Heine University Düsseldorf and GESIS, Germany
Msc. Roxanne ElBaff | German Aerospace Centre (DLR), Germany
Prof. Dr. Kai Erenli | University of Applied Sciences BFI Vienna, Austria
Dipl. Inf. Fabio Fracassi | Code.Berlin University of Applied Sciences, Berlin, Germany
Diana Francisco de Sousa | Joint Research Centre (JRC), Ispra, Italy
Dr. Maik Fröbe | Friedrich-Schiller-Universität, Jena, Germany
Prof. Dr. Fabian Geier | Code.Berlin University of Applied Sciences, Berlin, Germany
Priv.-Doz Dr. Christian Geminn | University Kassel, Germany
Prof. Dr. Michael Granitzer | University Passau, Germany
Prof. Dr. Christian Guetl | Graz University of Technology, Austria
Prof. Dr. Matthias Hagen | Friedrich-Schiller-Universität, Jena, Germany
Dr. Tobias Hecking | German Aerospace Centre (DLR), Germany
Prof. Dr. Denis Helic | Graz University of Technology, Austria
Msc. Gijs Hendriksen | Radboud University, Netherlands
Prof. Dr. Andreas Henrich | University Bamberg, Germany
Prof. Djoerd Hiemstra | University of Twente, Radboud University, Netherlands
Msc. Phil Höfer | SUMA e.V., Germany
Prof. Dr. Robert Jäschke | Humboldt University Berlin & L3S, Hannover, Germany
Dr. Igor Jakovljevic | CERN, Geneva, Switzerland
Prof. Dr. Mohammed Kaicer | Faculty of Sciences Kenitra, Morocco
Dr. Jens Kersten | German Aerospace Centre, Jena, Germany
Dr. Jelena Mitrovic | University of Passau, Germany
Prof. Dr. Engelbert Niehaus | University Koblenz-Landau, Landau, Germany
Prof. Dr. Monica Palmirani | Università di Bologna, Italy
Prof. Dr. Melanie Platz | Saarland University, Germany
Prof. Dr. Martin Potthast | Kassel University, Germany
Prof. Dr. Mirko Presser | Aarhus University, Denmark
Prof. Dr. Georg Rehm | German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
Dr. Renée Ridgway | Aarhus University, Denmark
Prof. Dr. Adam Roe | Code.Berlin University of Applied Sciences, Berlin, Germany
Dr. Tim Smith | CERN, Geneva, Switzerland
Prof. Dr. Benno Stein | Bauhaus University, Weimar, Germany
Dr. Nicola Tonellotto | University of Pisa, Italy
Eng. Marco Verile | European Commission, Joint Research Centre (JRC), Ispra, Italy
Dr. Stefan Voigt | Open Search Foundation, Germany
Dr. Andreas Wagner | CERN, Geneva, Switzerland

Conference Chairs

Prof. Dr. Michael Granitzer | University Passau, Germany
Prof. Dr. Christian Guetl | Graz University of Technology, Austria
Dr. Megi Sharikadze | Leibniz Supercomputing Centre, Munich, Germany
Dr. Stefan Voigt | Open Search Foundation, Starnberg, Germany
Dr. Andreas Wagner | CERN, Geneva, Switzerland

Publication Chairs

Sebastian Guertl | Graz University of Technology, Austria
Elias Sandner | CERN, Geneva, Switzerland
Dr. Andreas Wagner | CERN, Geneva, Switzerland

Contents

Preface i
Impressum i
More Information i
Foreword ii
Symposium Organisation iii
Research Track Information v
Papers 1
AIS-P01 - AI-Assisted Search String Creation Module for NeutrinoReview 1
AIS-P02 - Evaluating the Impact of Eligibility Criteria Reformulation on LLM-Based Literature Screening 9
CRA-P01 - Estimating Absolute Web Crawl Coverage From Longitudinal Set Intersections 17
CRA-P02 - A Focused Crawling Strategy to Enhance Geoscience Data Retrieval 23
ETS-P01 - An Ethics-by-Design Approach for the Open Web Index 28
ETS-P02 - Co-Creating Inclusive AI Search Frontends: Findings from the KIsu-Project with children, young people, and older adults 34
ETS-P03 - Open, to What End? A Capability-Theoretic Perspective on Open Search 39
LEA-P01 - Quo Vadis Digital Regulation? An Overview of the Digital Regulatory Landscape in Europe and what it means for the European Web Data Infrastructure 45
LEA-P02 - Data protection responsibility of online platforms for user-generated content – recognising sensitive data, preventing scraping and identifying users in light of the ECJ's 'Russmedia' judgement 53
RET-P01 - OURS & YOARS: Open Platforms for Sovereign Web Retrieval and Agentic Information Access 60
RET-P02 - An Open-Source Retrieval Layer for Institutional RAG: From Authenticated Crawl to Searchable Index 66
SPM-P01 - Proposed Methodology for Open Search Infrastructure leveraging federated RAG-systems and open-source Speech Intelligence for cross-domain collaboration featuring Disaster Management, Traffic Flow Prediction and Literature Studies 70
UBP-P01 - Hybrid Search Ecosystems: An Empirical Analysis of Online Search Behavior of Generation Z and Alpha in Germany 76
UBP-P01 10WDA-P01 - Designing a Power Bi Dashboard for Data-Driven Supplier Comparison in Manufacturing Procurement 85
Extended Abstracts 92
AIS-A01 - TraceableAI: An Open-Source Agentic Framework For Prisma-Compliant Literature Synthesis 92
CRA-A01 - Measuring Crawler Lock-in: Translating Scheduling Preferences into Language Coverage for Common Crawl 95
LEA-A01 - Design Conflicts in Open Search Infrastructures: A Design Pattern-Based Approach to Integrating Legal and User Requirements 97
LEA-A02 - Public Digital Infrastructure As Regulatory Stabilizer In A Recalibrated AI Landscape: Strategic Legal Navigation Under The EU Digital Omnibus On AI 99
RET-A01 - Adaptive Unsupervised Topic Modeling for Domain Focused Web Search 102
SPM-A01 - Extraction of Geospatial Disaster Related Information from Web News and Text Data 104
SPM-A02 - A Spatially Enabled Search System For Hydrological Publications: A Demonstration 106
UBP-A01 - Privacy-Enhancing Digital Infrastructures – Understanding User Preferences Regarding Internet Search and Their Inclusion into Systems Design 108
UBP-A02 - Renegotiating the Open Web: Emerging AI Preferences Standards 110
WDA-A01 - eMarketShield: Enhancing Consumer Protection in E-Commerce through Digital Innovation and Regulatory Oversight 112
WDA-A02 - SOURCE: Scalable, Open, and Comprehensive Detection of Disinformation Campaigns on the Web 113
Appendix 114
List of Authors 114

Research Track Information

AIS	-	AI in Science
CRA	-	Crawling
ETS	-	Ethics and Society
LEA	-	Legal Aspects of Open Web Search and AI training
RET	-	Retrieval
SPM	-	Spatial Modelling
UBP	-	User Behaviour and Preferences in Search and AI
WDA	-	Web Data Driven Applications

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22921381

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

AI-ASSISTED SEARCH STRING CREATION MODULE FOR NEUTRINOREVIEW

E. Sandner*[1,3], S. Pack†[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

The formulation of search strategies represents a critical step in systematic reviews, as the quality of the constructed search string directly determines the completeness and reliability of the retrieved evidence base. Developing effective search queries for bibliographic databases such as PubMed requires substantial expertise and is often a time-consuming and iterative process involving keyword identification, synonym expansion, and Boolean query construction. This paper presents an AI-assisted search string creation module integrated into the NeutrinoReview platform, developed in the context of a CERN–WHO research collaboration on systematic review automation. The module enables researchers to transform eligibility criteria into structured PubMed search strings and iteratively refine them through AI-generated synonym suggestions, natural language query modification, seed paper validation, and version tracking. To evaluate the proposed approach, a task-based usability study was conducted comparing the NeutrinoReview module with the conventional PubMed search interface. Ten participants performed search strategy development tasks using both systems. Results indicate improved perceived usability for the proposed system, with an average System Usability Scale (SUS) score of 79.5 compared to 51.75 for the PubMed interface. Participants also completed the tasks approximately five minutes faster on average and reduced the average number of retrieved records from 1201 to 506 while successfully retrieving all provided seed papers used for evaluation. These exploratory findings suggest that interactive AI-assisted tools may improve perceived usability and task efficiency in search strategy development while maintaining retrieval of the provided seed papers, highlighting the need to further assess transparency, user control, and retrieval quality.

INTRODUCTION

Systematic reviews sit at the heart of evidence-based research. They offer a structured and reproducible way of synthesizing existing literature [1,5]. Their methodological consistency relies heavily on transparent documentation and a clearly defined search strategy. This approach enables researchers to identify relevant studies and assess the available literature. Among the phases of a systematic review, the data retrieval stage is arguably one of the most important ones, as it determines the set of potentially relevant studies that will subsequently be screened and analyzed. An inadequately formulated search strategy may result in incomplete evidence retrieval, biased conclusions, or excessive manual effort during screening [2]. Consequently, the quality of the search strategy directly influences the reliability of the review findings. The creation of search strings for bibliographic databases such as PubMed[1] is inherently iterative and requires substantial expertise. Researchers must identify relevant keywords, derive appropriate synonyms, consider database-specific indexing terms, and construct syntactically correct Boolean expressions [10]. This process is often time-consuming and cognitively demanding. Moreover, different researchers may produce substantially different search strings for the same research question, potentially leading to divergent result sets [9]. The identification and combination of appropriate keywords is therefore a critical step in the review workflow.

Despite recent advances in AI support for systematic reviews, search strategy development remains a largely manual and expertise-dependent process. Addressing these issues is part of the NeutrinoReview project[2], which aims to support and partially automate the systematic review workflow. Within this workflow, the data retrieval phase represents a critical component with substantial potential for improvement. Supporting researchers during this stage may therefore reduce the overall workload. The NeutrinoReview platform was developed as part of this effort to support the systematic review process [8]. While other phases of systematic reviews, such as screening and data extraction, have already been subject to automation efforts [11], the search string creation phase previously lacked structured and iterative support. Providing tool support for this stage therefore represents an important step toward a more integrated systematic review workflow. This paper focuses exclusively on the search strategy creation phase within the NeutrinoReview platform and does not address the automation of subsequent review phases such as deduplication or literature screening. The developed solution enables users to transform eligibility criteria into structured PubMed search strings and supports iterative refinement through synonym enrichment, natural language based modifications, seed paper validation, and version management. It is designed to support researchers

* elias.sandner@cern.ch
† sebastian.pack@student.tugraz.at

[1] https://pubmed.ncbi.nlm.nih.gov/
[2] https://neutrino-review.web.cern.ch/

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22921381

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

during the search strategy development process while preserving transparency and user control.

The main contribution of this paper is the design and evaluation of an AI-assisted search string creation module integrated into the NeutrinoReview platform. The proposed approach combines AI-assisted search string generation with iterative user-guided refinement and is evaluated in a comparative usability study against the conventional PubMed search interface. The objective of this work is to investigate whether the proposed AI-assisted workflow can improve usability and efficiency during search strategy development.

Therefore, this paper aims to answer the following research questions:

- RQ1: To what extent can an AI-assisted interactive workflow improve the efficiency of PubMed search string creation for systematic reviews while retrieving the provided seed papers?
- RQ2: How can the assistance of AI help in terms of perceived usability?

The remainder of this paper is structured in several chapters, covering background and related work in this field, as well as the used methodology and details about the conducted user study. Finally, within the evaluation chapter, results are presented and discussed, and limitations of this paper are highlighted.

BACKGROUND AND RELATED WORK

Systematic reviews are conducted to synthesize evidence based scientific research in a structured and transparent manner. A key strength of such reviews lies in their reproducibility and traceability. To ensure these properties and to enhance the reliability of the review process, clearly defined procedures must be followed throughout the entire workflow [1, 5].

The workflow typically consists of several phases. Within the NeutrinoReview project, the process is structured into project protocol development, data retrieval, title and abstract screening, full-text screening, and data extraction. During the protocol development phase, the research question and eligibility criteria are defined and the methodological approach of the review is specified. The data retrieval phase focuses on identifying potentially relevant studies through database searches. In the subsequent screening stages, retrieved publications are evaluated according to the predefined eligibility criteria. Finally, relevant studies are analyzed and the extracted information is summarized [2, 5]. Each phase contributes to ensuring that the review process remains systematic and reproducible.

The development of a search strategy is a central component of the data retrieval phase. In this step, researchers define search queries that are executed against databases such as PubMed. The aim of the search strategy is to retrieve as many relevant studies as possible while minimizing the number of irrelevant results. Achieving this balance between sensitivity and specificity is essential for ensuring the coverage of relevant literature while maintaining manageable screening workloads [10].

Search string generation itself is typically an iterative process. The process usually begins with the identification of key concepts derived from the research question and eligibility criteria. These concepts are often structured using frameworks such as the PICO model, which distinguishes between population, intervention, comparison, and outcome elements [3]. For each concept, relevant keywords and synonyms are identified and grouped together. These groups are then combined using Boolean operators to form a search query [10].

In practice, search strategies often require multiple iterations. Initial queries are executed and the resulting records are examined in order to identify additional keywords that may improve the query. The search string is then refined and executed again until the results are considered sufficiently comprehensive. This iterative refinement process contributes to the overall quality of the systematic review but also increases the time and effort required for conducting the search [9]. When this process is performed without dedicated tool support, several steps have to be carried out manually. Previous search string versions have to be documented outside the database interface, seed papers have to be checked manually against the retrieved result set, and retrieved records or result counts have to be inspected, copied, or downloaded for comparison. Although large language models (LLM) can support the refinement of search strings, for example by suggesting synonyms or reformulating query terms, using them outside the review platform requires users to copy research questions, eligibility criteria, search strings, and suggestions between separate user interfaces.

Another important concept in systematic search strategy development is the use of seed papers. Seed papers are known relevant publications that represent the research topic of interest. These papers can be used to validate the effectiveness of a search strategy by verifying whether the constructed query retrieves these publications from the database [12]. Seed papers can also be used to identify additional keywords, concepts, or indexing terms that may not have been considered during the initial search strategy development.

Recent advances in generative artificial intelligence and LLMs have further expanded the possibilities for supporting complex research workflows. LLMs have demonstrated the ability to generate and manipulate natural language text, summarize information, and assist with various research-related tasks [4, 6, 20]. These capabilities suggest potential applications in systematic review workflows, including supporting the formulation and refinement of search queries.

Several automated approaches to search string generation have already been proposed. One example is the Search String Generator (SeSG), which aims to reduce the effort required to formulate search queries for secondary studies. The SeSG approach combines text mining techniques, topic modeling, and keyword enrichment methods to generate search strings based on a set of input documents. The work-

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22921381

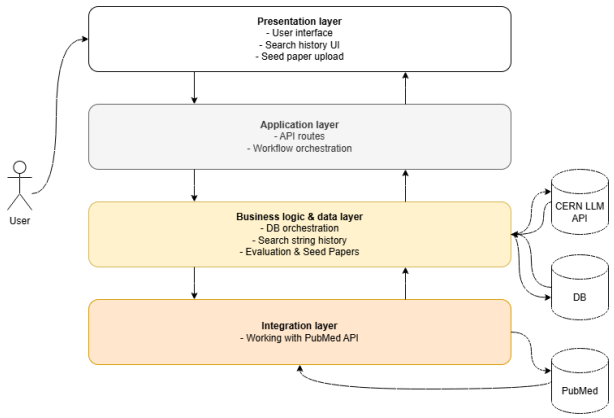


Figure 1: Conceptual architecture of the module

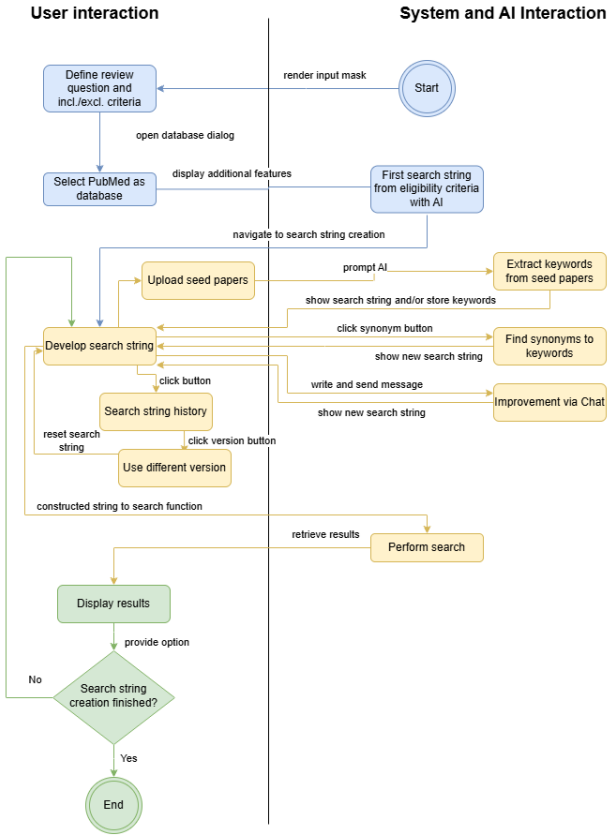


Figure 2: Search string creation process workflow

flow consists of preprocessing, topic extraction and enrichment, generation of the search string, and application of the generated query [7].

However, existing approaches primarily focus on fully automated search string generation and often provide limited possibilities for user interaction during the generation process. This can reduce transparency and control during search strategy development, which is an important requirement in systematic reviews. Consequently, interactive approaches that combine automated support with user-guided refinement represent a promising direction for improving the search string creation process. NeutrinoReview already provides functionality to retrieve publication data from PubMed through the PubMed API [21]. The contribution addressed in this paper therefore focuses on the structured and iterative creation, refinement, validation, and version tracking of PubMed search strings within this existing workflow.

METHODOLOGY

This section describes the methodological setup of the proposed approach. First, the design and integration of the search string creation module are presented, including the workflow, the interaction with PubMed, the use of seed papers, and the integration of the LLM. Subsequently, the user study is described, including the experimental setup, tasks, recorded metrics, and usability assessment.

Module Design and Integration

The implemented search string creation component extends the existing NeutrinoReview platform by enriching it with a structured workflow that supports users during the development and refinement of search strategies. Figure 1 illustrates the conceptual architecture of the module and Figure 2 summarizes the search string development process. The workflow is organized into three main phases: the input phase, the iterative search string creation phase, and the result phase. The input phase represents the starting point of the search string creation process. In this phase, users define the research question and the eligibility criteria of the systematic review. These criteria provide the conceptual basis for the subsequent search strategy development. Based on this input, the system automatically generates an initial search string draft with the help of a LLM that serves as a starting point for further refinement. As an example, Listing 1 provides the system prompt the LLM gets for the initial search string creation, whereas all other prompts are documented in the supplementary material[3]. Upon receiving an answer from the LLM, the process continues with the iterative search string creation phase. This phase represents the core functionality of the project. During this phase, users can iteratively refine the search strategy using several mechanisms. Keywords included in the search string can be extended with AI-generated synonyms in order to improve the coverage of the query. Additional keywords can also be added manually to further refine the search strategy. In addition to manual modifications, the system provides an improvement chat that allows users to modify the search string using natural language instructions. The chat interface enables users to request modifications such as adding synonyms, removing terms, or restructuring parts of the query. The system then generates an updated search string based on the provided instruction. Another important component of the workflow is the integration of seed papers. Seed papers represent known relevant publications related to the research question and can be used to validate the effectiveness of a search strategy [12]. Users can upload such papers to

[3] `https://zenodo.org/records/21640010`

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22921381

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

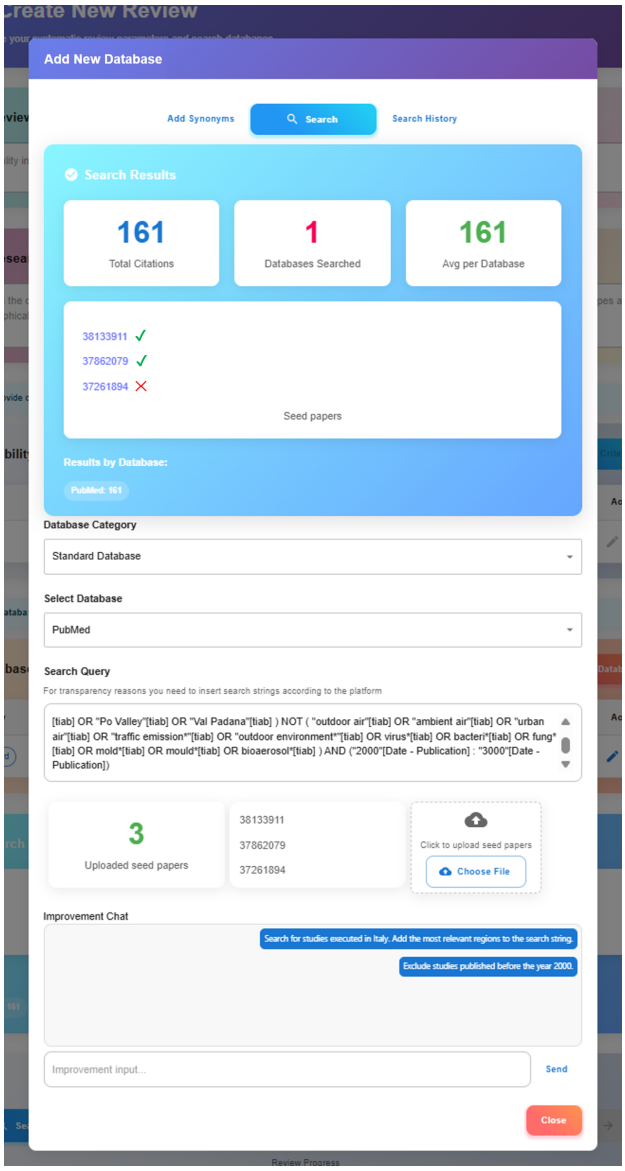


Figure 3: User interface for developing and refining search strategies

the system as a list of PMIDs, which then extracts relevant keywords and incorporates them into the search string creation process. Throughout the iterative process, the system maintains a search string history that stores all previously executed search queries together with their corresponding results and seed paper coverage. This history allows users to compare different query versions and trace the evolution of the search strategy. The final phase of the workflow is the result phase. In this phase, the constructed search query is executed against the PubMed database [13]. The system reports the number of retrieved records and indicates whether the uploaded seed papers appear within the result set. Based on this feedback, users can decide whether the search strategy should be refined further or finalized. The corresponding user interface is shown in Figure 3.

Listing 1: Prompt used for Initial Search String Creation based on Eligibility Criteria

```
prompt = """

--- Task:
Your task is to read keywords from so called
    eligibility criteria, consisting of
    inclusion criteria, exclusion criteria
    and labels that define how to group
    keywords.

--- Input:
The input is a list of objects in
    json-format, each containing an
    inclusion criteria, an exclusion
    criteria and a label.

--- Extraction:
From inclusion and exclusion criteria,
    extract only the most relevant keywords.
Inclusion and exclusion criteria can also
    contain several keywords or
    descriptions, usually separated by
    something similar to "," or ";".

--- Output:
The expected output is a syntactically
    correct PubMed search string.
Do not provide a programmatic approach,
    extract keywords and formulate the
    search string yourself!
Do not add any explanation! Do only return
    the created search string! Do not add
    anything to the valid and final search
    string!
Stick to the provided examples!
There are two approaches shown in the
    example. Find an approach that is
    inbetween regarding number of synonyms
    and the summarization of multiple
    structures/labels one group.

--- General rules:
Place "" around word groups!
Use the labels as a help to categorize the
    extracted keywords accordingly.
If exclusions or inclusions are empty, do
    not categorize!
Remove empty brackets!
Remove empty strings!
Do it like in the examples!

--- Examples:
Example 1 (short output): Input:
    {'inclusion': "interaction with
        chatbots, conversational agents",
        'exclusion': "", 'label':
        "population"},
    {'inclusion': "searching behaviour,
        search included", 'exclusion': "",
        'label': "intervention"},
    {'inclusion': "information about user
        satisfaction", 'exclusion': "",
        'label': "outcome"}
Output: (chatbot OR "conversational agent")
    AND search AND "user satisfaction"

Example 2 (longer output): Input:
    {'inclusion': "General-purpose LLMs
        covered in publicly available
```

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22921381

```
        literature", 'exclusion': "",
        'label': "population"},
    {'inclusion': "General Purpose LLMs
        applied for literature screening",
        'exclusion': "", 'label':
        "intervention"},
    {'inclusion': "Double blinded human
        screening", 'exclusion': "",
        'label': "Comparison"}
    {'inclusion': "Sensitivity, Workload
        reduction, Characteristics of
        evaluation datasets, papers since
        2021", 'exclusion': "", 'label':
        "outcome"},
Output:
("generative ai"[tiab] OR "generative
    artificial intelligence"[tiab] OR "large
    language model"[tiab] OR "large language
    models"[tiab] OR llm[tiab] OR gpt[tiab]
    OR chatgpt[tiab] OR chatbot[tiab] OR
    "general purpose llm"[tiab] OR "general
    language model"[tiab] OR openai[tiab] OR
    claude[tiab] OR llama[tiab] OR
    gemini[tiab] OR bard[tiab] )
AND ( screening[tiab] OR "literature
    screening"[tiab] OR "study
    selection"[tiab] OR "citation
    screening"[tiab] OR "title abstract
    screening"[tiab] OR "abstract
    screening"[tiab] OR "evidence
    synthesis"[tiab] OR "systematic
    review"[tiab] OR "meta analysis"[tiab] )
AND ("2021"[dp] : "3000"[dp])

--- Examples for wrong output:
Wrong Example 1: Input:
    {'inclusion': "I like fish",
        'exclusion': "", 'label': "food"}
Output: (fish) NOT ()
"""
```

The search string creation component was integrated into the existing NeutrinoReview system. The interface was implemented using the React framework, which supports the development of dynamic user interfaces for web applications [16]. The backend services were implemented using the FastAPI framework [14]. These services handle tasks such as generating initial search strings, processing natural language instructions, communicating with external APIs, and storing search-related data. Search string versions, search results, and seed papers are stored in a PostgreSQL database [17]. This design enables version tracking and supports the reproducibility of the search strategy development process. The system interacts with the PubMed API in order to execute search queries and retrieve publication metadata [13]. In addition, the system communicates with a LLM hosted within the CERN infrastructure. The final model that was used was qwen2.5-32B [15]. It is used to generate search string drafts, produce keyword synonyms, and interpret natural language instructions for modifying search queries. LLMs have demonstrated strong capabilities in generating and manipulating natural language text, which makes them suitable for supporting such tasks [4, 20].

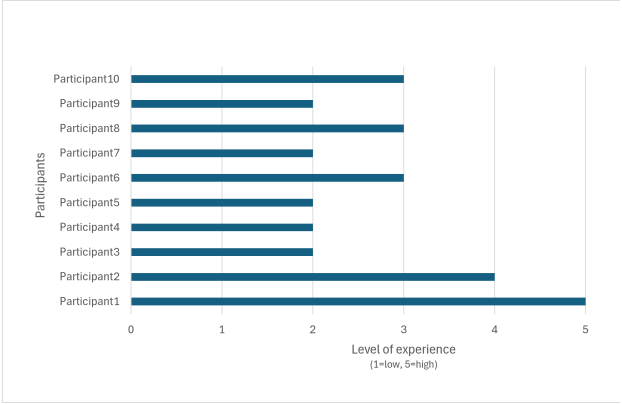


Figure 4: Self-evaluated level of experience of participants

User Study

To evaluate the usability and effectiveness of the developed search string creation module, a task-based usability study with ten participants was conducted. Participants comprised bachelor's, master's, and PhD students with varying levels of experience in systematic reviews. All participants had prior experience conducting literature searches and were familiar with the systematic review process, although their self-assessed level of experience with systematic reviews varied, as can be seen in figure 4. All data collected throughout the user study were anonymized. The study compared the AI-assisted search string creation workflow implemented in NeutrinoReview with the conventional PubMed search interface, which was selected as a baseline because it represents a common environment for manually constructing and testing PubMed queries. However, PubMed is a general-purpose biomedical search interface rather than a dedicated search-strategy assistance tool, and the comparison should therefore be interpreted as a comparison with common manual practice rather than with a specialized search-string generation system. The study followed an experimental design in which participants performed search string creation tasks using both systems. Such experimental approaches are commonly used to compare user performance across different system conditions [19]. Each participant therefore interacted with both systems in order to allow a direct comparison of the resulting search strategies and user experience. Participants were provided with predefined research questions and corresponding eligibility criteria. In addition, a set of seed papers was provided for each task. These seed papers represented known relevant publications that the search strategy should ideally retrieve. Participants were instructed to construct search strings that retrieved the provided seed papers while maintaining a reasonable number of overall search results. This reflects a common objective in systematic search strategy development, where researchers aim to balance comprehensive retrieval with manageable screening workloads [10]. Each participant completed two tasks. In one task, the search string was created using the NeutrinoReview system. In the other task, the search string was constructed using the standard PubMed interface. The exact task descriptions as

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22921381

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

well as the given seed papers are provided as supplementary material[4]. The order of the tasks was controlled to reduce potential learning effects. During the study, several quantitative metrics were recorded. These included the time required to complete the task and the total number of search results retrieved by the constructed query. These metrics provide insights into the efficiency and specificity of the generated search strategies. A task is considered completed once participants managed to retrieve all seed papers while keeping the number of retrieved records manageable.

In addition to these performance metrics, perceived usability was evaluated using the System Usability Scale (SUS) [18]. The SUS questionnaire consists of ten statements that participants evaluate on a Likert scale. The resulting score provides a standardized measure of perceived usability. Participants completed the SUS questionnaire after finishing each task. This allowed a direct comparison of perceived usability between the NeutrinoReview system and the PubMed interface. Furthermore, after completing both tasks, qualitative feedback was collected through additional questions. Participants were asked to describe their experience using both systems and to identify aspects they considered particularly helpful or problematic. This feedback provides additional insights into the strengths and limitations of the developed system. The questions raised to gather qualitative feedback are listed in the following:

- Compare the convenience of both workflows – where do you see benefits and drawbacks in the used search-string-creation automation feature?
- Do you trust the used AI system? If yes, what is convincing you, if not, what are your concerns?
- Which of the two workflows was more intuitive to use and why?
- Which further improvements would be beneficial?

Overall, the study design aims to evaluate both objective performance metrics and subjective user perceptions in order to provide a comprehensive assessment of the proposed search string creation approach.

EVALUATION

This section presents and discusses the results of the conducted usability study involving ten participants, focusing on task efficiency, perceived usability, and qualitative feedback.

Results

The SUS scores indicate a clear difference between the two evaluated systems. All achieved scores among all participants are displayed in figure 5. The SUS scores for NeutrinoReview ranged from 62.5 to 92.5, resulting in an average score of 79.5. In comparison, the PubMed interface achieved scores between 15 and 72.5, with an average value of 51.75. According to commonly used interpretations of the SUS scale, scores above 68 are considered above average usability [18]. Based on this threshold, the NeutrinoReview system can be considered to provide comparatively high usability, while the PubMed interface in this study falls below the commonly accepted usability benchmark.

In addition to usability scores, several performance-related metrics were collected during the study. One of these metrics was the time required to complete each search string creation task. The results show that tasks performed using the NeutrinoReview system were generally completed in less time than those performed using the PubMed interface. On average, NeutrinoReview reduced task completion time by nearly five minutes compared with PubMed. In some cases, participants required substantially more time when working with the standard PubMed interface, particularly when manually identifying synonyms and constructing Boolean search expressions. Another important metric was the number of search results returned by the constructed queries. Queries created using the PubMed interface typically returned substantially larger result sets. In the conducted tasks, the PubMed queries produced result sets ranging from 381 to 1916 records, with an average of retrieved results of 1201. In contrast, queries generated using the NeutrinoReview system returned considerably smaller result sets ranging from 43 to 1937 records whilst only averaging 506 retrieved results. The comparatively smaller result sets produced by NeutrinoReview suggest that the system supported the construction of more restrictive search queries in the evaluated tasks. At the same time, the seed papers provided during the tasks were successfully retrieved by the constructed search strings, indicating that the queries captured the predefined known relevant papers used in the evaluation.

Qualitative feedback collected from participants further supports these observations. Participants reported that the structured workflow of the NeutrinoReview system simplified the search string creation process. In particular, the ability to automatically generate synonyms and refine search strings using natural language instructions was perceived as helpful during the development of search queries. However, some participants also noted limitations of the system. In particular, they reported that it was sometimes unclear how specific keywords were introduced into the search string. This reduced transparency regarding the origin of individual query components.

Discussion

The evaluation results indicate that the proposed search string creation module has the potential to improve both usability and efficiency compared to the conventional PubMed interface. The higher SUS scores suggest that participants perceived the structured workflow and AI-assisted functionality of the system as beneficial during the search string creation process. One possible explanation for the improved usability is the reduction of cognitive effort required for constructing search queries. In traditional search interfaces, users must manually identify synonyms, construct Boolean expressions, and repeatedly modify queries. The Neutri-

[4] https://zenodo.org/records/21640010

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22921381

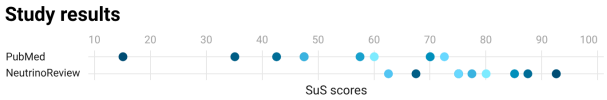


Figure 5: Results of the study

noReview system partially automates these steps by generating initial search string drafts and providing mechanisms for iterative refinement. The integration of AI-supported synonym generation appears to play an important role in this process. Identifying appropriate synonyms is a common challenge during search strategy development [10]. By automatically proposing additional keywords, the system assists users in expanding search queries without requiring extensive manual effort. This support may be particularly useful for researchers who are less experienced in constructing complex database queries, as it provides an initial structure that can subsequently be reviewed and refined by the user.

Another relevant observation concerns the size of the generated result sets. Queries created with NeutrinoReview returned substantially fewer results than those generated manually using the PubMed interface. While smaller result sets may initially appear to indicate reduced recall, the successful retrieval of the seed papers indicates that the system was still able to capture relevant publications in the evaluated tasks. The smaller result sets indicate that the constructed queries were more restrictive, although retrieval quality beyond the provided seed papers was not assessed. Reducing the number of irrelevant results can be beneficial during the screening stages of systematic reviews, as large result sets often require extensive manual screening and represent a major source of workload [9]. More targeted search strategies may therefore contribute to reducing the overall effort required to complete a review.

At the same time, the findings show that transparency remains essential when integrating AI-assisted functionality into systematic review workflows. Since search strategies must be reproducible and justifiable, users need to understand the origin and rationale of suggested terms, including whether they were derived from eligibility criteria, seed papers, or AI-generated suggestions.

LIMITATION AND FUTURE WORK

Several limitations should be considered when interpreting the results of the conducted evaluation. The most important limitation concerns the relatively small number of participants involved in the usability study. With only ten participants, the results cannot be considered statistically representative. Instead, the findings should be interpreted as exploratory insights into the usability and potential benefits of the developed system. Another limitation relates to the background of the study participants. All participants were affiliated with computer science and therefore had similar academic backgrounds. Researchers from other disciplines may have different levels of experience with systematic literature searches and may approach search strategy development differently. Consequently, additional studies involving participants from a broader range of research domains would be necessary in order to evaluate the system under more diverse conditions. Furthermore, the evaluation was conducted using predefined tasks and predefined seed papers. While this approach allowed for a controlled comparison with common manual PubMed use, PubMed is not designed as a dedicated search-strategy assistance tool. The results should therefore not be interpreted as showing superiority over specialized search-string generation systems, and real-world systematic reviews often involve more complex research questions and evolving search strategies.

Future studies could therefore investigate the system in real systematic review projects conducted over longer time periods. The presented search string creation module itself also reveals several areas for further improvement. In particular, participants reported that it was sometimes unclear how individual keywords were introduced into the search string. For systematic reviews, transparency and reproducibility of the search strategy are essential requirements. Improving the traceability of automatically generated terms therefore represents an important direction for future development. Future work may also extend the functionality of the module by improving the visualization of search strategies and providing more detailed explanations for AI-generated modifications. Additional evaluation studies with larger participant groups would further help to assess the effectiveness of the approach in practical systematic review workflows.

CONCLUSION

This work presented an AI-assisted search string creation module integrated into the NeutrinoReview platform. The system supports researchers during the development of search strategies by generating initial search string drafts, suggesting synonyms, enabling natural language based query refinement, and validating search results using seed papers. The results of the conducted usability study indicate that the proposed system can improve the usability and efficiency of search string creation compared to the conventional PubMed interface. Participants reported that the structured workflow and automated support mechanisms simplified the process of constructing and refining search queries. At the same time, the evaluation highlights the importance of maintaining transparency and user control when integrating automated functionality into systematic review workflows. Ensuring that users understand how search queries are constructed remains essential for maintaining the reproducibility and reliability of systematic reviews. Overall, the findings suggest that interactive AI-assisted tools represent a promising approach for supporting systematic review workflows. By combining automated assistance with user-guided refinement, such systems have the potential to reduce the effort required for search strategy development while preserving methodological rigor.

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

https://doi.org/10.5281/zenodo.22921381

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ACKNOWLEDGEMENTS

The joint CERN and WHO ARIA project is funding the PhD project, in the context of which this paper was written. Furthermore, we greatly thank OpenWebSearch.EU[5] project and the members for their help and support with this publication.

REFERENCES

[1] N. R. Haddaway, M. Land, and B. Macura. "A little learning is a dangerous thing": a call for better understanding of the term systematic review. *Environment International*, 99:356–360, 2017.

[2] A. P. Siddaway, A. M. Wood, and L. V. Hedges. How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology*, 70:747–770, 2019.

[3] Hosseini, Mohammad-Salar and Jahanshahlou, Farid and Akbarzadeh, Mohammad Amin and Zarei, Mahdi and Vaez-Gharamaleki, Yosra. Formulating research questions for evidence-based studies. Journal of medicine, surgery, and public health, 2024

[4] F. García-Peñalvo and A. Vázquez-Ingelmo. What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in generative AI. *International Journal of Interactive Multimedia and Artificial Intelligence*, 2023.

[5] L. S. Uman. Systematic reviews and meta-analyses. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 20(1):57, 2011.

[6] C. Ebert and P. Louridas. Generative AI for software practitioners. *IEEE Software*, 40(4):30–38, 2023.

[7] L. F. Alves, F. J. S. Vasconcellos, and B. M. Nogueira. SeSG: a search string generator for secondary studies with hybrid search strategies using text mining. *Empirical Software Engineering*, 27(5):105, 2022.

[8] E. Sandner, A. Wagner, I. Jakovljevic, and C. Gütl. NeutrinoReview: Concept proposal for an open source review management tool. *CERN*, pages 1–6, 2024.

[9] J. C. Carver, E. Hassler, E. Hernandes, and N. A. Kraft. Identifying barriers to the systematic literature review process. In *2013 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 203–212. IEEE, 2013.

[10] E. Aromataris and D. Riitano. Constructing a search strategy and searching for evidence. *American Journal of Nursing*, 114(5):49–56, 2014.

[11] I. J. Marshall and B. C. Wallace. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8:1–10, 2019.

[12] J. J. Wang, S. X. Shao, and F. Y. Ye. Identifying seed papers in sciences. *Scientometrics*, 126(7):6001–6011, 2021.

[13] National Library of Medicine. PubMed. Available at: `https://pubmed.ncbi.nlm.nih.gov/`. Accessed: 2026-02-19.

[14] FastAPI. FastAPI documentation. Available at: `https://fastapi.tiangolo.com/`. Accessed: 2026-02-21.

[15] Qwen Team. Qwen2.5: A Party of Foundation Models. Available at: `https://qwenlm.github.io/blog/qwen2.5/`. Accessed: 2026-03-02.

[16] Meta Platforms Inc. React. Available at: `https://react.dev/`. Accessed: 2026-01-03.

[17] The PostgreSQL Global Development Group. PostgreSQL. Available at: `https://www.postgresql.org/`. Accessed: 2026-01-03.

[18] J. Brooke. SUS: A quick and dirty usability scale. In *Usability Evaluation in Industry*, pages 189–194, 1996.

[19] G. Charness, U. Gneezy, and M. A. Kuhn. Experimental methods: between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81(1):1–8, 2012.

[20] T. B. Brown et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[21] E. Sandner, I. Illicic, U. Sharma, I. Jakovljevic, A. Simniceanu, L. Fontana, A. Henriques, A. Wagner, and C. Gütl. Automating Systematic Reviews: API-Powered Bibliographic Data Retrieval Module for NeutrinoReview. In *7th International Open Search Symposium (OSSYM 2025)*, page 13, 2025.

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

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22925155

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

EVALUATING THE IMPACT OF ELIGIBILITY CRITERIA REFORMULATION ON LLM-BASED LITERATURE SCREENING

D. Scharf[*,†,1], E. Sandner[†,1,2], I. Jakovlevic[2], A. Simniceanu[3], L. Fontana[3], A. Henriques[2], A. Wagner[2], C. Gütl[1]

[1] Graz University of Technology, 8010 Graz, Austria

[2] CERN, 1211 Geneva, Switzerland

[3] WHO, 1211 Geneva, Switzerland

Abstract

Systematic reviews are one of the most reliable methods for synthesizing all available relevant evidence for a specific set of research questions, but their execution remains time-consuming and resource-intensive. Recent advances in Large Language Models (LLMs) have created new opportunities to automate aspects of the systematic review pipeline, particularly during the screening phase. However, the impact of how eligibility criteria are formulated on the performance of LLM-based screening approaches has received limited attention. Building on a set of existing LLM-based screening algorithms, an evaluation framework was created enabling the systematic comparison of different criteria rewriting strategies. An initial set of seven rewriting strategies is evaluated, including one baseline, three rule-based restructuring approaches, and three LLM-based rephrasing approaches. The results demonstrate that eligibility-criteria formulation can significantly influence screening performance. Across 30 systematic review datasets, no strategy simultaneously achieved the highest sensitivity and the greatest reduction in screening workload, illustrating a trade-off between identifying relevant studies and reducing manual effort. Reformulating criteria as yes/no questions improved screening sensitivity but reduced the potential for workload reduction. In contrast, using only inclusion criteria reduced screening workload while maintaining high sensitivity, although performance varied more substantially across datasets. Overall, the findings demonstrate that the formulation of eligibility criteria can influence the effectiveness of LLM-based screening and highlight the need for further research into alternative rewriting and restructuring strategies. While the evaluation is limited to a specific set of screening algorithms, rewriting strategies, and LLMs and to a single runs per configuration, the results motivate further investigation of alternative and combined rewriting strategies. Future work should also evaluate newer models and repeated runs to assess the robustness and generalizability of the observed effects.

INTRODUCTION

Systematic reviews are an essential evidence-based practice aimed at identifying all relevant and available primary literature for a set of predefined research questions and eligibility criteria. They are widely regarded as one of the most reliable ways to identify, select, evaluate, and synthesize evidence from primary studies to consolidate existing knowledge, identify gaps, and evaluate the underlying research questions [1]. This high degree of systematic work and detail comes with the downside of being very time- and cost-intensive [2, 3]. LLMs, with their inherent reasoning capabilities, have opened the door for new ways to automate certain aspects of the systematic review pipeline. The Title-and-Abstract screening phase is a key focus of systematic review automation, as it involves a substantial screening workload while relying on concise and relatively standardized information, making it particularly suitable for automated relevance assessment.

While various approaches have been proposed and evaluated, existing research has primarily focused on prompt engineering to improve the decision-making process [4–7]. Comparatively little attention has been given to the review-specific input provided to the algorithm as natural-language text: the eligibility criteria, and how their formulation, structure, and presentation ultimately guide the LLM's screening decisions. Building on this identified research gap, this paper addresses the following research questions:

RQ1: To what extent do eligibility criteria rewriting strategies improve sensitivity and workload reduction of LLM-based screenings compared to the original eligibility criteria?

RQ2: What recommendations for structuring and formulating eligibility criteria can be derived from the evaluation results?

The remainder of this paper presents the relevant background and related work, describes the proposed evaluation framework and experimental evaluation, and discusses the resulting findings, limitations, and implications. Further details on the methodology, implementation, and evaluation can be found in the corresponding Master's thesis [8] and its supplementary material[1].

BACKGROUND AND RELATED WORK

Given that the main advantage of a systematic review lies within its structured nature, multiple different institutions [9, 10] as well as dedicated guidelines [1, 11, 12] have proposed

[*] daniel.scharf@tugraz.at

[†] Both authors contributed equally to this work.

[1] https://github.com/Dansch723/eligibility_criteria_evaluation

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22925155

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

their own, often slightly different, methodology of a systematic review, but they can all be broken down to the following essential phases:

1. *Review Planning and Protocol Development:* Defines the research question, eligibility criteria, search strategy, and databases in a predefined protocol to ensure a transparent, reproducible, and bias-reduced review process.
2. *Literature Retrieval and Deduplication:* Systematically retrieves relevant literature from multiple databases and removes duplicate records to create the collection used for screening.
3. *Title-and-Abstract (TiAb) Screening:* Applies the predefined eligibility criteria to titles and abstracts to exclude clearly irrelevant studies while retaining uncertain studies for full-text assessment.
4. *Full-Text (FT) Screening:* Evaluates the full publications against the eligibility criteria to determine final inclusion or exclusion.
5. *Data Extraction:* Systematically collects relevant study characteristics, methods, and outcomes using standardized extraction forms to prepare the evidence for analysis.
6. *Data Synthesis and Reporting:* Combines and interprets the findings of included studies through narrative or quantitative synthesis to answer the research question.

The Title-and-Abstract (TiAb) screening phase is particularly well suited for automation as it needs to assess a large volume of records given a standardized format of titles, abstracts, and eligibility criteria. Most existing LLM-based literature screening approaches for TiAb screening represent these eligibility criteria in textual form and incorporate them directly into the screening prompt, but differ primarily in how the screening task is formulated. These approaches can broadly be divided into two groups: those that treat screening as a binary classification task, where each study is either included or excluded, and those that formulate it as a multi-class classification task and subsequently apply a decision-making mechanism, often based on predefined thresholds, to derive the final screening decision.

Representative examples for multi-class screening approaches include the approaches by Dennstädt et al. [5] and Sandner et al. [6]. Dennstädt et al. [5] utilize an LLM to assign each record a five-point Likert relevance score, which is subsequently converted back into a binary screening decision based on a predefined threshold. In contrast, Sandner et al. [6] frame screening as a five-tier classification task, with each tier indicating the degree of confidence that a study meets the eligibility criteria, followed by a threshold-based decision-making mechanism that again obtains a binary final screening decision. Both approaches utilize the eligibility criteria as they were published by their respective systematic reviews, without any restructuring or reformulation.

Existing LLM-based binary-classification approaches similarly only use the original eligibility criteria during the screening without any reformulation or major restructuring. In their respective works, Wang et al. [13] and Guo et al. [4] both investigate the accuracy of LLMs for binary screening decision-making utilizing zero-shot settings. While both approaches achieve competitive accuracy, they primarily focus on the classification of eligibility criteria and do not investigate whether reformulating the criteria themselves could further improve screening performance. Similarly, the CAL-X screening algorithm, proposed by Sandner et al. [7], also used the original eligibility criteria, but additionally adds a second layer to the LLM-based binary screening consisting of a next-token-likelihood (NTL) based reclassification of all excluded records, to further improve sensitivity.

Eligibility criteria are typically formulated by human reviewers in natural language and are primarily intended to support human screening decisions. As a result, their formulation may reflect domain-specific conventions, reviewer preferences, or implicit assumptions that are understandable to human screeners but are not necessarily optimal for LLM-based screening. To improve the consistency and understanding of eligibility criteria, several structured frameworks have been proposed to organize them into categories, which can also support LLM-based screening. One of the most prominent ones is the PICO framework, which structures eligibility criteria into the four categories Population, Intervention, Comparison, and Outcome [14]. These criteria frameworks are subsequently used in LLM-based screening approaches to assist human reviewers in assessing studies and to provide LLMs with a structured basis for their screening decisions [15].

The utilization of LLMs for rephrasing criteria strings has also been investigated in the context of LLM-based question-answering frameworks, where LLMs are used to rephrase criteria strings into explicit yes-or-no questions, which are then used for prompting [16]. Similarly, LLMs have been employed for prompt rephrasing, enabling them to enhance and restructure prompts to improve their clarity and facilitate better interpretation by the screening LLM [17].

As existing studies on prompt and eligibility-criteria rewriting have largely focused on individual use cases, a systematic framework for comparing their effectiveness in the context of systematic reviews is still lacking. This gap motivates the development of a systematic evaluation framework, which is subsequently used to assess an initial set of eligibility-criteria rewriting strategies.

METHODOLOGY

To answer the proposed research questions, a structured evaluation framework was created to enable a systematic and reproducible evaluation of an initial set of eligibility criteria rewriting strategies. These strategies were evaluated on a set of three existing LLM-based literature screening algorithms, which were re-implemented in the framework. This section describes the systematic review datasets used during the evaluation, the three implemented screening algorithms, and

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22925155

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

the seven rewriting strategies, which form the initial set used for the evaluation of this paper.

Datasets

The initial version of the proposed evaluation framework supports datasets from two sources of published systematic reviews. The first is the SYNERGY study selection dataset collection, comprising 26 systematic review datasets spanning domains including medicine, psychology, computer science, chemistry, and biology [18]. The second is a collection of four clinical systematic review datasets from the medical domain [4]. In total, the framework includes 30 datasets, all of which were previously used to evaluate the CAL-X screening algorithms [7] and are subsequently reused for the evaluation presented in this paper. Table 1 provides an overview of the 30 evaluation datasets, comprising a total of 169,726 records. It also reports the number of inclusion and exclusion criteria defined for each dataset.

Table 1: Overview of the 30 evaluation datasets supported by the initial version of the evaluation framework.

Origin	Dataset ID	Records	Incl. Crit.	Excl. Crit.
SYNERGY [18]	Appenzeller_Herzog_2019	2,192	6	2
	Bos_2018	4,816	5	1
	Brouwer_2019	37,643	7	3
	Chou_2003	1,797	3	0
	Chou_2004	1,433	5	5
	Donners_2021	249	3	0
	Hall_2012	8,792	4	2
	Jeyaraman_2020	1,171	4	3
	Leenaars_2019	5,717	2	7
	Leenaars_2020	6,793	1	5
	Meijboom_2021	825	5	1
	Menon_2022	969	6	3
	Moran_2021	5,156	4	3
	Muthu_2021	2,695	3	3
	Nelson_2002	357	2	1
	Oud_2018	935	1	1
	Radjenovic_2013	5,934	4	5
	Sep_2021	270	5	5
	Smid_2020	2,512	6	4
	van_de_Schoot_2018	4,467	4	0
	van_der_Valk_2021	715	1	3
	van_der_Waal_2022	1,957	4	1
	van_Dis_2020	8,948	2	5
	Walker_2018	46,625	4	4
	Wassenaar_2017	7,228	5	9
	Wolters_2018	3,893	3	0
Clinical Reviews [4]	IVM	258	4	2
	LPVR	1,314	3	1
	RAYNAUDS	844	4	1
	SSRI	3,221	4	1
	Total	169,726	114	81

Screening Algorithms

The initial version of the evaluation framework supports three screening algorithms, which are subsequently used in the evaluation presented in this paper. The *5-Tier Prompting Approach* proposed by Sandner et al. [6] serves as the basis for the first two screening algorithms, while the *CAL-X Screening Algorithm* proposed by Sandner et al. [7] forms the basis for the third. These approaches were selected because both utilize textual representations of eligibility criteria during screening, allowing rewritten criteria strings to be readily integrated into the screening process. Furthermore, they employ different screening objectives: the 5-Tier prompting approach formulates screening as a multi-class classification task followed by threshold-based decision making, whereas the CAL-X screening algorithm employs a binary classification setup followed by NTL-based reclassification. The three implemented screening algorithms are:

- **5-Tier-Single-Prompt Screening Algorithm**
  This screening algorithm is a basic re-implementation of the 5-Tier prompting approach as proposed by Sandner et al. [6]. Utilizing a system prompt consisting of role-play, rule-set, and response instructions, the original binary screening problem is transformed into a five-class classification setting. For each record, based on its title, abstract, and the current eligibility criteria, the LLM should assign it to one of the five relevance tiers ranging from "1 Highly Relevant" down to "5 Not relevant". Following screening, the returned relevance tiers are mapped back to binary screening decisions, with responses corresponding to tiers 1-4 classified as *include* and responses corresponding to tier 5 classified as *exclude*.
- **5-Tier-Multi-Prompt Screening Algorithm**
  The 5-Tier-Multi-Prompt screening algorithm is a variation of the previously explained 5-Tier-Single-Prompt screening algorithm. Rather than prompting the LLM once using all eligibility criteria and obtaining a single response for each record, the LLM is prompted separately for each criterion. The individual screening responses are then collected and interpreted using the same approach as in the single-prompt scenario. If none of the individual responses for a record is assigned to tier 5, the final screening decision is *include*, otherwise, it is *exclude*.
- **Thresh-X Screening Algorithm**
  The Thresh-X screening algorithm is an adaptation of the CAL-X screening algorithm [7]. The first phase remains unchanged from the CAL-X algorithm, being a classical binary LLM-based classification into either *include* or *exclude*. This binary classification task is set up via a system prompt consisting of role-play and response instructions. In the second phase, records initially classified as *include* retain their original classification, while records classified as *exclude* undergo an additional assessment based on the next-token likelihood (NTL) of the original response, which serves as a measure of the LLM's confidence in its initial decision. If the NTL exceeds a predefined threshold, the record remains *exclude*, otherwise, it is reclassified as *include*. This reclassification step therefore aims to identify and recover uncertain exclusions, thereby reducing the risk of incorrect exclusion decisions. For the predefined threshold values, instead of utilizing threshold values that are calibrated for different sensitivity

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22925155

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

values, heuristically chosen values were used. This design decision was made as otherwise the algorithm would need to be recalibrated for every evaluation run, which would result in a substantial increase in computational cost and execution time given the large-scale evaluation conducted in this paper. The new threshold values are 50, 75, and 90. These values were chosen as they provide intuitive decision boundaries.

Rewriting Strategies

The initial set of eligibility-criteria rewriting strategies evaluated in this study comprises seven approaches: one baseline, three rule-based strategies, and three LLM-based strategies. This paper presents an overview of these seven strategies. The detailed prompts used for each strategy can be found in the supplementary material [2].

- **BASIC** (Baseline)
  The baseline approach uses the eligibility criteria as they were published in their original systematic reviews, without any rewriting or restructuring.
- **IF-ELSE-STRUCTURE** (Rule-based)
  This strategy transforms the original eligibility criteria into an explicit if-else construct string, thereby reflecting an intuitive decision-making process. An example if-else construct string is shown in Listing 1.

```
if (Studies that used ivermectin for
prophylaxis of COVID-19) {
    Exclude!
}
else{
    if (Randomized controlled trials and
    non-randomized comparative cohort studies)
    {
        if (Compared ivermectin against
        standard of care or control
        group/placebo) {
            Include!
        } else {
            Exclude!
        }
    } else {
        Exclude!
    }
}
```

Listing 1: Restructured eligibility criteria given a subset of the criteria of the `IVM` dataset for the 5-Tier-Single-Prompt screening algorithm.

- **INCLUDE-ONLY & EXCLUDE-ONLY** (Rule-based)
  These two strategies build upon the concept of using either only inclusion criteria or only exclusion criteria [19]. For these two strategies, the corresponding criteria type is discarded: *INCLUDE-ONLY* uses only the original inclusion criteria, while *EXCLUDE-ONLY* uses only the original exclusion criteria. Datasets lacking the respective criteria type are excluded from the evaluation of the corresponding strategy.
- **QUESTION-REPHRASING** (LLM-based)
  The *QUESTION-REPHRASING* strategy is motivated by related work in question answering and reformulation frameworks [16], and rewrites each criterion into a question that can be answered either with *yes* if the criterion is fulfilled and *no* otherwise. An LLM is utilized for the rewriting itself using a system prompt consisting of a combination of role-play and rule-set instructions. For the evaluation conducted in this work, the same LLM is always used for both the rewriting and the actual screening itself. Table 2 displays one example of an original eligibility criterion and its rewritten yes-or-no question counterpart.

Original	Exclude: Studies that used ivermectin for prophylaxis of COVID-19.
Question	Was ivermectin used for prophylaxis of COVID-19 in this study?

Table 2: Eligibility criterion example and its rewritten yes-or-no question counterpart given a criterion of the IVM dataset.

- **INCLUSION-REWRITING & EXCLUSION-REWRITING** (LLM-based)
  Similar to their lightweight counterparts, *INCLUDE-ONLY* and *EXCLUDE-ONLY*, these two strategies follow the principle of using only a single eligibility-criteria type. However, instead of discarding the respective counterpart criteria, they transform them into the selected criteria type. Specifically, *INCLUSION-REWRITING* transforms all exclusion criteria into inclusion criteria, whereas *EXCLUSION-REWRITING* transforms all inclusion criteria into exclusion criteria. As with the *QUESTION-REPHRASING* strategy, an LLM is used to perform the rewriting based on a system prompt containing role-play, rule-set, and response-format instructions. The same LLM is used for both the criteria rewriting and the subsequent screening process.

Original Include	Included laboratory-confirmed, hospitalized COVID-19 patients
Rewritten Exclude	Studies that do not include laboratory-confirmed, hospitalized COVID-19 patients.
Original Exclude	Studies that used ivermectin for prophylaxis of COVID-19
Rewritten Include	Studies that did not use ivermectin for the prophylaxis of COVID-19.

Table 3: Eligibility criterion examples and their rewritten counterparts given criteria of the IVM dataset.

[2] `https://github.com/Dansch723/eligibility_criteria_evaluation`

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22925155

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

EXPERIMENTAL EVALUATION

The described rewriting strategies were experimentally evaluated. This section presents the utilized LLMs, the employed performance metrics, and the detailed evaluation plan, including the measures taken to make the evaluation computationally feasible given the available resources.

Large Language Models (LLMs)

Three locally hosted open-weight LLMs were used for the evaluation conducted in this paper, namely the Llama 3.1 8b Instruct model [20], the Qwen 2.5 32B model [21], and the GPT-OSS 20B model [22].

Evaluation Metrics

For the evaluation, sensitivity and workload reduction were selected as the primary evaluation metrics, as defined in Equations 1 and 2.

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (1)$$

$$\text{Workload Reduction} = \frac{\text{True Negatives} + \text{False Negatives}}{\text{N}} \quad (2)$$

Sensitivity is the proportion of relevant studies correctly identified and was chosen because systematic review screening aims to minimize the number of eligible studies that are missed. Workload reduction measures the framework's ability to reduce the number of studies requiring manual review.

Evaluation Plan

As the proposed evaluation spans multiple dimensions, including 30 datasets, seven rewriting strategies, three screening algorithms, and three LLMs, an initial feasibility test was conducted to estimate the overall execution time. Using the small IVM dataset, comprising 258 records, execution times were measured and extrapolated to the remaining datasets, resulting in an estimated total runtime of 385 days with the available computational resources. Consequently, an evaluation plan was developed to enable the evaluation of the presented rewriting strategies within a reasonable time frame.

This evaluation plan consists of two parts: a multi-stage pre-evaluation using subsets and a final full evaluation utilizing only a subset of the original parameter configurations. For the pre-evaluation first, the experiments are conducted on a subset of each dataset with a maximum size of 2,000 records. For each dataset, all originally included records are retained, after which excluded records are added until either all records have been included or the 2,000-record limit is reached. This reduces the total number of records to 25.5% of the original size and lowers the estimated runtime to approximately 91 days. As this time frame remains too long for the scope of the pre-evaluation, the multi-stage evaluation with progressive parameter elimination is introduced.

The multi-stage pre-evaluation progressively eliminates underperforming parameter configurations to reduce computational requirements, rather than evaluating all configurations across all datasets. In the first two stages, configurations are primarily assessed based on sensitivity, while workload reduction is incorporated from the third stage onward, with decisions based on average performance across the evaluated datasets. The evaluation consists of four stages, using 5, 10, 20, and 30 datasets, respectively. While the number of datasets increases in each consecutive stage, only the most promising LLMs, screening algorithms, and rewriting strategies are retained for subsequent stages.

RESULTS

This section first reports the intermediate results of each stage of the multi-stage pre-evaluation, followed by the results achieved using the final parameter configuration on all 30 full datasets. For the scope of this paper, only those intermediate results of the pre-evaluation are presented that resulted in a reduction of one or more parameter dimensions. A full summary of all evaluation results can be found in the supplementary material [3].

Multi-Stage Pre-Evaluation

- **Stage 1**
  In the first stage, the *5-Tier-Multi-Prompt* and the *Thresh-50* screening algorithm showed the lowest average sensitivity values across all rewriting strategies, LLMs, and datasets, as it is displayed in Table 4. These two screening algorithms were therefore excluded from all subsequent evaluation stages, while the better-performing algorithms were retained. The total execution time for Stage 1 was 149.26 hours, of which 14.12 hours were required by *Llama 3.1 8B Instruct*, 30.10 hours by *Qwen 2.5 32B*, and the remaining 105.04 hours by *GPT-OSS 20B*.

Screening Algorithm	5-tier-single	Thresh-50	Thresh-75	Thresh-90	5-tier-multi
Average Sensitivity	85.85	63.63	66.48	67.61	61.21

Table 4: Evaluation results pre-evaluation stage 1 for the screening algorithm-dimension.

- **Stage 2**
  In stage 2 the *GPT-OSS 20B* LLM was excluded from subsequent evaluation stages due to its substantially lower average sensitivity of 62.86% compared to 99.08% of the *Llama 3.1 8B Instruct* model and the 75.26% of the *Qwen 2.5 32B* model. An overview of the average sensitivity values per LLM for the first two stages is shown in Table 5. The total execution time of stage 2 was 106.94 hours, whereas a large portion of this time could again be attributed to the experiments

[3] https://github.com/Dansch723/eligibility_criteria_evaluation

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22925155

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

conducted using the *GPT-OSS 20B* LLM. It is noticable that the elimination of two screening algorithms in stage 1 reduced the overall execution time of this stage by 28.35%. The time allocation per LLM is 7.74 hours for the *Llama 3.1 8B Instruct*, 15.18 hours for the *Qwen 2.5 32B*, and 84.02 hours for the *GPT-OSS 20B LLM*.

LLMs	Stage 1	Stage 2	Average
Llama 3.1 8B Instruct	98.50	99.65	99.08
Qwen 2.5 32B	62.77	87.75	75.26
GPT-OSS 20B	58.66	67.05	62.86

Table 5: Average Sensitivity evaluation results pre-evaluation stage 2 for the LLM-dimension.

- **Stage 3**
  In Stage 3, the parameter space was reduced along two dimensions.
  First, three rewriting strategies (*IF-ELSE-STRUCTURE*, *EXCLUDE-ONLY*, and *EXCLUSION-REWRITING*) were removed from subsequent stages. This reduction was based on the project team's decision that the final parameter configuration should include both LLM-based and rule-based strategies. Accordingly, the two lowest-performing rule-based strategies, measured by average sensitivity, were removed. Among the LLM-based strategies, the one with the lowest combined average sensitivity and average workload reduction was eliminated. The corresponding average metric values is shown in Table 6.

Rewriting Strategies		Sensitivity	Workload Reduction
Baseline	BASIC	86.98	47.87
Rule-based	IF-ELSE	84.63	38.90
	INCL. ONLY	84.81	52.60
	EXCL. ONLY	83.94	24.63
LLM-based	QUEST. RPHR	85.14	41.68
	INCL. RW	87.41	47.64
	EXCL. RW	85.92	35.91

Table 6: Evaluation results pre-evaluation stage 3 for the rewriting strategy-dimension.

Second, in stage 3, the *Qwen 2.5 32B* LLM was excluded from further evaluation stages after consistently achieving lower average sensitivity values than its Llama counterpart across all configurations. Across all configurations the *Qwen 2.5 32B* LLM achieved only an average sensitivity of 72.60% compared with 98.49% of the *Llama 3.1 8B Instruct* model, as is shown in Table 7.
The total execution time of the experiments of stage 3 was 43.66 hours. Even though the number of datasets used during this stage doubled compared to its two predecessors, the required time decreased substantially, only needing 40.83% of the time needed in stage 2.

Rewriting Strategy / LLM	Llama 3.1 8B Instruct	Qwen 2.5 32B
BASIC	98.61	75.32
IF-ELSE	99.49	69.77
INCL. ONLY	96.51	73.11
EXCL. ONLY	99.80	68.09
QUEST. RPHR	99.02	71.26
INCL. RW	97.82	76.99
EXCL. RW	98.21	73.63
Average	98.49	72.60

Table 7: Average Sensitivity evaluation results pre-evaluation stage 3 for the LLM-dimension.

- **Stage 4**
  Finally, in stage 4, *Thresh-90* was selected as the screening algorithm for the final parameter configuration due to its highest average sensitivity value of 99.27%, outperforming *Thresh-75* and *5-Tier-Single-Prompt* as detailed in Table 8. The total execution time of the experiments of stage 4 was 9.06 hours. After all four stages of the pre-evaluation, the total execution time for all experiments summed to 308.91 hours.

Screening Algorithm	Rewriting Strategies				Average
	BASIC	INCL. ONLY	QUEST. RPHR	INCL. RW	
5-tier-single	97.10	95.77	97.56	97.09	96.88
Thresh-75	98.40	95.51	99.43	98.16	97.88
Thresh-90	99.43	98.50	99.95	99.22	99.27

Table 8: Average Sensitivity evaluation results pre-evaluation stage 4 for the screening algorithm-dimension.

Final Full Evaluation

After the multi-stage pre-evaluation the final configuration was as follows:

- LLM: *Llama 3.1 8B Instruct*
- Screening Algorithm: *Thresh-90*
- Rewriting Strategies: *BASIC, INCLUDE-ONLY, QUESTION-REPHRASING,* and *INCLUSION-REWRITING*

These parameter configurations were subsequently used for the final full-dataset evaluation, in which all 30 datasets were evaluated in their entirety. The experiments required a total execution time of 55.88 hours. The results of this final evaluation are presented in Tables 9 and 10.

DISCUSSION

The full-dataset evaluation provides a clear answer to RQ1, showing that reformulating eligibility criteria can affect

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22925155

Screening Algorithm	Metrics	Rewriting Strategies			
		BASIC	INCL. ONLY	QUEST. RPHR	INCL. RW
Thresh-90	Sensitivity	99.43	98.50	99.95	99.22
	Workload Reduction	19.86	27.27	6.66	19.09

Table 9: Average Sensitivity and Workload Reduction values (%) across all 30 datasets per rewriting strategy.

Screening Algorithm	Sensitivity Metrics	Rewriting Strategies			
		BASIC	INCL. ONLY	QUEST. RPHR	INCL. RW
Thresh-90	Maximum	100.00	100.00	100.00	100.00
	Minimum	93.33	81.82	98.61	92.86

Table 10: Maximum and minimum sensitivity values (%) from all 30 datasets per rewriting strategy.

both the sensitivity and workload reduction of LLM-based screening. The baseline *BASIC* strategy, which uses the original criteria, achieved an average sensitivity of 99.43% and a workload reduction of 19.86% across all 30 datasets. *QUESTION-REPHRASING*, which reformulates each criterion as a yes/no question, achieved the strongest overall performance, increasing average sensitivity to 99.95% while maintaining a workload reduction of 6.66%. Notably, its sensitivity never fell below 98.61% across the evaluated datasets. In contrast, *INCLUDE-ONLY*, which discards all exclusion criteria, achieved the highest workload reduction (27.27%) but a lower average sensitivity of 98.50%, with sensitivity dropping as low as 81.82% on individual datasets. *INCLUSION-REWRITING*, which converts exclusion criteria into logically equivalent inclusion criteria, performed similarly to the baseline, with 99.22% sensitivity and 19.09% workload reduction.

Overall, the results demonstrate that eligibility-criteria reformulation can improve screening sensitivity compared with using the original criteria, with question-based reformulation providing the most robust improvement. At the same time, the results reveal a clear trade-off between sensitivity and workload reduction: more conservative reformulations can achieve very high sensitivity but reduce the number of records filtered out, whereas more aggressive simplifications can substantially reduce workload at the expense of sensitivity and robustness. Thus, the answer to RQ1 is that eligibility-criteria reformulation can improve LLM-based screening performance, but the magnitude and direction of the improvement depend strongly on the chosen rewriting strategy.

For RQ2, the results suggest several practical recommendations for structuring and formulating eligibility criteria. Across both the multi-stage evaluation and the full-dataset evaluation, QUESTION-REPHRASING showed the most consistent performance, achieving sensitivity comparable to or exceeding the original criteria and never falling below 98.61% across the 30 datasets. This indicates that expressing eligibility criteria as explicit yes/no questions can provide a robust formulation for LLM-based screening. The evaluation further indicates that inclusion-oriented formulations are preferable to exclusion-oriented formulations. Both *INCLUDE-ONLY* and *INCLUSION-REWRITING*, which rely exclusively on inclusion criteria, generally performed better than their exclusion-oriented counterparts. This suggests that, where possible, eligibility criteria should be formulated positively as conditions that a study must satisfy rather than primarily as conditions that lead to exclusion. However, the results also demonstrate that this choice involves a trade-off: *INCLUDE-ONLY* achieved the highest workload reduction but showed substantially lower sensitivity and greater variation across datasets.

LIMITATIONS AND FUTURE WORK

The evaluation is subject to several limitations. First, the findings are restricted to the three implemented screening algorithms and the evaluated rewriting strategies and therefore cannot be generalized to other approaches. Second, the evaluation relies on three LLMs that were selected at the planning stage and are no longer state-of-the-art, meaning that results may differ when using newer models. The single-threaded LLM API also increased execution time, although it did not affect the evaluation methodology or screening correctness. Finally, each configuration was evaluated only once due to the high computational cost, leaving the results potentially affected by the inherent nondeterminism of LLMs. Repeated runs would provide more robust estimates of performance.

Future work should primarily extend the modular evaluation framework with additional eligibility-criteria rewriting strategies, particularly approaches that can improve or maintain the high sensitivity achieved by *QUESTION-REPHRASING* while increasing workload reduction. Combining existing strategies, such as *INCLUDE-ONLY* and *QUESTION-REPHRASING*, represents a promising direction for achieving a more favourable trade-off between sensitivity and workload reduction. Further research should also evaluate newer and domain-specific LLMs, repeat experiments to account for LLM nondeterminism, and integrate additional screening algorithms and systematic review datasets. Expanding these evaluation dimensions would provide a broader basis for assessing the effectiveness and generalizability of eligibility-criteria rewriting strategies.

CONCLUSION

Systematic review screening represents a promising area for the application of Large Language Models, but the effectiveness of LLM-based screening depends not only on the underlying model and screening approach but also on how the eligibility criteria are formulated. This work addressed this aspect by introducing a modular evaluation framework for systematically investigating eligibility-criteria rewriting strategies in LLM-based systematic review screening. The framework enables different rewriting approaches, screening algorithms, and LLMs to be evaluated across a broad range of systematic review datasets and can be readily extended

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

https://doi.org/10.5281/zenodo.22925155

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

with additional strategies and evaluation dimensions. The evaluation demonstrates that reformulating eligibility criteria can substantially influence both screening sensitivity and workload reduction. In particular, expressing criteria as explicit yes/no questions provided the most consistent sensitivity, while using only inclusion criteria offered greater workload reduction at the cost of lower and more variable sensitivity. These results highlight a fundamental trade-off between minimizing missed relevant studies and reducing the number of records requiring manual assessment. Consequently, there is no single rewriting strategy that is optimal for all screening objectives, and the preferred formulation depends on the desired balance between sensitivity and workload reduction.

Overall, the findings show that eligibility criteria should be considered an important component of LLM-based screening rather than merely a static input inherited from the original systematic review. The proposed framework provides a foundation for systematically studying this aspect and facilitates the development and comparison of further criteria-rewriting approaches.

ACKNOWLEDGEMENTS

The joint CERN and WHO ARIA[4] project is funding the PhD project, in the context of which this paper was written. Additionally, this funding supported a two-month internship at CERN for one of the authors, during which parts of this work were developed. Furthermore, we greatly thank the OpenWebSearch.EU[5] project and its members for their help and support with this publication.

REFERENCES

[1] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele University and Durham University, EBSE Technical Report EBSE-2007-01, version 2.3, 2007, https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf

[2] M. Michelson and K. Reuter, "The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials," *Contemporary Clinical Trials Communications*, vol. 16, p. 100 443, 2019, doi:doi.org/10.1016/j.conctc.2019.100443

[3] B. Nussbaumer-Streit et al., "Resource use during systematic review production varies widely: A scoping review," en, *J. Clin. Epidemiol.*, vol. 139, pp. 287–296, 2021.

[4] E. Guo, M. Gupta, J. Deng, Y.-J. Park, M. Paget, and C. Naugler, "Automated paper screening for clinical reviews using large language models: Data analysis study," en, *J. Med. Internet Res.*, vol. 26, e48996, 2024.

[5] F. Dennstädt, J. Zink, P. M. Putora, J. Hastings, and N. Cihoric, "Title and abstract screening for literature reviews using large language models: An exploratory study in the biomedical domain," en, *Syst. Rev.*, vol. 13, no. 1, p. 158, 2024.

[6] E. Sandner et al., "Screening automation for systematic reviews: A 5-tier prompting approach meeting cochrane's sensitivity requirement," in *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, 2024, pp. 150–159, doi:10.1109/FLLM63129.2024.10852425

[7] E. Sandner et al., "Cal-x: Enhancing systematic review screening with llms and next-token likelihood calibration," in *2025 3rd International Conference on Foundation and Large Language Models (FLLM)*, 2025, pp. 1071–1078, doi:10.1109/FLLM67465.2025.11391146

[8] D. Scharf, "Evaluating the impact of eligibility criteria reformulation on LLM-based systematic review screening," Master's Thesis, Graz University of Technology, 2026, doi:10.3217/5w7jg-0ah69

[9] J. P. T. Higgins et al., Eds., *Cochrane Handbook for Systematic Reviews of Interventions*, Version 6.5. Cochrane, 2024, Updated August 2024, https://www.cochrane.org/handbook

[10] M. J. Page et al., "The prisma 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, 2021, doi:10.1136/bmj.n71

[11] E. Calderon Martinez et al., "Ten steps to conduct a systematic review," en, *Cureus*, vol. 15, no. 12, e51422, 2023.

[12] A. Pollock and E. Berge, "How to do a systematic review," en, *Int. J. Stroke*, vol. 13, no. 2, pp. 138–156, 2018.

[13] S. Wang, H. Scells, S. Zhuang, M. Potthast, B. Koopman, and G. Zuccon, *Zero-shot generative large language models for systematic review screening automation*, 2024, https://arxiv.org/abs/2401.06320

[14] W. S. Richardson, M. C. Wilson, J. Nishikawa, and R. S. Hayward, "The well-built clinical question: A key to evidence-based decisions," *ACP J Club*, vol. 123, no. 3, A12–A13, 1995.

[15] S. K. Vallamchetla et al., "Do it faster with picos: Generative ai-assisted systematic review screening," *Journal of Biomedical Informatics*, vol. 168, p. 104 860, 2025, doi:doi.org/10.1016/j.jbi.2025.104860

[16] O. Akinseloyin, X. Jiang, and V. Palade, "A question-answering framework for automated abstract screening using large language models," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 1939–1952, 2024, doi:10.1093/jamia/ocae166

[17] Y. Deng, W. Zhang, Z. Chen, and Q. Gu, *Rephrase and respond: Let large language models ask better questions for themselves*, 2024, https://arxiv.org/abs/2311.04205

[18] J. De Bruin, Y. Ma, G. Ferdinands, J. Teijema, and R. Van de Schoot, *SYNERGY - Open machine learning dataset on study selection in systematic reviews*, version V1, 2023, doi:10.34894/HE6NAQ

[19] A. Sciurti et al., "Compact large language models for title and abstract screening in systematic reviews: An assessment of feasibility, accuracy, and workload reduction," en, *Res. Synth. Methods*, vol. 17, no. 2, pp. 332–347, 2026.

[20] Meta, *Llama-3.1-8b-instruct*, https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct, Hugging Face model repository, accessed 2026-06-22, 2024.

[21] Qwen et al., *Qwen2.5 technical report*, 2025, https://arxiv.org/abs/2412.15115

[22] OpenAI et al., *Gpt-oss-120b & gpt-oss-20b model card*, 2025, https://arxiv.org/abs/2508.10925

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

[5] https://openwebsearch.eu

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22925265

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ESTIMATING ABSOLUTE WEB CRAWL COVERAGE FROM LONGITUDINAL SET INTERSECTIONS*

M. Paris† [1,2], G. Paris[3], F. Baumann[4]
[1]Common Crawl Foundation, San Francisco, USA
[2]Humboldt-Universität zu Berlin, Berlin, Germany
[3]Independent Researcher, Berlin, Germany
[4]University of Pennsylvania, Philadelphia, USA

Abstract

Web archives preserve portions of the web, but quantifying their completeness remains challenging. Prior approaches have estimated the coverage of a crawl by either comparing the outcomes of multiple crawlers, or by comparing the results of a single crawl to external ground truth datasets. We propose a method to estimate the absolute coverage of a crawl using only the archive's own longitudinal data, i.e., the data collected by multiple subsequent crawls. Our key insight is that coverage can be estimated from the empirical URL overlaps between subsequent crawls, which are in turn well described by a simple urn process. The parameters of the urn model can then be inferred from longitudinal crawl data using linear regression. Applied to our focused crawl configuration of the German Academic Web (15 semi-annual crawls between 2013–2021), we find a coverage of approximately 46% of the crawlable URL space for the stable crawl configuration regime. Our method is extremely simple, requires no external ground truth, and generalizes to any longitudinal focused crawl.

INTRODUCTION

Focused web crawls—archives targeting specific domains like academic institutions, government websites, or cultural heritage—must terminate at some point. The German Academic Web (GAW), for instance, crawls approximately 150 German university seeds semi-annually, stopping after collecting roughly 100 million records. This raises the question: how much of the *crawlable*[1] German Academic Web (cGAW) is archived by the GAW?

A web crawl is defined by its crawl configuration, which specifies the heuristics and rules to collect websites. For instance, it defines the termination threshold M, i.e., the total number of URLs to be collected in the crawl. What remains unknown, however, is the set of URLs (N) that could, in principle, be collected by a given crawler and its crawl configuration. As a result we are faced with the fundamental limitation of a single, isolated crawl: it does not allow us to determine its coverage $c = M/N$ of the underlying web, i.e. it remains unknown how much of the crawlable web space does it actually capture.

* Parts of this research was funded by the German Federal Ministry of Education and Research (BMBF) in the REGIO project (grant no. 01PU17012D)

† micha@commoncrawl.org

[1] defined by the crawls delimitations (time, termination, revisits. etc.)

Table 1: Positioning of our approach relative to prior work.

Method	Output	Requirements
Bharat & Broder [1]	Relative: $S(A)/S(B)$	Multiple crawlers
Paris & Jäschke [3]	Recall of entity class	Ext. ground truth
This paper	**Absolute:** $c = M/N$	**Longitudinal data**

Several strategies have addressed related problems. Bharat and Broder [1] estimated *relative* search engine sizes by comparing URL sets across multiple engines. Their method yields ratios like "Engine A is 2.3× larger than Engine B" but cannot determine what fraction of the total web either captures. Entity-based approaches use external ground truth (e.g., named entities from reference databases) to measure recall of known items. This requires domain-specific datasets and measures coverage only for that entity class, not the full URL space. Quality frameworks like SHARC [2] assess temporal coherence and resource completeness of archived pages, but not overall domain coverage.

Here we provide an interpretable method to estimate absolute coverage from longitudinal crawl data alone. The contribution is twofold. First, we gain theoretical insight by modeling the crawling process using a simple urn model. And second, we estimate coverage by fitting the model to the longitudinal crawl data.

The remainder of the paper is structured as follows. Section reviews related work on set similarity measures and coverage estimation. Section formalizes the random sampling model and derives the self-intersection interpretation. Section applies our method to the German Academic Web, showing the heatmap, observed decay patterns, and coverage estimates. Section discusses implications and limitations.

RELATED WORK

Estimating an unknown population size from overlapping samples dates to Petersen [4] and Lincoln [5], who formalized the estimator $\hat{N} = n_1 n_2/m$ for wildlife abundance, later bias-corrected by Chapman [6]. Seber [7] provides the standard reference covering closed- and open-population models, including the Jolly–Seber model for populations with births and deaths. Fienberg [8] reconceptualized multiple-recapture as incomplete contingency tables, enabling heterogeneous-catchability extensions.

These ecological methods were first applied to the web by Lawrence and Giles [9], who used pairwise URL overlap across six search engines to estimate the indexable web size

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22925265

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

via Lincoln–Petersen. They noted that positive dependence between engines inflates overlap, making their estimate a lower bound. Lawrence and Giles [10] complemented this with random IP sampling, and Bharat and Broder [1] derived relative size ratios from random-query sampling, yielding only relative sizes and requiring multiple engines. Dobra and Fienberg [11] reanalyzed the data with Bayesian log-linear models incorporating Rasch-type heterogeneity, estimating roughly 2.5× the original figure and demonstrating that ignoring capture heterogeneity causes severe underestimation.

While these studies compared multiple engines at a single point in time, Bar-Ilan [12, 13] conducted the first longitudinal study, tracking a query across engines over five months and categorizing URLs as lost, dropped, forgotten, or recovered. Bar-Ilan [14] formalized measures for temporal self-overlap and relative coverage.

Central to all overlap-based approaches is the choice of similarity measure. The symmetric Jaccard index [15] and Dice–Sørensen coefficient [16, 17] contrast with the asymmetric containment $C(A, B) = |A \cap B|/|A|$ introduced by Broder [18], who also proposed MinHash sketches for its efficient estimation [19]. Garg et al. [20] showed that asymmetric measures often outperform symmetric ones in retrieval tasks.

The probabilistic foundation comes from urn model theory: Kalinka [21] showed that expected intersection sizes under independent sampling follow the hypergeometric distribution, with Mahmoud [22] providing comprehensive treatment of Pólya urn models. This connection to capture-recapture is established in ecology [7] but has not previously been applied to web crawl coverage.

A key assumption of urn-based sampling is a stable population, yet the web undergoes continuous turnover. Koehler [23, 24] found a web page half-life of approximately two years; Gomes and Silva [25] modeled URL persistence; Fetterly et al. [26] and Wren [27] quantified change and decay rates at scale.

Complementary work on web archive quality—capture "blur" and coherence [2], missing embedded resources [28], and temporal inconsistency [29]—focuses on individual page fidelity rather than domain coverage. As Baack [30] found, even the Common Crawl team expressed fundamental uncertainty, with their director noting the web is "practically infinite." Overall, existing work either (1) uses entity-based proxies requiring external datasets [3]; (2) compares multiple engines for only relative sizes [1, 9]; or (3) studies temporal decay without connecting it to coverage [13, 23]. Our contribution is to extract absolute sampling fractions from longitudinal self-intersections of a single archive.

THEORETICAL INTUITION AND URN MODEL

Let u_t denote the set of URLs captured in a given crawl at time t, with $|u_t| = M$ being the number of URLs captured in the crawl. Furthermore, let N denote the size of the crawlable web, which is represented by the size of the set of all URLs within the archive's scope that could potentially be harvested. Note that while we assume to know the value of M, we do not have access to the value of N, as a crawl will not capture all crawlable URLs. The ratio $c = M/N$ is defined as the coverage of the crawl.

The urn represents a population of N unique URLs. Time advances in discrete steps $t = 1, \ldots, T$, which corresponds to successive crawls. At each time step, the urn process consists of two steps: First, turnover occurs in the underlying population. This means that a fraction of $(1 - \alpha)$ of the N URLs is removed and replaced by the same number of newly introduced URLs, such that N remains constant. In other words, the parameter $\alpha \in [0, 1]$ captures the temporal persistence of URLs across two subsequent crawls. Second, conditional on the resulting population, a crawl sample u_t is generated by drawing M URLs uniformly at random without replacement from the urn.

Crucially, this simple model allows us to derive an analytical expression for the overlap between two crawls at different times, u_i and u_j. First, let us consider an arbitrary URL that appears in the crawl at time $t = 1$. The probability that this URL survives the population turnover for $T - 1$ steps is α^{T-1}. Second, conditional on surviving until time T, the probability that it is included in the crawl at time T corresponds to M/N, i.e., the uniform sampling probability from the urn. Since the initial crawl contains M URLs, the expected size of the overlap between the first and the T-th crawl is therefore given by

$$\mathbb{E}[|u_1 \cap u_T|] = M \cdot \alpha^{T-1} \cdot \frac{M}{N} = \frac{M^2}{N}\, \alpha^{T-1}. \tag{1}$$

Dividing by M, and substituting $c = M/N$, yields the expected fraction of URLs in the initial crawl that reappear at time T,

$$f(T) = c\, \alpha^{T-1}. \tag{2}$$

The important feature of Eq. 2 is that it can readily be used to fit the empirical observations of temporal crawl overlaps to extract the value of c. Interestingly, Eq. 2 is closely linked to previous work, specifically, Broder's containment measure, defined as [18]

$$g(u_i, u_j) = \frac{|u_i \cap u_j|}{|u_i|}. \tag{3}$$

As defined in Eq. 3, $g(u_i, u_j)$ denotes the fraction of crawl u_i that also appears in crawl u_j, and has been used for cross-sectional multi-engine comparisons [1]. In contrast to its original conception of g, our model allows to interpret Eq. 3 as the definition of URL overlap between two crawls separated by $t = i - j$, hence

$$g^* = g(u_i, u_i), \tag{4}$$

is interpreted as "self-intersection", which we recover from Eq. 2 by setting $T = 1$ and thus get $g^* = c$. Equivalently, this result can be obtained directly from Eq. 3 by assuming two independent draws ($i = j$) of size M from a population of

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22925265

N URLs, where the expected intersection size follows from the hypergeometric distribution [21], i.e., we get[2]

$$g^* = \mathbb{E}\left[\frac{|u_1 \cap u_T|}{M}\right] = \frac{1}{M}\frac{M^2}{N} = c\,. \tag{5}$$

Importantly, this interpretation holds even if the crawl itself is non-uniform. Consider a fixed crawl u with $|u| = M$ obtained by any sampling method (uniform, breadth-first, preferential, etc.) from population U with $|U| = N$. If we compare against a uniform sample $\hat{u}$ from U, the expected intersection is:

$$\mathbb{E}\left[\frac{|u \cap \hat{u}|}{M}\right] = \frac{M}{N} = c. \tag{6}$$

This follows because each element of u has probability M/N of being sampled into $\hat{u}$, regardless of how u was constructed.

Thus the coverage estimate c has three equivalent interpretations: (i) the fraction of the crawlable web captured by the crawl, (ii) the expected overlap fraction with a uniform sample, and (iii) the normalized self-intersection under independent resampling.

Extending this framework to pairs of crawls, we compare the observed relative overlap $g(u_i, u_j) = |u_i \cap u_j|/M$ against the theoretical baseline of uniformly sampled counterparts, denoted $\hat{u}_i$ and $\hat{u}_j$. Under the urn model with independent uniform sampling, the expected relative overlap is determined solely by the coverage c and the temporal decay α:

$$\mathbb{E}[g(u_i, \hat{u}_j)] = \mathbb{E}[g(\hat{u}_i, \hat{u}_j)] = c \cdot \alpha^{|i-j|} \tag{7}$$

However, real-world crawling processes often exhibit temporal correlations—for instance, through persistent seeds that force the crawler into the same local subgraph across repetitions. We model this deviation by introducing a *crawler bias coefficient* R_{ij}, defined as the ratio between the actual and theoretical expectations:

$$R_{ij} = \frac{\mathbb{E}[g(u_i, u_j)]}{\mathbb{E}[g(u_i, \hat{u}_j)]} = \frac{\mathbb{E}[g(u_i, u_j)]}{c \cdot \alpha^{|i-j|}} \tag{8}$$

The overlap ratio R_{ij} serves as a diagnostic metric for crawler behavior: $R_{ij} \approx 1$ implies a uniform process, while $R_{ij} > 1$ reveals positive correlation where the crawler systematically revisits specific subgraphs. This bias is empirically isolated as the residual term in the log-linear regression:

$$\underbrace{\log R_{ij}}_{\text{Residual}} = \log g(u_i, u_j) - (\log c + |i-j| \log \alpha) \tag{9}$$

RESULTS

We apply our methodology to the German Academic Web (GAW), a longitudinal focused crawl of German university websites. The part of the GAW considered here consists of 15 semi-annual focused crawls conducted between December 2013 and December 2021. Each crawl started from approximately 150 seed domains corresponding to German universities with doctorate-granting rights. The Heritrix crawler [31] was used with a breadth-first traversal policy. Crawls terminate after collecting approximately 100 million records [3].

The data of our longitudinal crawls of the cGAW allows us to empirically observe the containment between two crawls $g(u_i, u_j)$ for $i > j$. The results are depicted in Fig. 1, which shows two views of the pairwise containment measures. Figure 1(a) displays the 15×15 matrix of $g(u_i, u_j)$ values as a heatmap. The off-diagonal entries show clear temporal structure, and values decrease with distance from the diagonal, i.e. for increasing time difference between two crawls, reflecting URL turnover. This visualization makes explicit our estimation target, namely the hypothetical intersection of two independent samples drawn at the same moment, i.e., $\Delta t = 0$. Figure 1(b) depicts a scatter plot showing the log-transformed containment measure $\log g(u_i, u_j)$ as a function of the time difference between two crawls for every crawl pair.

For the GAW, the observed decay is in excellent agreement with our model, as depicted in Fig. 2(a), where we have fitted Eq. 2 via ordinary least squares regression to the log-transformed containment values ($R^2 = 0.95$). In particular, we find an average decay rate of $\alpha \approx 0.73$ per year, which corresponds to approximately 27% annual URL turnover, consistent with prior studies of web persistence [23]. Most importantly, we estimate coverage as 0.46, i.e., an average of 46% of the crawlable academic web are captured. To demonstrate the validity of the analytical expression of the urn model, we compare Eq. 2 with stochastic simulations of the model in Fig. 2, which show perfect agreement.

Figure 3 examines whether coverage varies over time by fitting separate regressions for each crawl's comparisons with all subsequent crawls. Panel 3a depicts the individual regression lines and shows consistent decay behavior with approximately parallel slopes. Panel 3b presents the y-intercepts estimating the sampling fraction c for each starting crawl. This shows a rising trend over time, indicating that crawl coverage has improved over the observation period. This improvement in coverage is either driven by an increase in crawl efficiency or a reduction in the crawlable population size N.

DISCUSSION

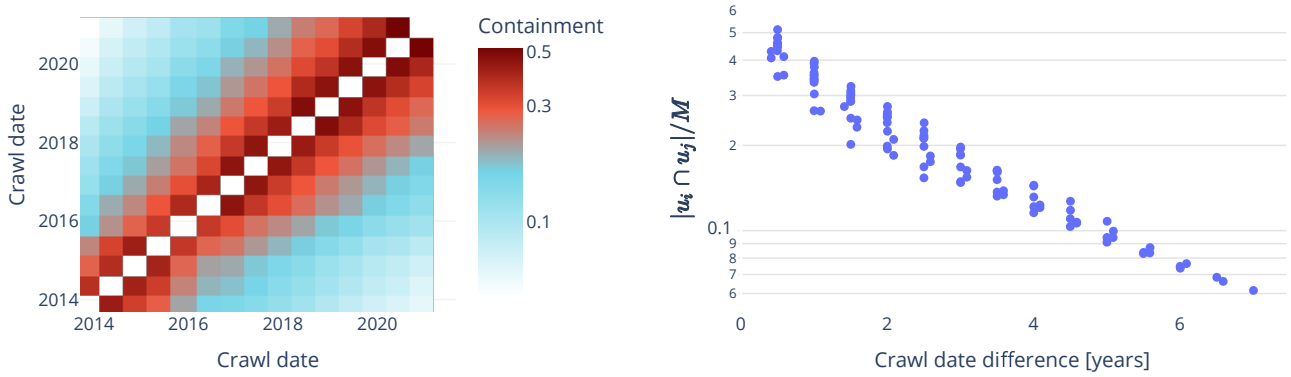
We presented the first method to estimate absolute web crawl coverage from longitudinal data alone, i.e., from a series of subsequent crawls. By analyzing pairwise URL set intersections over time across crawls we were able to estimate the coverage of the crawls. A simple urn model reproduces key features of the empirical data and therefore provides theoretical grounding for our ability to estimate coverage from the longitudinal GAW crawl data.

[2] Note that the hypergeometric (sampling without replacement) and binomial (independent Bernoulli trials) models give identical expected intersections. They differ only in variance—hypergeometric variance is smaller by the finite population correction $(N - M)/(N - 1)$. For point estimation, the models are equivalent.

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22925265

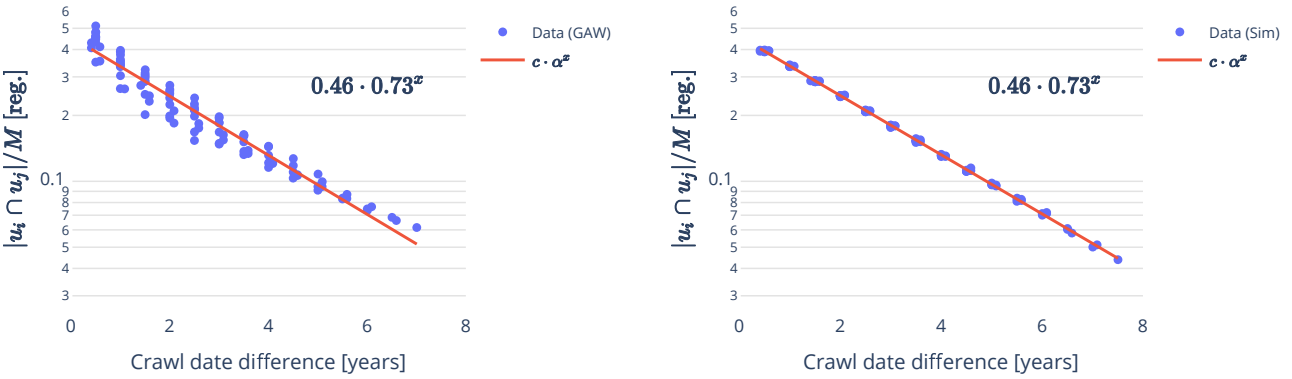
Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI



(a) Heatmap displays the containment between pairs of crawls.

(b) Compressed representation of the left panel onto Δt.

Figure 1: Pairwise URL set intersections for the 15 GAW crawls. (a) The heatmap reveals temporal decay structure; the diagonal shows estimated self-intersection values. (b) Log-transformed containment vs. time difference exhibits a linear relationship.



(a) Empirical GAW data.

(b) Simulated urn model data.

Figure 2: Validation via simulation. (a) Regression on GAW data. (b) Simulation with urn model parameters recovers the same relationship, confirming the method.

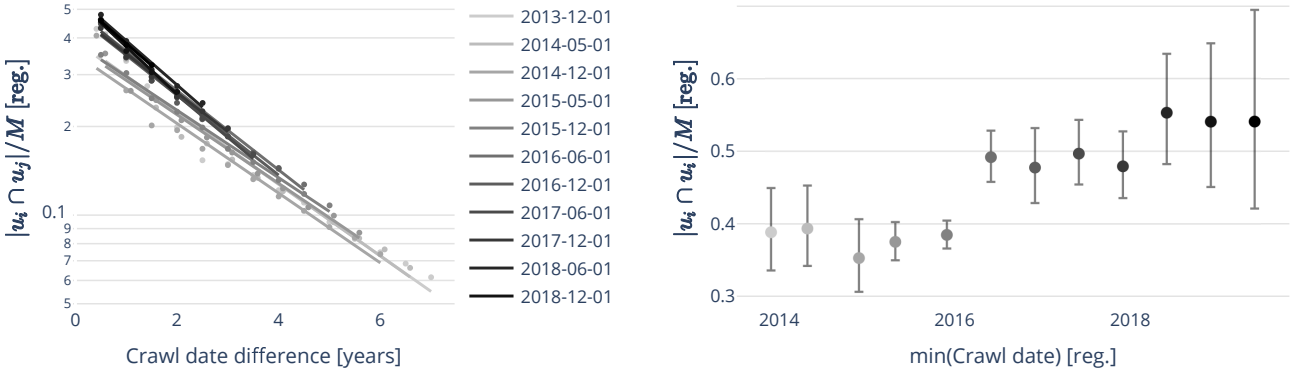
The key insight is that temporal variation in a longitudinal archive of web crawls serves as a substitute for the multiple independent samples required by traditional capture-recapture methods. In particular, by observing how URL sets overlap across time, the urn model allows us to decompose the (i) coverage (y-intercept of the fitted log relationship c, i.e. what fraction of the total we capture), and the (ii) decay component α, slope of the log relationship, i.e., how quickly content turns over. As a case study, we use our method to estimate our coverage of the German Academic Web. We find that each crawl captures approximately 46% of the complete crawlable academic web, with content persistence of approximately 73% per year (27% annual URL turnover).

Our work extends Bharat and Broder [1] from cross-sectional comparison of search engine indices to longitudinal analysis of crawled URL sets. Where they probed multiple search engines through query interfaces and could only compute relative sizes, we extract absolute coverage from direct temporal self-intersections within a single crawl archive. Furthermore, our extracted decay parameter $\alpha \approx 0.73$ per year—representing the annual URL persistence rate—implies a half-life of approximately 2.2 years for URL survival, closely aligning with Koehler's [23] longitudinal finding of approximately 2-year half-lives for web page persistence.

The method enables ongoing assessment of crawl coverage without external resources. Specifically, it provides archive operators a self-contained method to assess coverage without knowledge of external ground truth datasets, content extraction or entity matching, and most importantly, without the ability to compare to other crawlers. Researchers may benefit by being able to estimate what fraction of the target web it captures, informing confidence in downstream tasks, i.e., training LLMs. In addition, the temporal stability analysis (Fig. 3b in Section) reveals whether the target population is in a "replacement regime" (stable size) or "growth regime" (expanding), guiding crawl frequency decisions.

The way a crawl is modeled here is independent of the type of flux in the URL space; the flux of the web only affects how we may fit the overlap data. Corresponding the y-intercept to the crawl size normalized self-intersection is independent of the flux. If the cGAW grows over time, the sampling fraction $c_t = M/N_t$ decreases. We would observe

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22925265



(a) Individual regression lines per starting crawl.

(b) Estimated c vs. starting crawl date.

Figure 3: Per-crawl regression analysis. (a) Separate regressions for each starting crawl yield approximately parallel lines. (b) The y-intercept (estimated c) shows a rising trend over 2013–2021.

the coverage estimate declining. Conversely, if N_t shrinks, coverage estimates would increase. The coverage estimate across starting years provides an internal test for dynamics of N_t. In the case of the GAW, with a non-revisit policy, the web is crawled outwards, exploring further regions without resampling previously seen URLs. The sampling bias is accounted for through the log residual, providing a measure for an intersection's deviation from the urn model.

Persistence Cohorts and the Containment Spectrum

For the GAW, the containment curve on a log-y scale is approximately linear, justifying the single-urn fit. However, when the same method is applied to broader crawls such as Common Crawl, the curve should be *concave*: steeper at small Δt and shallower at large Δt. This concavity reflects the fact that the web is not homogeneous—some URLs persist for days, others for decades.

The concavity can be interpreted as the superposition of multiple urn processes with different persistence rates. Define a *persistence cohort* of fraction p: the $M_p = p \cdot M$ URLs per crawl that survive at least some threshold (e.g., one year), drawn from an unknown population of N_p such URLs. Each cohort has its own sampling fraction $c_p = M_p/N_p$ and persistence rate α_p, simulating a separate urn. By fitting the model interval-wise—starting from the largest Δt values and working backwards—one can extract separate (c_p, α_p) pairs for each persistence class.

This decomposition yields a *persistence spectrum*: not just "the crawl covers c% of the web" but "c_{fast}% of fast-turnover URLs and c_{long}% of long-living URLs." For a crawl operator, this is a richer diagnostic with different operational implications per class. We leave the full treatment of cohort decomposition and its application to Common Crawl for future work.

Limitations

Like any inference based on indirect observation, our approach rests on a set of assumptions and data requirements that delimit its applicability. The method assumes crawls sample approximately uniformly from the cGAW. Real crawls exhibit biases: breadth-first traversal favors well-connected pages; scope restrictions exclude content outside seed domains. The coverage estimate should be interpreted as coverage weighted by the crawl configuration, not uniform coverage of all theoretically reachable URLs. The approach further requires multiple crawls over time. It cannot assess coverage of a single snapshot. Archives with a small number of crawls may have insufficient data for reliable inference of the crawl coverage. We also assume the cGAW size remains approximately constant between start and termination of a crawl. In rapidly growing domains (e.g., social media), this assumption may not hold, and the intercept interpretation would be confounded by population change. Finally, coverage is measured at the URL level, not content level. A URL may persist while its content changes substantially. Content-based similarity measures could complement URL intersection analysis.

Extensions include non-uniform sampling models, application to other longitudinal archives such as Common Crawl, and content-based similarity measures.

CONCLUSION

We introduced a self-contained method to estimate absolute web crawl coverage using longitudinal data alone, leveraging temporal self-intersections of URL sets as a substitute for independent samples. By modeling overlap dynamics with a simple urn process, we showed that the intercept of the empirically observed decay in containment directly recovers the sampling fraction, while the slope captures content turnover. Applied to the German Academic Web, the method indicates that individual crawls capture approximately 46% of the crawlable URL space, with an annual persistence rate of about 73%. Because the approach requires neither external ground truth nor comparison across crawlers, it provides archive operators and researchers with an interpretable and

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

https://doi.org/10.5281/zenodo.22925265

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

general tool for assessing crawl completeness in longitudinal web archives.

Data availability. Metadata for the German Academic Web (URLs and timestamps) is available at `https://german-academic-web.de/`.

ACKNOWLEDGEMENTS

The authors thank Robert Jäschke for conceiving and maintaining the German Academic Web archive and for providing access to the crawl data.

REFERENCES

[1] Krishna Bharat and Andrei Broder. A technique for measuring the relative size and overlap of public web search engines. *Computer Networks and ISDN systems*, 30(1-7):379–388, 1998.

[2] Dimitar Denev, Arturas Mazeika, Marc Spaniol, and Gerhard Weikum. The sharc framework for data quality in web archiving. *The VLDB Journal*, 20(2):183–207, 2011.

[3] Michael Paris and Robert Jäschke. How to assess the exhaustiveness of longitudinal web archives: A case study of the german academic web. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 85–89, 2020.

[4] C. G. J. Petersen. The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station to the Home Department*, 6:1–48, 1896.

[5] Frederick C Lincoln. Calculating waterfowl abundance on the basis of banding returns. *US Department of Agriculture Circular*, 118:1–4, 1930.

[6] Douglas G. Chapman. Some properties of the hypergeometric distribution with applications to zoological censuses. *University of California Publications on Statistics*, 1:131–160, 1951.

[7] G. A. F. Seber. *The Estimation of Animal Abundance and Related Parameters*. Edward Arnold, London, 2nd edition, 1982.

[8] Stephen E. Fienberg. The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, 59(3):591–603, 1972.

[9] Steve Lawrence and C. Lee Giles. Searching the World Wide Web. *Science*, 280(5360):98–100, 1998.

[10] Steve Lawrence and C. Lee Giles. Accessibility of information on the web. *Nature*, 400(6740):107–109, 1999.

[11] Adrian Dobra and Stephen E. Fienberg. How large is the World Wide Web? In Mark Levene and Alexandra Poulovassilis, editors, *Web Dynamics*, pages 23–45. Springer, 2004.

[12] Judit Bar-Ilan. On the overlap, the precision and estimated recall of search engines: A case study of the query "Erdos". *Scientometrics*, 42(2):207–228, 1998.

[13] Judit Bar-Ilan. Search engine results over time—a case study on search engine stability. *Cybermetrics*, 2/3(1):paper 1, 1999.

[14] Judit Bar-Ilan. Methods for measuring search engine performance over time. *Journal of the American Society for Information Science and Technology*, 53(4):308–319, 2002.

[15] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société vaudoise des sciences naturelles*, 37:547–579, 1901.

[16] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[17] Thorvald Julius Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. *Biologiske skrifter*, 5:1–34, 1948.

[18] Andrei Z Broder. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences*, pages 21–29. IEEE, 1997.

[19] Andrei Z. Broder, Moses Charikar, Alan M. Frieze, and Michael Mitzenmacher. Min-wise independent permutations. *Journal of Computer and System Sciences*, 60(3):630–659, 2000.

[20] Ankita Garg, Catherine G Enright, and Michael G Madden. On asymmetric similarity search. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 649–654. IEEE, 2015.

[21] Alex T Kalinka. The probability of drawing intersections: extending the hypergeometric distribution. *arXiv preprint arXiv:1305.0717*, 2013.

[22] Hosam M. Mahmoud. *Pólya Urn Models*. CRC Press, 2008.

[23] Wallace Koehler. Web page change and persistence—a four-year longitudinal study. *Journal of the American Society for Information Science and Technology*, 53(2):162–171, 2002.

[24] Wallace Koehler et al. A longitudinal study of web pages continued: a consideration of document persistence. *Information Research*, 9(2):9–2, 2004.

[25] Daniel Gomes and Mário J Silva. Modelling information persistence on the web. In *Proceedings of the 6th International Conference on Web Engineering*, pages 193–200, 2006.

[26] Dennis Fetterly, Mark Manasse, Marc Najork, and Janet Wiener. A large-scale study of the evolution of web pages. *Software: Practice and Experience*, 34(2):213–237, 2003.

[27] Jonathan D Wren. Url decay in medline—a 4-year follow-up study. *Bioinformatics*, 24(11):1381–1385, 2008.

[28] Justin F Brunelle, Mat Kelly, Hany SalahEldeen, Michele C Weigle, and Michael L Nelson. Not all mementos are created equal: measuring the impact of missing resources. In *International Journal on Digital Libraries*, volume 16, pages 283–301, 2015.

[29] Scott G Ainsworth, Michael L Nelson, and Herbert Van de Sompel. Only one out of five archived web pages existed as presented. In *Proceedings of the 26th ACM Conference on Hypertext and Social Media*, pages 257–266, 2015.

[30] Stefan Baack. A critical analysis of the largest source for generative ai training data: Common crawl. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2199–2208, 2024.

[31] Internet Archive. Heritrix – The Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project. `https://github.com/internetarchive/heritrix3`, 2020.

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22922310

A FOCUSED CRAWLING STRATEGY TO ENHANCE GEOSCIENCE DATA RETRIEVAL

R. Frühwirth[1], T. Hecking [2], A. Nussbaumer[1], C. Gütl [1]
[1]Graz University of Technology, Graz, Austria
[2]German Aerospace Center: Cologne, Germany

Abstract

This paper presents a lightweight approach for web crawling focussing on particular themes. Instead of supervised models for website classification the system evaluates the relevance of web resources based on similarity with a small set of reference documents based on embeddings produced by a pre-trained language model to guide exploration. Initial evaluations demonstrate that our approach identifies relevant web documents significantly faster than baseline methods while maintaining minimal resource consumption. By enabling rapid, data-efficient crawling tailored to dynamic requirements, this work advances a practical paradigm for resource-aware real-time web analysis with immediate utility for ad-hoc monitoring scenarios, including crisis response and event-driven information gathering.

INTRODUCTION

Web crawling is an essential building block not only for building search indexes for web search engines, but also for web analytics, and the creation of training data for AI models. Traditional general-purpose web crawlers operate with a greedy paradigm, systematically discovering and indexing pages across broad sections of the web. However, for numerous information retrieval and monitoring applications, such exhaustive exploration is neither practical nor necessary. Use cases frequently require only a highly specific fraction of the web: language-specific archives (e.g., Frisian-language content), domain-specific monitoring (e.g., academic physics literature or regulatory documents), or services demanding frequent updates. Focused crawlers address this inefficiency by deliberately prioritizing the crawl frontier and managing hyperlink exploration to target pages satisfying a defined property. This selective strategy yields substantial operational advantages. By filtering out irrelevant content early, focused crawlers drastically reduce network traffic, storage overhead, and computational demands, often enabling deployment on commodity hardware rather than specialized infrastructure. Furthermore, concentrating crawl resources on a curated set of highly reliable information sources allows for shorter crawl intervals, ensuring that collected document collections remain both high-quality and timely.

Despite these well-documented benefits, constructing focused crawlers for highly specialized or niche topics remains a persistent challenge. Conventional approaches typically rely on supervised classification models that require extensive, manually labelled training corpora. Such resources are frequently unavailable for emerging, narrow, or linguistically underserved domains. Compounding this limitation, real-world information ecosystems are inherently dynamic; topic boundaries shift, new authoritative sources emerge, and crawling targets must be continuously adapted without interrupting the crawl process. Existing systems often lack the agility to reconfigure targeting criteria rapidly, creating a gap between the theoretical promise of focused crawling and its practical deployment in time-sensitive, data-scarce environments.

To address these challenges, we propose a lightweight focused crawling framework that operates effectively without large-scale labelled training data. Our core idea centers on leveraging unsupervised classification techniques guided by a minimal reference corpus of a few seed documents. By computing semantic and structural similarities between frontier candidates and the reference set, the crawler dynamically prioritizes URLs that align with the target property, enabling rapid adaptation to novel or evolving topics. An initial evaluation shows that our approach is indeed capable to locate relevant web documents faster than baseline methods. Furthermore, we outline a practical application of the approach to ad-hoc discovery of web information during natural disasters.

RELATED WORK

Open Web Index and Search

To overcome the issue of closed web data infrastructures by large commercial search engines contributing to information infrastructure sovereignty, the Open Web Search (OWS) initiative aims for an open web search ecosystem. To give private, public and commercial stakeholders maximum freedom, six guiding principles for an Open Web Index (OWS) are followed: (1) distributed as open data; (2) transparent and extensible index creation; (3) collaborative creation, (4) tracking content usage while privacy protection; (5) usage control by content owners; (6) legal compliance along the date processing pipeline. [1]

The simplified conceptual architecture of the open web search initiative follows the search engine concept taking into account above stated principles. Accessible websites are indexed by the open web crawler OWLer, a distributed system, and stored as WARC (Web Archive) files. This enables also third parties to contribute web data in the same format. The collected WARC files are pre-processed by the RESILIPARSE library, content and metadata extracted as well as enriched with further metadata. Parquet file format, a column-oriented file format for efficient data storage and retrieval, is used for storing the pre-processed content and metadata. This is followed by index creation using the Com-

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22922310

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

mon Index File Format (CIFF), a standard format for storing and exchanging index data. Both, Parquet and CIFF representation of the web content files can be downloaded and used in various applications, such as domain specific search applications. [2][3]

Narrowing down to the crawling pipeline, the OWler is a distributed web crawler which is an extension of the StormCrawler running in Apache Storm distributed over date centers. It is designed to run in two cluster tiers: (1) singleton tier coordinating the multi-tier crawling process, managing the URL crawling list and statistics; (2) crawling cluster tier accessing the URL resources and creating the WARC files. [3][4]

Focused Thematic Crawling

The decentralized and highly dynamic nature of the World Wide Web (WWW) requires web crawlers to uncover the web graph and keep information updated. Because of its design, fast emerging technologies, commercialization of the web and security threats, challenges include (1) finding proper seed URLs, (2) downloading and extracting content including hyperlinks, (3) dealing with complex dynamic content, (4) dealing with personalized content, AI-generated content and data pollution, (5) crawler traps, and (6) recrawling and keeping the index fresh. Specifically, the latter requires strategies such as highly parallelization or prioritization for high-quality, popular fast changing web sites. [5][6][7]

Another prioritization strategy is focused crawling or topical crawling which selectively targets web pages relevant to a pre-defined theme, avoiding the rest of the internet.

Based on approaches and mechanisms to determine relevance and guiding the path, they can be classified into the following groups: (1) Keyword-driven crawlers build their selection process based on keyword filters, which is also seen as the simplest and most traditional approach. (2) In a more advanced approach, semantic and ontology-based crawlers utilize domain-specific knowledge representations by concept matching and metadata abstraction. (3) Classification-based crawlers use machine learning classifiers for applying relevance scoring. (4) Adaptive and learning-based crawlers learn dynamically by exploring and processing the web space and refining their strategy to maximize its harvest rate applying reinforcement learning and genetic algorithms. (5) Link structure and context graph crawlers analyze the hyperlink topology and the neighborhood of a web page to identify content most central to a community of interest. [8][9][10]

REQUIREMENTS AND DESIGN

The main motivation for our focussed crawler is to enable fast and resource efficient on-demand crawling of very specialized topics. As in other focussed crawler one requirement is to reduce resource consumption in terms of network traffic and storage capacity by only retrieving and storing relevant information. Focus on a few but highly reliable information sources enables maintaining task-specific high-quality and timely document collections on commodity hardware.

However, most focussed crawlers rely on supervised classification models to assess page relevance, which relies on the availability of a sufficient amount of training data and limits flexibility because one model has to be trained for each topic of interest. A requirement for our crawler is to overcome this limitation allowing to perform ad-hoc focussed crawls of very specialized niche topics with little effort.

To this end, the general design idea behind our solution is that a user only has to give a few examples of texts of a topic of interest and the crawler takes this as a basis to find semantically similar documents in the web.

The approach is straight forward and depicted in Figure 1. Provided documents of interest (e.g. paper abstracts on a certain scientific topic) form a reference corpus. By using a pre-trained document embedding model the reference corpus is transformed into a point cloud in the embedding vector space. If the reference documents are thematically similar enough one can assume that the resulting point cloud is coherent in the sense that the reference document embeddings are closer to each other than to embeddings of irrelevant documents. In order to decide if an unknown document is relevant given the reference corpus, we check if its embedding is close enough to the reference point cloud or if constitutes an outlier. This decision is based on the well-known Local Outlier Factor [11] (LOF).

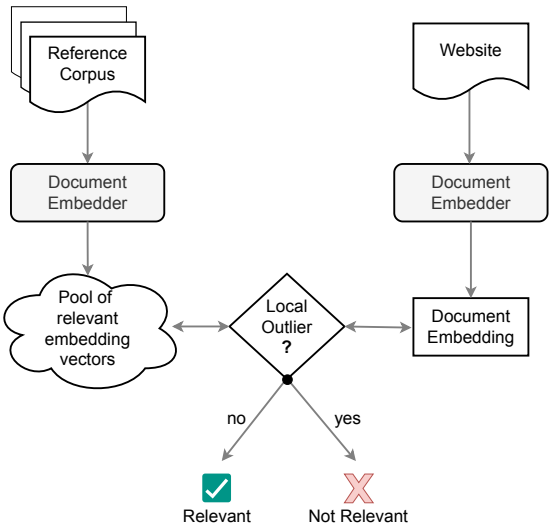


Figure 1: Unsupervised method to decide if a website is relevant based on Local Outlier Factor (LOF)

$LOF_k(x)$ is the fraction of the local density (similarity) of a data point x given its k nearest neighbours and the average local densities of those neighbours. A $LOF_k(x) > 1$ indicates that the reference embeddings form a more dense region than x and its k nearest neighbours, and thus, x can be considered as an outlier. In practice the decision threshold can be adjusted depending on how tight the decision boundary should be.

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22922310

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

Through this lightweight approach focussed crawls of highly specialized topics can be performed on-demand without the need for costly training of classification models.

FOCUSED CRAWLING PROTOTYPE

The focused crawling prototype consists of two complementary systems: a Domain Embedding System (DES) and a Focused Crawling Pipeline that extends the OWLer crawler. The DES automates the construction of a domain-specific reference corpus and handles the selection of high-quality seed URLS, which guide the subsequent crawl. The Focused Crawling Pipeline integrates into OWLer's StormCrawler-based topology and steers the crawl toward relevant content by comparing paragraph-level BERT [12] embeddings against the reference corpus using LOF novelty detection as described in the last section. A design choice is the use of a general-purpose, pre-trained BERT model (bert-base-uncased) without domain-specific fine-tuning. Rather than requiring labeled training data for each target domain, LOF learns the distribution of the reference embeddings and flags content as irrelevant, which makes the approach applicable across domains.

Domain Embedding System

The DES produces the two inputs the pipeline depends on, the reference corpus and the seed URLs in three stages. In the URL gathering stage, the system queries multiple search engines (Google, Bing) and scientific repositories (OpenAlex, DOAJ) for the user-specified domain and filters the returned URLs by the robots directives and MIME type, which results in an initial set of crawlable, domain-related pages. The link analysis stage construct a web graph over these URLs and applies the HITS [13] algorithm, which assigns hub and authority scores to every node. Only high-hub and high-authority URLs are kept, as they provide high-quality entry points into the target domain and are exported as the seed list for the focused crawl. In the page embedding stage, each selected page is downloaded, cleaned and segmented into logical content blocks using Boilerpipe [14] and every block is embedded with BERT. The resulting embeddings form the domain-specific reference corpus against which crawled content is later compared.

Focused Crawling Pipeline

The Focused Crawling Pipeline extends OWLer's StormCrawler topology with four additional bolts while reusing the existing fetching, parsing, status updating and WARC storage components. A page segmentation bolt applies Boilerpipe to strip boilerplate from the HTML such as navigation bars and advertisements and to divide each page into several logical content blocks which enables relevance assessment at block rather than page granularity. This distinction matters on large pages where only individual sections are on-topic. An embedding bolt converts each block into a 768-dimensional vector using bert-base-uncased with mean pooling over token embeddings. To execute BERT inside the Java-based crawler, the model is exported to the ONNX format and run via the ONNXRuntime. A classification bolt then scores each block embedding with the LOF mode fitted on the reference corpus. Since no efficient Java implementation of LOF exists, the classifier runs as an external Python service exposing a Flask-based API around scikit-learn's LOF model. The bolt submits the block embeddings and receives an outlier score and a binary relevance label per block. Finally, a relevance evaluation bolt aggregates the block-level results into page and link priorities that steer the URL frontier. Pages that rely on client-side JavaScript are rendered with Selenium in headless mode before segmentation, this way no content is lost.

Link prioritization is based on a relevance propagation scheme inspired by Shark Search. The priority of a discovered link combines the relevance of its parent page, the relevance of the content block containing the link and the relevance of the anchor text, which are weighted equally at 0.33 each. The combined score is mapped onto the frontiers scheduling so that links from relevant contexts are crawled first. Because relevant regions of the web graph are not always connected, the pipeline additionally implements a tunnelling mechanism which allows the crawler to traverse sequences of irrelevant pages to reach isolated relevant components. Tunnelling is bounded by a depth limit that is decremented whenever the LOF model flags a page as an outlier. The combination of block level relevance propagation and depth-limited tunnelling lets the crawler exploit relevant neighbourhoods aggressively while also having the ability to escape locally irrelevant regions.

EVALUATION AND SAMPLE APPLICATIONS

Evaluation

The crawler is evaluated in two phases: a document similarity evaluation to select the classification approach and a focused crawl evaluation to measure end-to-end crawling performance.

The focused crawl evaluation runs on a single machine (Ubuntu 24.04, AMD Ryzen 5 3600X, 32 GB RAM, NVIDIA GTX 1070) with Docker providing isolated, reproducible environments. All crawls respect the robots.txt directive and are restricted to English-language content. To ensure reproducibility, crawling is restricted to a controlled web graph of approximately 130.000 unique URLs containing 700 manually verified Earth Observation target pages. The ground truth is built by scanning every page for domain keywords, matched pages are sent to OpenAI API resulting in roughly 2.000 candidates and are manually reviewed which yields the final 700 targets. A custom OWLer URL Filter prevents navigation outside of the web graph. Ten randomly sampled seed sets of 25 URLs each are used to initialize the crawls and every crawl stops after 1.000 fetched URLs. Target recall is the primary metric which measures the fraction of all target URLs found. The secondary metric is domain diversity, which measures how many unique domains the

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22922310

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

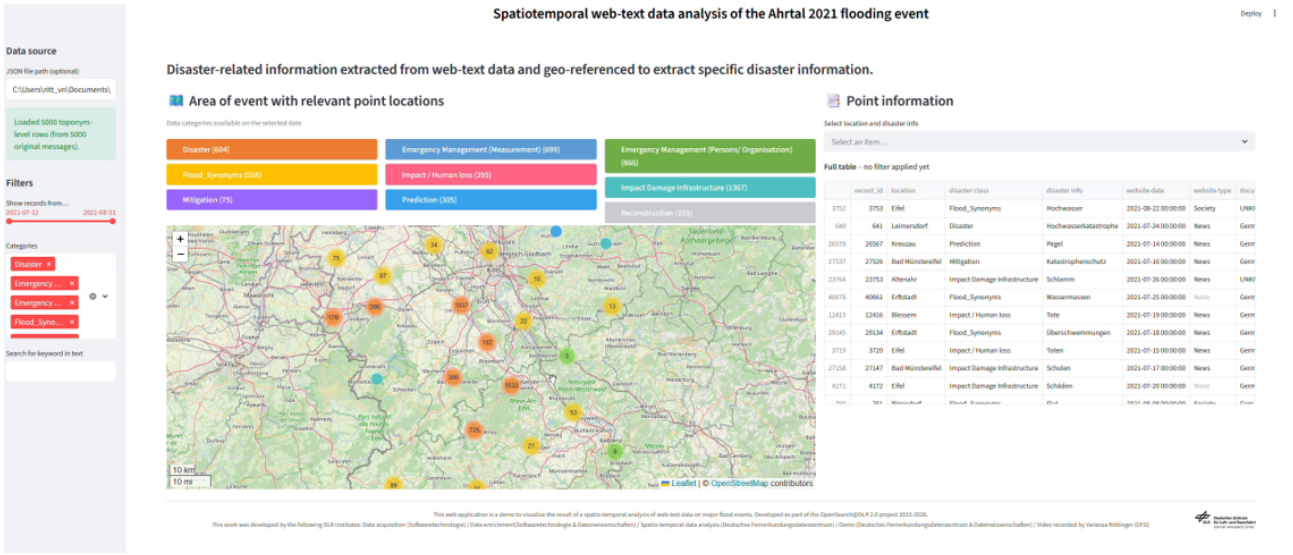


Figure 2: Dashboard mapping georeferenced web documents on a flood event

crawler visited. Three approaches are compared: a TF-IDF [15] clustering baseline, the BERT embedding clustering approach of the pipeline and a supervised, fine-tuned BERT classification variant. dsew Table 1 reports the average number of targets found per seed set. BERT clustering discovers about 30 targets on average and is the most consistent across seed sets, a 58 % improvement over the TF-IDF baseline (19). BERT classification averages 27 targets but produces the single best run, discovering over 50 targets on one seed set. All approaches perform near-identically over the first 200 URLs, the gaps widen as the crawls progress deeper into the graph. Notably, TF-IDF discovers the most unique domains (138, against 87 for BERT clustering and 82 for BERT classification) while finding the fewest targets, which indicates that keyword-based approaches explore more broadly but less effectively than semantic approaches, which concentrate on relevant clusters within fewer domains. The main limitations are the computational overhead of BERT embedding generation, which constrains its scalability.

Table 1: Average number of targets found per seed set after 200, 500 and 1,000 crawled URLs.

Approach	200	500	1,000
BERT Clustering	5	18	30
BERT Classification	5	15	27
TF-IDF Baseline	4	10	19

Example Application

To further highlight the utility of our approach, we showcase an example application that utilized ad-hoc focussed crawls in the case of extreme weather events. Figure 2 shows a dashboard depicting localized web information on the catastrophic flood event that took place in 2021 in the Ahrweiler region in Germany.[1] Data for this visualization is fed in by focused crawls initiated with a corresponding reference corpus. This example shows that our focused crawler in combination with geotagging of web resources can be used to gather ad-hoc information in real time from the web during crisis events to inform authorities and emergency services, or to support analyses in the aftermath of the event with high-quality web datasets.

CONCLUSION AND FUTURE WORK

The initial results presented in this work demonstrate that an unsupervised, reference corpus driven prioritization strategy can effectively guide a focused crawler toward application specific target content without relying on costly training of topic-specific classification models. This work thus constitutes a step towards resource efficient real-time web analysis which can be especially valuable for use cases where dynamic web information has to be gathered quickly and on demand, for example, during crisis events.

In order to learn more about proper parameter configurations (e.g. similarity thresholds), best practices, and limitations in different scenarios, the evaluation has yet to be expanded including experiments in a broad spectrum of application domains, ranging from stable academic disciplines to highly dynamic, rapidly evolving domains. This includes also investigating how topic drift, lexical sparsity, and domain-specific structural patterns influence classification accuracy.

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).

[1] 2021 European floods - Wikipedia

https://doi.org/10.5281/zenodo.22922310

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

REFERENCES

[1] 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, 2023. 10.1002/asi.24818

[2] M. Granitzer et al., "Openwebsearch.eu - building an open web index on eurohpc ju infrastructures," *Procedia Computer Science*, vol. 255, pp. 43–52, 2025, Proceedings of the Second EuroHPC user day. https://doi.org/10.1016/j.procs.2025.02.259

[3] Open Web Search Community, *Creating the owi*, https://openwebindex.eu/book/content/dnt/pipeline.html, Last accessed on 2026-05-04, 2026.

[4] Open Web Search Community, *Crawling*, https://openwebindex.eu/book/content/dnt/crawling.html, Last accessed on 2026-05-04, 2026.

[5] S. M. Mirtaheri, M. E. Dinçtürk, S. Hooshmand, G. V. Bochmann, G.-V. Jourdan, and I. V. Onut, "A brief history of web crawlers," in *Proceedings of the 2013 Conference of the Center for Advanced Studies on Collaborative Research*, 2013, pp. 40–54.

[6] Z. Chang, "A survey of modern crawler methods," in *Proceedings of the 6th International Conference on Control Engineering and Artificial Intelligence*, 2022, pp. 21–28. 10.1145/3522749.3523076

[7] C. Olston and M. Najork, "Web crawling," *Found. Trends Inf. Retr.*, vol. 4, no. 3, pp. 175–246, 2010. 10.1561/1500000017

[8] S. Chakrabarti, M. van den Berg, and B. Dom, "Focused crawling: A new approach to topic-specific web resource discovery," *Computer Networks*, vol. 31, no. 11, pp. 1623–1640, 1999. https://doi.org/10.1016/S1389-1286(99)00052-3

[9] M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles, and M. Gori, "Focused crawling using context graphs," in *Proceedings of the 26th International Conference on Very Large Data Bases*, 2000, pp. 527–534.

[10] J. Liu, X. Li, Q. Zhang, and G. Zhong, "A novel focused crawler combining web space evolution and domain ontology," *Know.-Based Syst.*, vol. 243, no. C, 2022. 10.1016/j.knosys.2022.108495

[11] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," *SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, 2000. 10.1145/335191.335388

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186. 10.18653/v1/N19-1423

[13] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, 1999. 10.1145/324133.324140

[14] C. Kohlschütter, P. Fankhauser, and W. Nejdl, "Boilerplate detection using shallow text features," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 2010, pp. 441–450. 10.1145/1718487.1718542

[15] K. Spärck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22915695

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

AN ETHICS-BY-DESIGN APPROACH FOR THE OPEN WEB INDEX

A. Nussbaumer[1], C. Plote[2], R. Ridgway[3], N. Pöchhacker[4,5], M. Granitzer[6], M. Dinzinger[6], S. Zerhoudi[6], S. Heineking[7], M. Potthast[7,8,9]

[1]Graz University of Technology, Austria
[2]Open Search Foundation e. V., Starnberg, Germany
[3]Digital x Data Research Centre, London, UK
[4]University of Graz, Austria
[5]University of Vienna, Austria
[6]University of Passau, Germany
[7]University of Kassel, Germany
[8]hessian.AI, Darmstadt, Germany
[9]ScaDS.AI, Leipzig, Germany

Abstract

This paper presents an ethics-by-design approach and methodology applied to the Open Web Index (OWI). This methodology consists of a stakeholder analysis, a collection of potential issues, ethical requirements elicitation, and an elaboration of technical and organisational measures. It results in an ethics framework consisting of relevant stakeholders, requirements, and measures to mitigate ethical risks and problems. The elaboration of this framework has been conducted in parallel with the technical design and development, which allowed the ethics framework and the technical design to mutually influence each other.

INTRODUCTION

Digital information search systems are key infrastructures of today's information age [1, 2]. They influence access to knowledge, public discourse, and political participation. Yet their development is mostly driven by commercial interests and opaque decision-making. The most popular of such systems are "black boxes" [3], generating profit and power while denying users insight into their inner workings and their logic. At the same time, these search systems extract data from users' digital traces and exploit them for their own benefit [4]. Besides their problematic treatment of user data, there are further ethical problems related to the handling of web data. Most importantly, they are functioning as a kind of gatekeeper for users searching and accessing web data.

In order to overcome these problems, the OpenWebSearch.eu project has established an Open Web Index (OWI) that provides free and open access to web data and stimulates the creation of an ecosystem of applications and data products using these web data. The technology of the OWI consists of three core processes, which are (a) crawling web data, (b) pre-processing and analysing web data, and (c) building a shareable index [5]. While the overall approach has the potential to solve some ethical problems, such as market dominance and the lack of transparency, there is still a need for a careful design of the technology informed by ethical requirements.

This paper presents an ethics-by-design approach and methodology that informed the technical development of the OWI, which was developed during the OpenWebSearch.eu project [6]. This methodology consists of a stakeholder analysis, a collection of potential issues, ethical requirements elicitation, and an elaboration of technical and organisational measures. These steps have been undertaken in parallel with the technical design and development, where they mutually influenced each other. The outcome of these steps forms the ethics framework and is presented in the next sections.

ETHICS-BY-DESIGN METHODOLOGY

This section describes an ethics-by-design methodology that guides the integration of ethical requirements with the design and development of the OWI. This methodology is based on the Value-Sensitive Design (VSD), which aims to influence the technological design by explicitly identifying human values to be considered and integrated throughout the design process [7, 8]. The formalised VSD methodology consists of a three-part iterative process, which includes the investigation of conceptual, empirical and technical issues specific to a particular design. An application of this approach and its integration in the research process is described in [9], where the procedure of informing the technical design by ethical requirements and values is outlined.

Our ethics-by-design methodology (see Figure 1) consists of a structured step-by-step approach that is integrated in the design and development phase of the OWI. In the first step, a stakeholder analysis is conducted that aims to identify all people and user groups that are involved in the development of the OWI, that make use of the OWI, and that are affected by the operation of the OWI. In the second step, a general issue collection is undertaken by analysing the positive and negative impact of the OWI on the different stakeholder groups. In this step a conceptual design of the technology is used for the analysis, in order to identify problems before the technology development has started. Based on the collected issues, in step three a detailed analysis of concrete risks is undertaken along with setting up a list of values as

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22915695

an ethical foundation. Together they form the ethical requirements. Finally, technical and organisational measures are elaborated that provide concrete guidance on how to adhere to the ethical requirements. These measures are a high-level description of how the requirements can be fulfilled with either technology or organisation. These measures concern the development of the technology, the use and sharing of the technology and data, and third-party users of the index.

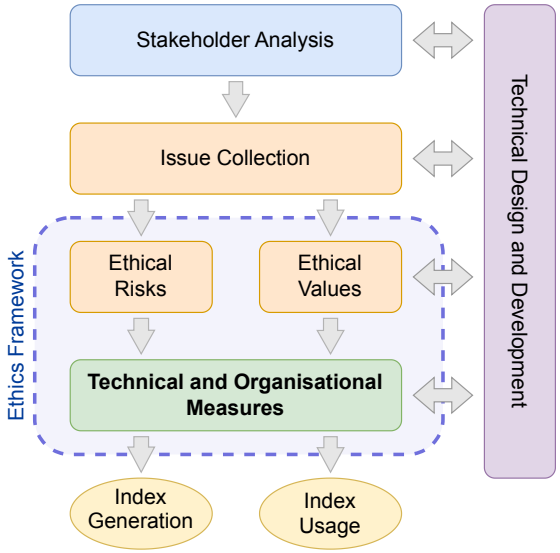


Figure 1: Ethics-by-Design Approach

This methodology results in an ethics framework that consists of concrete instantiations of ethical requirements (ethical risks and values) and measures to address these requirements. While the ethics-by-design methodology describes the overall process to elaborate the ethics framework in the context of the technical design and development, the ethics framework includes concrete information to inform the technical design and development. An important aspect is the co-design character of both the ethics framework creation and the technical development. They have been performed over the whole project lifetime mutually influencing each other.

STAKEHOLDER ANALYSIS

This section lists the stakeholders of an Open Web Index from an ethical perspective. The list focuses on three different types of stakeholders. The first consists of those stakeholders who make technical or organisational decisions. The second consists of those who are impacted by the decisions and technology in positive or negative ways. Finally, the third consists of stakeholders with a governance role. The detailed list below was used to analyse potential risks and harm, as well as required values and their enforcement.

Technology and index developers. This user group includes people who contribute to the creation and operation of the index, such as developers, data centres, or data curators.

Index users. This group consists of those who use the OWI in some way, such as developers creating search applications or data products, researchers analysing web data, or organisations maintaining products created with the OWI.

Regular end users. These are end users of the search applications or data products created with the OWI, who simply consume the various services with a corresponding user interface.

Content providers and content creators. This group includes people and organisations that create and publish content on the web.

People mentioned on the web. This group includes people whose data, such as name, address, or activities, are published on the web.

Contextual researchers and institutions. This group includes researchers and institutions on cross-cutting issues, for example researchers on legal, ethical or social aspects, educational institutions, or communication departments.

Governance and management. This group takes care for compliance of ethical and legal standards or for adapting these standards, such as policy makers.

ETHICAL REQUIREMENTS

This section lists and clusters ethical risks and values related to the OWI, whereby each cluster has multiple facets and relations to each other. The compilation of the risks and values form the ethical requirements.

Ethical Risks

An ethical risk is the chance that a decision, action, or policy will cause harm, break moral principles, or damage trust and integrity. In the context of the OWI, we identified hypothetical scenarios where stakeholders are harmed. These scenarios are the basis for informing the design, in order to avoid or mitigate them.

1) Control of indexed websites. The question of which websites should and should not be included in the index is a key issue that has a major impact on the quality of the index, on people and on society, and the risk of (perceived) censorship comes into play. The question of who should decide about removal is crucial. The following cases are particularly relevant: (a) Website owners don't want their sites included in the index for any reason. (b) Website owners want to have their sites included, but they are missing. (c) Websites containing information loosely related to people are indexed against the will of the people affected. (d) Websites dealing with aspects of social minorities are missing from the index. (e) Websites from questionable sources or with problematic content are indexed. (f) Overall, the index contains biased information or information that lacks plurality.

2) Flawed content in the OWI. The quality of content has a major impact on the information retrieved from the index. For example, the following issues may occur. (a) Websites may contain false, misleading, or manipulative content, as

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22915695

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

well as unreliable AI-generated content and disinformation that are disseminated on constantly growing content farms. (b) Websites may contain information that offends other people or institutions. (c) Websites contain information that some people do not want to be published.

3) Inaccurate metadata in the OWI. The quality of metadata affects the further processing of search applications or data products created by third parties. This can reduce trust in the OWI and in alternative search engines. This risk is particularly likely to arise in the following situations: (a) Original metadata in web pages is incorrect or not adequate for the content of a page. (b) Original metadata from web pages is not included correctly in the index. (c) Metadata created in the pre-processing steps is incorrect or misleading.

4) Misuse of the OWI by third parties. Third parties using the OWI, particularly application developers or creators of data products, may not adhere to the required ethical practices and standards. Examples are: (a) Third parties use the index data to analyse, extract, or modify information that is used in a way that is unjust or harmful to individuals or to society. (b) Third parties create applications that do not respect ethical or legal guidelines, such as violating privacy and data protection rules.

5) Unsuitable operation and governance of the OWI. Problematic practices in relation to the technical administration and maintenance of the OWI can cause the following issues: (a) Management and control structure of the OWI operation might be unsuitable and cause problems related to the openness and unbiased OWI data. (b) Security breaches might lead to unwanted changes in the OWI data or cyber attacks might harm the functionality of the OWI. (c) Government, corporate, or commercial influence can infringe on the unbiased operation of the technical processes.

Ethical Values

An ethical value is a core principle or standard that can be used as a moral compass to decide what is right, fair, and good. The following list contains the key ethical values considered most relevant to the development and operation of an Open Web Index. They form the basis and key requirements to inform the design and development of the OWI. Please note that most of the values are closely related, and the order does not necessarily represent a prioritisation.

1) Transparency. As an ethical value that spans science, engineering, business and the humanities, transparency is the act of making it easy for others to see what actions are being or have been performed. Transparency implies openness, communication and accountability. Transparency also includes the concepts of explainability, observability, understandability and controllability. *Relevance for OWI:* Transparency in the OWI context refers to the data and algorithms available in the OWI and related applications. Metadata, data analysis, algorithmic decisions and the overall approach should be made transparent to users in a way so that they can easily understand the matter, without causing cognitive overload (e.g., due to excessive technical data). Furthermore, users should have access to the data and algorithmic operations, making them as accessible and understandable to the user as possible. *Related concepts*: Explainability, Understandability, Controllability, Interoperability, Sustainability, Social Responsibility

2) Safety and Security. Security refers to protection or resilience against potential harm, hostile forces or unwanted coercion. At the core of information security is information assurance, the act of maintaining the confidentiality, integrity, and availability of information, ensuring that information is not compromised in any way when critical issues arise. *Relevance for OWI:* The crawling, processing, storage, and sharing must adhere to the basic principles of security, data integrity, confidentiality and availability. This includes measures to ensure that the search infrastructure is resilient against manipulation and misuse. *Related concepts*: Responsibility, Trustworthiness, Justice, Accountability

3) Openness. Openness is a virtue of people or organisations that stimulates progress and growth, as opposed to closed-mindedness, narrowness and rigidity. Openness can be practised in at least three directions. First, openness is people-centered and receptive to others, inclusive rather than exclusive and welcoming diversity. It values (not just "tolerates") others and seeks to discover the gifts and talents of those others. The second direction is toward ideas (intellectual openness), which includes openness to creativity, innovation and novelty. Thirdly, openness is directed toward criticism. In addition to criticism on technical or economic grounds, ethical criticism asks whether a technology or business decision is good or bad for people or organisations. *Relevance for OWI:* The concept of openness in the context of OWI refers to receptiveness to the integration of new concepts and technical contributions, but also to the use and exploitation of the data and services provided by the OWI. It addresses a wide range of different types of stakeholders located globally. Openness includes also the acceptance of ethical criticism of software, concepts and services developed. Hence, openness is also relevant to i) establishing the Open Web Index as a critical infrastructure and ii) building an ecosystem around it. *Related concepts:* Interoperability, Accessibility, Transparency

4) Responsibility. Responsibility refers to both the duty and obligation to take action to avoid harm. The actions should be taken in advance, such as designing and planning processes that avoid damages or that are taken if damages have occurred. *Relevance for OWI:* Responsibility is an important value when it comes to the design of technology or organisational processes. The technology should be designed in such a way that it mitigates disadvantages or harm of users. Processes should be designed that allow to respond accordingly if an unforeseen harm has occurred. *Related concepts:* Reliability, Accountability, Trustworthiness

5) Privacy. Privacy refers to the right of individuals to retain certain personal information without disclosure and to be protected from unauthorised access of any information about themselves, even if it is collected with their consent. *Rele-*

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22915695

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

vance for OWI: In the context of OWI, this refers to i) the personal data collected through indexing and end-user interactions, and ii) the responsible handling of that data. Applications such as search engines must respect the informational self-determination and privacy of each individual. Furthermore, individuals have the right to make informed choices about data collection, advertising and personalisation tools. This also means that there must be easy ways to obtain consent and remove content. *Related concepts:* Transparency, Security/Safety, Openness, Responsibility, Trustworthiness, Sovereignty/Autonomy, Sustainability/Social Responsibility

6) Justice. Justice means treating everyone fairly and equally. It means making sure that decisions and actions are fair and that everyone has the same opportunities and access to resources and information without bias. This is specifically important in diverse groups with differences and similarities between people related to gender, age, religion, ethnicity, etc. *Relevance for OWI:* Search engines and other applications created by index users, in particular third parties, should be able to be used in a fair and non-discriminatory way. All stakeholders should be treated fairly in all situations. *Related concepts:* Diversity and Inter-cultural Sensitivity, Reliability, Accountability, Fairness, Non-discrimination

7) Trustworthiness. Trustworthiness is related to ethics in two specific dimensions, namely integrity and benevolence. A trustworthy party is one who does not unfairly exploit the vulnerabilities of the other party in the relationship. Trustworthiness relates with trust, as a person can trust someone or something that is trustworthy. Trustworthiness is also related with reliability in terms of consistency and dependability of a process, system, or measurement to perform its intended function or produce consistent results in time. *Relevance for OWI:* Trustworthiness refers to the processes of OWI that ensure the integrity of data and services. This includes careful and reflective design of processes and algorithms, and responsiveness to problems. In particular, it must focus on the data and algorithms not only from a functional point of view, but also from the perspective of the stakeholders who use the system. Trustworthiness is also related to usability and simplicity. Systems that are designed in a usable and simple way are easier to understand, which increases trust. *Related concepts:* Usability and Simplicity; Reliability, Quality, Accuracy, Fairness

8) Accountability. Accountability refers to the obligation of admitting and reacting in case of damage or harm to individuals or to society has occurred. It refers to actions to be taken in case of problems becoming visible. It includes liability, transparency, admitting mistakes and taking corrective action. *Relevance for OWI:* The operators or maintainers of the Open Web Index need to respond to incidents and problems that emerge in relation with the OWI. Swift actions should be undertaken in response to critical information from external stakeholders about the information they provide. *Related concepts:* Responsibility, Fairness, Transparency

9) Sustainability. Sustainability is a social goal for human co-existence on Earth over a long period of time. Specific definitions of this term are controversial and have varied with literature, context and time. Experts often describe sustainability as having three dimensions: environmental, economic and social, with a specific focus on the environmental dimension. *Relevance for OWI:* In the context of the Open Web Index, all three dimensions of sustainability are relevant. Environmental sustainability refers to energy efficiency of the software used and the hosting in data centres. Economic sustainability refers to the business-oriented applications that rely on the OWI. Social sustainability refers to the provision of a critical search infrastructure that is of value to society. *Related concepts:* Transparency, Responsibility

10) Sovereignty/Autonomy. Digital sovereignty and autonomy refer to the right and ability of individuals and entities to control their digital presence, data, and online interactions without undue influence from external forces. It emphasises independence in managing one's digital footprint and ensuring that digital practices respect individual and collective autonomy. *Relevance for OWI:* Internet search as a key resource in our knowledge society must be independent of political and economic influence. *Related concepts:* Transparency, Justice, Responsibility

TECHNICAL AND ORGANISATIONAL MEASURES

Based on the ethical risk and value analysis described in the last section, technical and organisational measures have been developed that aim to mitigate the ethical risks and adhere to the ethical values. This informs the development of the technical components, the operation of the OWI including crawling and data sharing, and the usage of the index data downloaded from the OWI. In this way these measures address the impact on the stakeholders as identified in the last section and help reduce harm from them.

M1 Website Management. Website management is the regulation and processing of which websites are indexed and which are not. It also regulates how the content of websites is processed, used and shared. It involves several stakeholders, such as website owners and webmasters, as well as end-users who are affected by the content of the websites. Several technical and organisational measures have been put in place to enable website management that meets the ethical requirements. (a) Opt-out mechanism: Site owners or webmasters can choose not to have their websites included in the index. By adjusting the robots.txt file of their webpages, the "no_index" flag can be set, which prohibits further processing of the site's content, or the "no_genai" flag indicates that this site must not be used for training AI models. (b) Blacklisting: Websites can be blacklisted in the index and thus excluded from further processing. Blacklisting is based on an initial seed list of blacklisted URLs or on the detection of malicious content in a website (see M3). (c) OWI Dashboard: The dashboard is the central hub that provides transparency to site owners, webmasters and the public. It provides an overview of all indexed sites, e.g., allowing site owners and webmasters to check whether their site is indexed or not. A

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22915695

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

search can be done by URL, title, metadata or content (e.g., a person mentioned on a page). The dashboard also allows site owners or webmasters to explicitly add or remove their site from the index. Adding a URL will result in that site being crawled and indexed. (d) Prioritisation: The prioritisation mechanism ensures certain sites are crawled and indexed earlier. For instance, popular sites or sites of high public value (an indicator of this is a link from Wikipedia, for example) are crawled with higher priority. Overall, these measures provide transparency to site owners and the general public about which sites are included in the index. They provide fairness to site owners by giving them control over how their sites are used and processed and they demonstrate responsibility by helping to prevent harm to end users. This approach also demonstrates openness, as new content can be added to the index, for example by webmasters.

M2 Take-down Management. Closely related to website management, take-down management ensures that requests to remove websites or web pages from the index are checked and, if justified, carried out. The ability to submit take-down requests has already been implemented: Both website owners and affected individuals and stakeholders can request that a website be removed from the index. (a) A website to be removed can be searched for in the OWI Dashboard. (b) If the page is available, a request for removal can be sent to the OWI operators. The reason for the request must be given. Possible reasons include, for example, a website owner requesting the removal of one of their own web pages, or any person discovering illegal content requesting the removal of the website in question. (c) Validation of requests is necessary. For example, site owners would have to prove their identity. Or take-down requests citing illegal content would require personal identification to prove the validity and would need to be evaluated. A validation of requests is not currently implemented due to a lack of resources. Human and technical resources to validate the requests need to be considered in the future. (d) Websites take-downs are recorded in log files. Overall, these measures address the ethical risk of erroneous content. It also demonstrates responsibility and accountability for the crawling activities.

M3 Content analysis. Content analysis is performed at different levels and in several phases. Content analysis is particularly important, as the web index might include websites that contain ethically problematic content, such as fake news, false information, or psychological triggers. Automatic detection of such websites results in flagging these websites, so that index users can pay attention to them. In particular, the following has been achieved: (a) Web pages are analysed after crawling for malicious or adult content, which leads to blacklisting or tagging. (b) In the pre-processing phase, research has been conducted and methods have been implemented to identify quality issues, trigger warnings, and advertisement. Trigger warnings are used to inform viewers of the presence of potentially disturbing content. Content issues are detected by analysing the content quality and assigning a score. (c) The open architecture of the pre-processing pipeline allows the inclusion of plug-ins that analyse content for further ethical and legal issues. (d) The external project (VERITAS) developed a method to detect disinformation on websites [10]. This or similar method can be included in the pre-processing pipeline. (e) The wealth of content available in the Open Web Index enables researchers to analyse large chunks of web data in detail externally. This offers a completely new way of gaining an overview of the presence of problematic content on the Internet.

Overall, content analysis contributes to the trustworthiness of the OWI. Tagging or removing flawed content is part of the responsible design and avoids harm to end-users.

M4 Access control. An access control is set up for the OWI, which regulates who has access to the index data. (a) To retrieve index data from the OWI, users must register with an EUDAT B2ACCESS account. This account identifies users with their home organisation account (e.g., a university), but also allows social media accounts. Furthermore, users can register with their email account. (b) In case of misuse, index users can be immediately blocked and excluded from further retrieval of index data. Overall, access control allows for the mitigation of ethical and legal misuse of the index data, as accounts can be blocked if users are found to be in breach of ethical and legal terms of use.

M5 Transparency guideline. A transparency guideline and label has been developed that helps developers of third-party (search) applications to adopt and integrate transparency features to increase trust of end-users, enhance accountability and fairness, ensure compliance with regulatory standards and mitigate potential risks or misuse. (a) A guideline has been created that consists of 27 aspects of how to improve transparency in search applications, such as providing information about the funding and business model of the application. This guideline suggests four different ways to address the aspect, which are explanation, exploration, processed data and reflection. (b) The guideline is realised as a questionnaire. Answers are presented in a visualisation, resulting in a transparency label. The label can be displayed prominently in the own application. It allows the end-user to quickly get information about the transparency features and to compare applications related to this label. Overall, the transparency guideline and label increase the transparency of an application and should also increase trust in the application. Trust is built through the four methods of explanation, exploration, processed data and reflection. It also supports the ethical use of index data.

M6 Data protection and privacy. A data protection and privacy concept has been created for application developers. (a) User data collected on the search client is sent to backend services as anonymised and vectorised queries. Only user data that is essential for further processing is sent to the services. Overall, this ensures privacy and protection of user data. It provides a model that can be reused and adopted by other applications.

M7 Safety and security management. Regular security meetings are held where the development and infrastruc-

https://doi.org/10.5281/zenodo.22915695

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ture teams discuss and analyse potential security threats and safety issues of the software components and infrastructure settings. (a) During the meetings, a security analysis of all technologies (e.g., libraries and tools) used in the OWI's infrastructure is performed on a regular basis in order to identify potential security risks. This also includes monitoring of new information on security issues (e.g., via newsletters). (b) An incident reporting and management mechanism has been set up to allow website owners to complain about automatic crawling of their website. Webmasters can exclude the crawler (OWLer) in the robots.txt file. They can also contact the crawling team by email. The security meetings address the security value, as they help to avoid security risks that could lead to erroneous content or misuse of content. Incident management addresses the security value by mitigating bugs and misbehaviour of the crawler.

M8 OWI License. The license is a contract between the OWI operators and the index user and defines the rights and obligations of the index user. Though the license is primarily a legal contract, it also has a strong ethical dimension, as it helps reduce the misuse of the index data. Three types of OWI licenses have been created to regulate the terms and conditions under which the OWI may be used by third-party index users. (a) The first type is the research license that allows uses to use the index only for research purposes. This is the recommended license under which index users can try out the OWI, explore the available data, or develop prototype applications. (b) The second type is the commercial license that allows users to use OWI data for commercial purposes. This is the type of license if index users want to include OWI data in a commercial application or for other commercial purposes. Granting a commercial license, however, poses some liability risks to the OWI operator, for example, if copyright material is part of the OWI data. (c) The third type is the data access license, that regulates the use of raw crawled data. Sharing raw data needs specific attention, as it is the original data of website owners and other people. Overall, this should reduce the risk of misuse of index data.

M9 Legal and ethical self-assessment. The self-assessment ensures that index users understand the ethical and legal requirements. Legal and ethical self-assessment questions have been defined that summarise the most important aspects of legal and ethical use of the index data. Index users are required to fill out the questionnaire, which makes them aware of the legal and ethical requirements. The assessment will be done per user account. Based on the answers an ethics score is calculated that is presented to the respective users only. Overall, this should reduce the risk of misuse of index data.

CONCLUSION

This paper presents an ethics-by-design approach and methodology describing how ethical requirements were integrated into the design and development of the Open Web Index. The outcome of the ethics-by-design methodology is an ethics framework that provides information about the ethical requirements in the form of risks and values, as well as guidance on how the requirements can be addressed. An important aspect of the ethics-by-design approach is the mutual influence between the creation of the ethics framework and the technology design: rather than simply imposing ethical requirements on the technology, this approach benefited from continuous exchange between the two. Although not all ethical requirements could be fully addressed through technical or organisational measures, the most important ones were taken up. The framework can serve as a basis for further development of the OWI, and as a template for developing other technologies.

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).

Some of 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.

REFERENCES

[1] M. Castells, *The Rise of the Network Society*. Wiley-Blackwell, 2010. http://doi.org/10.1002/9781444319514

[2] A. Halavais, *Search Engine Society*. polity, 2008.

[3] F. Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Information*. Harvard University Press, 2015.

[4] S. Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Public Affairs, 2019.

[5] 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

[6] A. Nussbaumer, C. Plote, and M. Granitzer, "Deliverable 6.3: Final ELSA-Catalogue & Code of Conduct for Open Web Search", OpenWebSearch.eu, Tech. Rep., 2026.

[7] M. L. Cummings, "Integrating ethics in design through the value-sensitive design approach", *Science and Engineering Ethics*, pp. 701–715, 2006. 10.1007/s11948-006-0065-0

[8] B. Friedman, P. H. Kahn, A. Borning, and A. Huldtgren, "Value sensitive design and information systems", in *Early engagement and new technologies: Opening up the laboratory*. 2013, pp. 55–95. 10.1007/978-94-007-7844-3_4

[9] A. Nussbaumer, A. Pope, and K. Neville, "A framework for applying ethics-by-design to decision support systems for emergency management", *Information Systems Journal*, vol. 33, no. 1, pp. 34–55, 2023. https://doi.org/10.1111/isj.12350

[10] P. Bisconti, "VERITAS project final activity report", OpenWebSearch.eu, Tech. Rep., 2025. https://doi.org/10.5281/zenodo.17588890

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22920023

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

Co-Creating Inclusive AI Search Frontends: Findings from the KIsu-Project with children, young people, and older adults

M. Platz, S. Malone, V. Wolf, K. Altmeyer, M., S. Heel, K. Reese, L. Schick, Saarland University, Saarbrücken, Germany
T. Gottsmann, M. Wiesner, fragFINN e.V., Berlin, Germany
J. Gabelmann, Saarbrücken, German Research Center for Artifical Intelligence, Saarbrücken, Germany
A. Rodrigues Eismann, Saarbrücken, Centigrade GmbH, Saarbrücken, Germany
U. Gmelch, C. Plote, Open Search Foundation, Starnberg, Germany

Abstract

Across all age groups and ability levels, internet searches present significant hurdles for users who struggle to translate their information needs into effective queries. Children, older adults, and people with special needs are disproportionately affected: vague or incomplete search terms frequently return irrelevant results, causing frustration and discouraging further engagement with the web. Beyond usability, this gap in search competence carries a deeper risk – without the skills to evaluate results critically, users are particularly susceptible to misinformation and algorithmically shaped information asymmetries.

The KIsu project ('Kinderleicht KI-basiert suchen und finden' – Finding and Searching with AI, Easy and Child-friendly) was a 19-month cooperative research initiative (June 2024–December 2025) between Saarland University and fragFINN e.V., funded by the German Federal Ministry for Education, Family Affairs, Senior Citizens, Women, and Youth. It developed an AI-supported chatbot using the open-source Llama 3.1 8b Instruct model, integrated with the child-safe fragFINN.de search index, alongside comprehensive Open Educational Resources (OER) and empirically grounded design principles for inclusive AI search frontends. This paper presents the completed project: its methodology, key findings from co-creative workshops with over 100 participants and a controlled intervention study (n=71), central results, and implications for future research and practice. It builds upon and extends a preliminary conference contribution presented at the Open Search Symposium 2025 in Helsinki [6].

INTRODUCTION

The global search engine landscape is heavily concentrated: Google holds the dominant market position, with Bing supplying the underlying index for most competing services and gaining further ground through integration with the ChatGPT AI assistant. Neither provider discloses its ranking algorithms or index composition, creating systemic opacity around transparency, accessibility, data protection, and security. This opacity can produce bias, information asymmetries, and data-driven discrimination [1, 2]. Despite their central role in everyday information access, most people interact with search engines unreflectively, without understanding the mechanisms that shape their results. The rise of generative AI compounds these effects: language models introduce new layers of opacity while accelerating the pace at which misleading content can be produced and distributed. At the same time, commercial incentives give major providers little reason to invest in research on, or infrastructure for, genuinely barrier-free access.

The KIsu project was conceived as a direct response to this situation. Its core ambition was to equip all citizens – irrespective of age, gender, or disability status – with the knowledge and tools to navigate online searches confidently and autonomously. To that end, the project combined the development of targeted learning materials and participatory workshops with the construction of an AI-powered search interface. A particular focus was placed on identifying the specific obstacles facing each target group and on addressing them with both pedagogical and technical means.

Against this backdrop, according to the 2024 KIM Study [3], more than two-thirds (69%) of 6- to 13-year-olds in Germany use Google at least once a week. According to the 2025 JIM Study [4], an equally high proportion (70%) of young people (aged 12–19) use AI applications such as ChatGPT for schoolwork and other purposes.

Original sources are checked less frequently when AI is used, and the risk of uncritically adopting AI outputs without source verification is especially pronounced for socially disadvantaged children without academic backgrounds. For older adults, the lack of AI competencies leads to significant participation problems and heightened vulnerability to disinformation.

OBJECTIVES

With KIsu, we aimed to strengthen the digital sovereignty and participation of people with special needs, children, and older people, and to enable as many user groups as possible to use search engines efficiently and inclusively. Concretely, the project pursued five goals:

1. Promoting the competent and responsible use of internet searches through developing and testing learning and information materials.
2. Identifying factors and specific needs in dealing with search engines, and deriving implications for the

https://doi.org/10.5281/zenodo.22920023

development of learning materials and low-barrier AI search frontends.
3. Fine-tuning a needs-based AI language model and developing a model-agnostic data pipeline.
4. Provision of AI frontends for optimized, barrier-free search (technology maturity level 5).
5. Provision of the data collected in the project and publication of research results under open access principles.

PROJECT STEPS

The project ran for 19 months (01.06.2024–31.12.2025) and was structured into four steps:

- *Step 1 – Material and frontend development:* Learning, training, and information materials were co-creatively designed with groups from civil society. The first intuitive frontend prototypes were developed using a modular design approach.
- *Step 2 – Material and frontend optimization:* Materials and frontends were tested with a controlled intervention study (n=71) and iteratively optimized based on results and user feedback.
- *Step 3 – Dissemination and final development:* A training concept and OER were prepared in cooperation with civil society groups. Design principles for low-barrier AI search frontends were derived, implemented, and published. A closing conference was held with all project participants.
- *Step 4 – Provision of data and publication:* Collected data was processed, anonymized, and submitted to a research data repository. Research results were published under open access.

RESEARCH DESIGN AND RESULTS

The project adopted an Action Design Science Research (ADSR) framework [5]. Rooted in the philosophy of pragmatism, Design Science Research addresses real-world problems through the creation of purposeful artefacts; the ADSR variant places co-creative collaboration between researchers and users at its center. Empirical data collection combined quantitative instruments (standardized questionnaires, structured pre- and post-tests) with qualitative methods (semi-structured interviews, participant observation, and co-creative workshop protocols), in line with a mixed-methods design [6]. An overview of the project design and early-stage results was previously presented at #ossym2025 [6]; this paper reports on the fully completed project.

Quantitative data were analyzed descriptively and inferentially. Qualitative data were analyzed using Design Thinking methods, such as the development of Personas [8], combined with qualitative content analysis [9]. The diagnostic phase analyzed the problem and solution spaces in mutual agreement with the researcher-user team. The design phase then identified and conceptualized artefacts (learning materials and low-barrier AI search frontends) through iterative cycles. The implementation phase evaluated efficiency and effectiveness in practice.

Co-creative Workshops

User involvement was central to the project methodology throughout. Drawing on the principle that end-user participation yields more relevant and usable outputs while reducing development risk [10], the project adopted participatory design and conducted iterative prototype testing guided by direct user feedback. Workshops were structured using a Design Thinking framework [8], with participants working in groups of three to four on small whiteboards, using marker pens and symbol cards to sketch their ideal search companion. To avoid anchoring participants to familiar solutions, the prompt described this tool as a 'digital assistant' rather than a 'search engine', encouraging genuinely open-ended and creative responses.

Overall, co-creative workshops were conducted with over 100 participants across all target groups with attention to balanced gender composition:

• Children (2nd grade, ca. 7 years old): 15 children (many children with German as a second language); and 20 children.

• Mathematically interested children (3rd/4th grade, 9–10 years old): 14 children.

• Young people (7th grade, ca. 13 years old): 19 participants; and 5th grade (ca. 11 years old): 17 participants.

• Older adults (aged 64–87): two groups of 14 and 16 participants.

Additional public outreach formats significantly broadened participation. A 'Children's University' lecture reached 510 children aged 8–12, of whom 345 participated in a subsequent questionnaire study. Further events included the Open Campus at Saarland University, make-it.saarland: inklusiv, Inklusionstage Berlin, and the Tag der Deutschen Einheit in Saarbrücken.

The workshops yielded a rich set of insights specific to the target groups. Children placed strong emphasis on social warmth and friendly interaction in their ideal assistant, while for older participants, trustworthiness emerged as the single most important attribute. Voice-based input and output proved appealing across all groups. Among the young adolescent participants – who showed the highest baseline knowledge of search strategies, AI, and digital tools – there was a clear demand for adaptive responses: the assistant should calibrate the depth and tone of its answers to the user's actual needs, including adjusting to a more playful register when the search topic calls for it. A cross-cutting finding with implications for curriculum design was that girls consistently showed lower confidence and engagement with AI tools, pointing to a structural gender gap requiring dedicated intervention.

Controlled Intervention Study

Based on the co-creative workshop findings, a controlled intervention study was conducted with 71 school students (grades 5–8, Montessori school, including children with special educational needs: ADHD, dyslexia, hearing impairment). The study addressed three research questions:

1. *Does the use of specifically developed learning materials and the customized frontend lead to measurable*

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22920023

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

improvements in 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 informational materials impact children's interaction with the frontend during internet searches?*

Participants were randomly assigned to two intervention groups. Intervention Group 1 (IG 1) first engaged with the learning materials, then completed a structured internet search task using the newly developed frontend, and finally evaluated its usability. Intervention Group 2 (IG 2) reversed the sequence: frontend use preceded engagement with the learning materials. Both groups completed pre- and post-tests assessing knowledge and attitudes toward internet searching and AI.

The results showed that even short, targeted self-learning significantly improves knowledge of AI-based search. Post-test results were better than pre-test results across both groups. A particularly notable finding was that higher knowledge about AI chatbots correlated with more frequent use, greater enthusiasm, and, at the same time, more critical evaluation. This suggests that AI literacy and critical media competence are mutually reinforcing rather than in tension with one another.

AI frontend

The 'Kisu' chatbot sits at the heart of the project's technical output. Its interface is designed to accept and interpret queries as users actually phrase them – including fragmented, imprecise, or context-dependent inputs typical of children, older adults, and people with special needs. An integrated LLM processes each query, refines it into an optimized search expression, and routes it to the fragFINN.de index. When the user's intent is unclear, the system initiates a brief clarifying dialogue rather than returning potentially irrelevant results.

Reaching Technology Readiness Level 5, the chatbot runs on Llama 3.1 8b Instruct, hosted by GWDG, chosen deliberately as an open-source alternative to services such as ChatGPT. Earlier development stages also explored Llama 3 and Mistral, fine-tuned using LoRA and QLoRA. The open-source route offered four decisive practical benefits: no dependency on a single commercial vendor, full visibility into model behavior, substantially lower operating costs, and straightforward auditability of data-handling practices. The completed API integration with fragFINN.de ensures that all child users receive results from a curated, age-appropriate index.

Specialized system prompts were developed for multiple query modes: broadly formulated queries, factual knowledge, current events, web navigation, small talk, intent-oriented queries, and incomplete queries. The frontend was continuously optimized based on workshop feedback. Key features added during iteration include automatic input field focusing, process indicators during AI processing ('I am thinking...'), hover explanations for buttons, and the option to clear the chat history. Hosting by GWDG was identified as an important factor ensuring better data protection, particularly relevant when working with sensitive target groups such as children.

LEARNING MATERIALS AND OPEN EDUCATIONAL RESOURCES

All learning materials developed in the project are published as Open Educational Resources (OER) and freely accessible via the project website (https://kisu.cs.uni-saarland.de). The materials comprise:

- A digital learning program on the KIsu website
- A KIsu handbook for primary school pupils
- A KIsu handbook for senior citizens

The materials cover three thematic areas: Artificial Intelligence, Search Engines, and AI-supported Search Engines. They promote competencies in the areas of Foundations, Input, and Output. The materials were developed iteratively in close collaboration with target group participants and refined based on empirical feedback from workshops and the intervention study.

KEY FINDINGS AND COMPETENCE AREAS

The project identified five central competence areas as crucial for all target groups:

1. *AI Literacy:* Understanding how AI systems and language models function. Higher AI knowledge correlates with more critical evaluation, meaning AI literacy and critical media competence are mutually reinforcing.
2. *Search Engine Competence:* The ability to formulate precise search queries and interpret results. Many users cannot formulate effective queries, leading to frustration and restricted internet use.
3. *Critical Information Evaluation:* The ability to assess the reliability of information. Given intransparent algorithms, commercial bias, and the increasing spread of AI-generated content, this competence is essential.
4. *Data Protection and Transparency Awareness:* Understanding user profiling and manipulative design patterns in commercial search systems, as a foundation for self-determined usage decisions.
5. *Accessibility Awareness (developer-side):* The ability to recognize and remove technical, cognitive, and informational barriers, particularly vital for enabling vulnerable groups to participate in digital society.

Across all groups, an important conceptual tension emerged: the balance between offering simplified, direct AI-generated answers (preferred especially by younger children) and fostering source verification and critical media competence. This tension has no simple resolution and requires continued pedagogical attention.

CONTRIBUTION TO THE COMMON GOOD

The KIsu project contributed to the common good on multiple levels. In terms of digital inclusion, it reduced search barriers for disadvantaged user groups – children,

https://doi.org/10.5281/zenodo.22920023

older adults, and people with special needs – actively counteracting digital exclusion. In terms of transparency and enlightenment, it provided education about non-transparent algorithms, user profiling, and manipulative designs in commercial search engines and AI tools, fostering critical media literacy across civil society.

In the spirit of Open Science, all learning materials were published as OER, research data and design principles were made available for the research community, and the AI frontend code is publicly accessible at https://github.com/kisu-saar/Kisu. Free access to AI-supported search tools, free of vendor dependency, contributes to the democratic availability of these technologies. By strengthening search engine competence and AI literacy across educational and social institutions, the project aims for a lasting societal impact.

CHALLENGES AND OPEN QUESTIONS

Challenges Encountered

Technical challenges included finding the right balance between usability and functionality and managing performance issues caused by longer AI response times. Accessibility could not be fully optimized for all target groups within the project period.

Usability hurdles were group-specific. Older adults had difficulty locating the input field and the search results; the need to click into the input field was not intuitive to them. Children often did not recognize that they could click on result links. Methodological challenges arose from the limited empirical research base on search behavior of heterogeneous target groups, requiring the project first to build a solid empirical foundation. Structural challenges included the intransparent algorithms of commercial search engines and the Digital Divide, which showed that socially disadvantaged children require more intensive support.

Open Questions for Future Research

Several important questions for follow-up research emerged from the project:

- *Long-term anchoring:* How can KIsu be sustainably integrated into school curricula and extracurricular educational offerings?
- *Gender disparity:* What additional measures are needed to address the finding that girls require targeted AI offerings systematically?
- *Transferability:* How can the derived design principles be transferred to other AI applications to multiply their benefit?
- *Advanced accessibility:* What further developments are needed to better support people with cognitive and physical limitations, e.g., optimized color design for color blindness, improved screen reader compatibility, and eye-tracking support?
- *Direct answers vs. source verification:* How can the balance between simplified AI-generated direct answers and fostering source literacy be optimally designed?
- *Scalability and sustainability:* How can the technology be scaled and sustainably financed after the project's conclusion, so that achieved progress is not lost?

CONCLUSION

The KIsu project completed its 19-month run in December 2025, achieving all stated project goals. Building on the objectives and early-stage design presented in [9], the completed project produced a functional open-source AI search chatbot, a comprehensive OER for multiple age groups, empirically validated design principles for inclusive AI search frontends, and new insights into the AI literacy needs of diverse user groups.

The project demonstrates that even short, targeted learning interventions can significantly improve AI literacy and critical search competence. It further shows that open-source LLMs can be successfully deployed as alternatives to proprietary systems for public benefit applications, combining data protection, cost-effectiveness, and sufficient output quality. The involvement of over 100 participants in co-creative workshops across all target groups – including children with special educational needs, older adults, and socially disadvantaged youth – provides a solid empirical basis for the derived design principles and learning materials.

The project can unfold its long-term impact through:

1. the use of training modules in training and train-the-trainer courses,
2. the publication and utilization of design principles and study results for the development of inclusive AI frontends,
3. building a multiplier network through training and OER, and
4. public awareness raising.

In the long term, we hope that AI will recognize users' search intentions and proactively support them in specifying queries, critically evaluating information, and making informed decisions – but this vision can only fully unfold with competent users, making the continued promotion of AI literacy a societal imperative.

ACKNOWLEDGMENTS

We thank all the participants in the co-creative workshops and surveys, the 'Familienbildungsstätte Saarlouis', the 'Lebenshilfe für Menschen mit Behinderung Obere Saar e.V.', the 'Grundschule Mellin' and the 'Montessori Schule Saar', 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

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

https://doi.org/10.5281/zenodo.22920023

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

gap in search engines. European journal of law and technology, 8(2),1–18.

[3] KIM (2024). *KIM-Studie 2024.* Mpfs. https://mpfs.de/studie/kim-studie-2024/

[4] JIM (2025). *JIM-Studie 2025* Mpfs. https://mpfs.de/studie/jim-studie-2025/

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

[6] Altmeyer, K., Hladky, M., Malone, S., Platz, M., Reese, K., Schick, L., Wolf, V., Gottsmann, T., Wiesner, M., & Plote, C. (2025). The Project KIsu: AI-based searching and finding – easy, inclusive, and self-determined. *Open Search Symposium 2025 (OSSYM2025),* CSC IT Center for Science, Helsinki, Finland. https://doi.org/10.5281/zenodo.17229191

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

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

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

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

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22920264

OPEN, TO WHAT END? A CAPABILITY-THEORETIC PERSPECTIVE ON OPEN SEARCH

Nicola Neophytou*, Independent Researcher, Tiohtià:ke / Montréal, Canada
Bhaskar Mitra†, Independent Researcher, Tiohtià:ke / Montréal, Canada

Abstract

The hegemony of control over our search platforms by a few large corporations raises justifiable concerns, particularly in light of emerging geopolitical tensions and growing instances of ideological imposition by authoritarian actors to manipulate public opinion. Recent movement for promote open search has emerged in response. This follows from past and ongoing push for openness to challenge corporate oligopolies (*e.g.*, open source and open AI models) which have seen significant ongoing negotiations and renegotiations to establish standards around what constitutes being *open*. These tensions have hindered these movements from effectively challenging power, in turn allowing powerful corporations to neutralize or co-opt these movements to further entrench their dominance. We argue that the push for open search will inevitably encounter similar conflicts, and should foreground these tensions to safefguard against similar challenges as these adjacent movements. In particular, we argue that the concept of *open* should be understood not with respect to what is being made open but through a capability-theoretic lens, in terms of the capabilities it affords to the actors the system is being opened to.

INTRODUCTION

Search systems inherently mediate what information receives exposure, and in doing so shape our political opinions, what we choose to consume, and how we perceive our place in society. The hegemony of control over our search platforms by a few large corporations has therefore triggered justifiable concerns, especially in light of emerging geopolitical tensions between the US and its allies [1, 2] and increasing instances of ideological imposition by authoritarian state and private actors to manipulate public political opinion [3–6]. These concerns are compounded by the increasing integration of generative AI models in information access, which are becoming increasingly effective in manipulating mass opinion [7]. In response, there has been recent calls for prioritizing societal needs over corporate interests [8] in search, counter-imagining hegemonic search as foundational infrastructure [9], establishing democratic digital public sphere [10], developing public interest search platforms [11], and safeguarding our information ecosystem from authoritarian and corporate capture [12, 13]. Each of these calls underscore the need for open search alternatives.

From source code to AI models, computing has a long tradition of pursuing openness as a mechanism to challenge capitalist and oligopolistic power. These past and ongoing movements represent sites of conflict between corporations seeking industry dominance and the impacted communities exercising resistance. Within these struggles, the standards of openness are constantly negotiated and renegotiated. These tensions can be observed, for example, in the divergence of the so-claimed technological pragmatism of the open source movement from the sociopolitical aspirations underpinning the free software movement [14]. Also notable in this struggle are the strategies employed by large corporations to defend their oligopolistic power—*incl.* smearing [15], sabotaging [16], co-option [17], and open-washing [18]. These tensions have been renewed in recent years within the open AI movement, where the definition of "open" is under contestation [19, 20] and open-washing is rampant [21, 22]. In response, several frameworks have been proposed to measure openness of AI models [22–24], grading them based on what aspects of the model and its production process are being made open. The open search movement is still in its infancy, yet we can observe renegotiations on the notion of openness in the process of its conception [9]. We posit that a reexamination of the concept of "open" is necessary because of the shared tensions between open search and previous open movements.

Movements for openness—be it for source code, models, data, or systems—are generally motivated by specific social, political, economic, or technological aspirations. However, the negotiations that follow often center *what* is being made open rather than whether the act of opening brings us closer to our stated motivations. This creates opportunity for powerful corporations to benefit reputationally from participating in the open ecosystem and extracting value from the movement to further entrench their dominance, without substantially satisfying any of the socioeconomic or political motivations of the open movement. Even if the system is being made open with sincere intentions, the inordinate focus on what is being made open detracts the community from undertaking the critical examination of: *to what end* does this opening contribute, and who benefits from it?

Towards this goal, we argue that openness should be understood and appraised with respect to the capabilities it affords to actors that the system is made open to. We draw from Amartya Sen's Capability Approach (CA) [25] to reframe *openness* centering desired well-beings, and corresponding capabilities actors must be granted to achieve them.

WHY DO WE NEED OPEN SEARCH?

No search platform is "neutral" as every decision about how to organize and surface information is saliently political.

* neophytounicola@gmail.com
† bhaskar.mitra@acm.org

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22920264

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

Every stage of their design involves normative choices that determine what information is deemed relevant and whose perspectives are considered worthy of exposure. Dominant search companies make these decisions behind closed-doors, sheltered from public scrutiny. This lack of transparency and accountability raises concerns that search systems may inadvertently (or deliberately) foster ideologies that align with those of the platform operators, or the state actors who have sway over them. *E.g.*, U.S. president Donald Trump's "Preventing Woke AI in the Federal Government" executive order imposes the Trump regime's own ideological agenda [3] on AI models, with implications for the search systems that incorporate them. Given existing evidence that search results can be instrumental in influencing political opinion [26] and emerging evidence on AI persuasion [7], this raises serious concerns for how public opinion is vulnerable to manipulation by private corporations and authoritarian regimes.

The European Union (EU)'s attempts to rein in Big Tech's power through regulations have been met with intense lobbying and political maneuvering [27–29]. Concurrently, rising geopolitical tensions involving the US [1, 2] has attracted further attention to the vulnerabilities of a digital ecosystem in which many countries rely heavily on the tools controlled by a few US-based firms. In response, several countries, including Canada [30] and in the EU [31], are accelerating **digital sovereignty** mandates to enhance the autonomy of their national data and digital infrastructure. These initiatives seek to reduce dependence on overseas tools by creating self-hosted and local alternatives, offering industry for local economies, and the opportunity to engrain a culture of compliance-by-design with local laws. Projects like the Open Web Search (OWS)[1], and the related Open Web Index (OWI)[2] initiative, funded by the EU and "based on European values" [32], are examples of how open and publicly governed search alternatives are emerging to address concerns around digital sovereignty and promote global competition.

An open search ecosystem may also encourage investments in niche markets and **addressing historically underserved linguistic, regional, and sociocultural needs** of communities, filling important gaps where the current dominant search platforms do poorly. This includes robust information access in and for under-served languages, ensuring appropriate representation of historically marginalized peoples and cultures, indexing and retrieving regionally relevant and community-specific knowledge and cultural artifacts, and addressing local and community-specific needs.

The development of open search also lowers the bar to entry for new search providers to compete by leveraging shared infrastructure without needing to build from scratch. This has the potential to encourage a faster pace of **innovation through market competition**. One can point to historic cases of innovation under open internet standards to support the movement for shareable and co-optable infrastructure [33]. This is particularly relevant in light of US Big Tech monopolies stifling competitors both in the US [34, 35] and globally—*e.g.*, in the EU [36, 37]. However, in the realm of AI, in spite of a burgeoning ecosystem of open models, smaller companies have struggled to be competitive, and the economic benefits of AI still flow exclusively through the major AI monopolies [38]. The open search movement must therefore consider that the reality of achieving real competition goes beyond just creating an open ecosystem, and must circumvent the levers of power that Big Tech instrumentalizes to maintain their dominance.

Open search also supports **promoting open science and rigor in information access research**, echoing the ethos of scientific reproducibility and lowering the barrier between academic research and production systems. Open search in this context may enable researchers to conduct their experiments under more realistic system settings, provide access to search logs and live traffic that academic researchers typically do not have access to, and make it convenient for the community to reproduce and build on each other's work.

Furthermore, an open search ecosystem also presents an opportunity to progress towards recent calls [8, 12, 39] for **realizing information access as a force for justice, emancipation, and democracy**. Profit motives have historically been the key driving force for proprietary platforms. *E.g.*, in the social media space, Meta has repeatedly evaded the responsibility to reduce misinformation and radicalization on its platforms [40], echoing how privately-owned platforms prioritize engagement and growth at the cost of information integrity and societal good. An open search ecosystem is an opportunity for us to deliberately consider how we can build information access platforms centering our societal and democratic needs, instead of being reactionary to potential social harms of emerging information technologies [8]. This requires radical rethinking of the design of our search systems and developing effective mechanisms for democratic governance of our platforms since we cannot simply rely on state regulations to uphold our ethical values [13]. Compliance with regulations does not necessarily imply that societal concerns are wholly accounted for, and does not ensure genuine emancipatory outcomes.

Next, we present a capability-theoretic perspective on open search to formalize the connection between these motivations and the capabilities required to attain them.

RETHINKING SEARCH OPENNESS

We present an alternative perspective on openness in search, shifting the focus from assessing transparency and access to individual stack components to instead evaluating openness by what *capabilities* are enabled by opening up the system. This theoretical move uncovers two key insights currently underrepresented in the discourse around open search; first, openness is a means to an end and should be appraised with respect to the affordances that result from it, and second, the capabilities that result from system openness are not universal and instead depend on the circumstances of actors that the system is purportedly being opened to. We

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

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22920264

draw from Amartya Sen's Capability Approach [25] from welfare economics to reframe openness for search. While our focus is on open search, we believe this reframing also speaks to the shared underlying tensions we observe across other movements for open source, data, models, and science. We posit that the open search movement would benefit from adopting the capability-theoretic lens to reexamine its motivations and how purported strategies of *opening* the system is supposed to help move us towards those goals.

The Capability Approach

CA outlines a philosophical framework for evaluating a person's well-being in terms of their capabilities and functionings [25]. Sen defines a *functioning* as an "achievement of a person: what he or she manages to do or be" which may contribute towards their subjective *well-being*. A *capability* then reflects a person's "ability to achieve a given functioning" [41] given the *commodities* available to them.

$$\textit{Commidity} \rightarrow \underset{\text{(to function)}}{\textit{Capability}} \rightarrow \textit{Functioning} \rightarrow \textit{Well-being}$$

CA emphasizes that commodities alone do not result in desired functioning or state. Importantly, the ability to achieve a functioning, given a commodity, diverges among people with varying circumstances. For example, individuals may require different resources to move around a building depending on their mobility needs. If everyone is presented with a staircase (*commodity*), their ability to move across floors (*functioning*) may not be the same. We leverage CA to underpin the idea that openness in search should be appraised with respect to its ability to enable desired functionings, and in light of the divergence in capability that it affords to different actors. This perspective is underrepresented in current open search discourse, and constitutes our primary contribution.

We begin by positioning the previously stated motivations for open search as desired *well-beings* within our framework, specifically, achieving digital sovereignty, addressing historically underserved linguistic, regional, and sociocultural needs, encouraging innovation through market competition, promoting open science and rigor in information access research, and realizing information access as a force for justice, emancipation, and democracy. We then track backwards to identify key corresponding *functionings*—a set of ideal 'beings and doings'—that should be afforded by open search to actors to move towards said well-beings, namely to (i) critique, (ii) copy, and (iii) co-opt and co-construct the search system. Next, we enumerate different *commodities* including code, data, models, tooling, compute, labor, and governance that open search may make available. We conclude this section with a discussion of how open commodities afford different *capabilities* to different actors.

Well-beings

The motivations for open search span a range of political, sociocultural, economic, scientific, and societal causes that we map to desired well-beings under CA. These intrinsically different motivations make saliently different demands of openness from search systems. For example, open search is not a strict prerequisite for **digital sovereignty** nor for **addressing historically underserved linguistic, regional, and sociocultural needs**. There are several examples of privately-owned search engines that focus on the needs specific to a language, country, or user demographic, *e.g.*, Baidu (for Chinese), Yandex (for Russian), Naver (for Korean), Parsijoo (for Persian), and KidRex (for kids). However, the path to digital sovereignty or meeting needs of local communities likely require challenging US Big Tech hegemony over search platforms and open search maybe critical to bring together industry, academia, government, and civil society to mount a more concerted offensive. Open search is also important if the goal is not to just address the needs of well-resourced individual states or communities; but to develop replicable strategies and shareable artifacts for those who may lack the adequate resources to build for their own needs.

If the goal is to encourage more **innovation through market competition** then providing open artifacts under appropriate commercial-usage licensing would be valuable. Commercial entities that leverage these artifacts have access to disparate level of resources of their own and therefore will draw different capabilities from the same commodities. *E.g.*, if a high-volume data firehose was made available as part of open search, not all actors may have the necessary infrastructure to consume it. Similarly, if the learned parameters of a massively large model was made open, not every actor will have the compute resources to deploy that model in production settings or finetune it for their specific application scenario. The criticality of these considerations depends on how seriously we want to safeguard against the possible emergence of an oligopoly of actors leveraging these commodities more effectively due to having outsized access to external resources relative to their competitors.

For **promoting open science and rigor in information access research**, open search can enable information access researchers to conduct their studies and test their hypotheses under reproducible real-world deployment settings. However, the additional overhead of working with large-scale production systems may also hinder researchers who may be more interested in access to a simpler prototype of the system that they can easily work with and not have to undergo a steep learning curve. So, the openness that academics may desire from search system are likely different from someone who wants to leverage its openness with commercial intent.

Finally, if our objective is **realizing information access as a force for justice, emancipation, and democracy**, then it is important to ensure equity of capabilities that different communities can draw from the artifacts made open. This is because marginalized communities, who are interested to leverage openness to critique and co-produce technology in support of their struggles, are also likely to be generally resource-poor. Johnson [42] makes a similar critique of the benefit of open data for democracy and justice that is often taken for granted; and instead argues that open data raises serious concerns of information justice. Furthermore, epistemic justice [43] and representational justice [44] may not

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22920264

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

be achievable simply by opening the system to critique and to be copied for adaptation. To challenge the hegemony of western epistemologies and the seeing of the world through the dominant Eurocentric, white, male gaze, we require marginalized communities to also be able to co-opt, deconstruct, and reshape the search system being made open [13].

Functionings

To achieve the aforementioned well-beings relevant to open search, we identify three key functionings within our capability-theoretic framework which we discuss next.

Critiquing. Critiquing refers to the study of sociotechnical systems like search engines to identify potential risks and mechanisms of marginalization and other harms from the system. The object of critique may include the system design (*e.g.*, accessibility of user interface), the materials employed in the production of the system (*e.g.*, datasets) and components of the technology stack (*e.g.*, models), the quality control and evaluation protocols (*e.g.*, metrics for decision-making), the institutional policies and guidelines for system development and operations (*e.g.*, moderation policies and guidelines for data labeling), the organizational processes involved (*e.g.*, red-teaming and system monitoring), the incentive structure that shapes the system (*e.g.*, monetization strategy), and how these aspects interact in complex ways.[3] Critiquing maybe conducted by academics and civil society to safeguard against societal harms, or by regulatory bodies to ensure legal compliance (*e.g.*, antitrust laws, consumer protection laws, data protection laws, and emerging AI regulations); and is emphasized as one of three practices in the emancipatory information access framework [12].

Copying. Copying refers to building alternative systems by leveraging components of an open search. The OWI is a relevant example of this. Its goal is to support the development of a range of new search systems by permitting their use of the open web index it makes publicly available. Such commodities can support the development of alternative systems and infrastructure to challenge the hegemony of big platforms and address unmet community needs in search.

While copying may be a possible path to address the needs of local communities not offered by dominant platforms, it cannot meaningfully mitigate the epistemic injustice and representational harms caused by said platforms. Building separate, distinct search engines does not solve the issue of harmful content, e.g., the sexualization of women of color in search results [45, 46], being proliferated on the dominant platforms if they maintain their monopolistic control over market share. Furthermore, it is also undesirable that every community needs to build a new system to mitigate the erasure of their knowledge or their harmful representation. Even if that is practically and economically feasible, it would produce a range of different systems that neither meaningfully challenge dominant platforms on their own nor benefit from the work of others without massive global coordination. Evidently, there are limitations to the possible gains from being able to simply copy from and make alternatives to dominant systems, particularly when evaluating whether openness satisfies the goal for information access as a force for justice. This leads us to consider the next functioning.

Co-opting and co-constructing. The most radical functioning that open search could enable is to allow for communities to co-opt and co-construct the open search platform itself, which Mitra et al. [13] argue is crucial if our goal is to realize information access as an emancipatory force. In this proposal, communities modify the open search platform to ensure epistemic and representational justice for themselves, and are not restricted to just make derivative systems from it. The platform in turn becomes a site upon which the efforts of several different communities can be put into meaningful arrangements to design information experiences that meet the needs of each. For such arrangements to emerge, the open platform design must hold space for inter- and intra-community negotiations and put in place platform governance mechanisms to ensure specific outcomes, such as foregrounding justice, emancipation, and democracy.

Commodities

The commodities that open search systems cann make available include code, data, models, tooling, compute, labor, and governance. Of these, code, data, models, and tooling constitute artifacts that has traditionally been considered for making open in computing. In the pursuit of open AI models, Widder et al. [38] emphasize the need for also making open the compute and the labor necessary for AI production. For open search, both the compute and the labor necessary for the development and running of search systems should likewise be subject to considerations for opening.

In this context, compute includes the hardware and computational resources necessary for indexing, model training, and online running of the retrieval pipeline; and the human labor necessary for search systems includes data labor for relevance labeling, content moderation, red-teaming, and other data annotations, as well as labor involved in system engineering and operations. Making labor available as a commodity may include allowing critique of working conditions and compensation; providing access to pool of crowdworkers and engineers for derivative system development; and the option to contribute external workforce to the pool. Additionally, we posit that open search should also make available the processes, tools, and other resources for platform governance. Here, labor and governance constitute the *social infrastructure* that are also candidates for opening.

[3] *E.g.*, If the deployment of new ranking models depends on ad revenue from clicks, the system may be at risk of optimizing for the set of users generating the most revenue, who tend to belong to historically privileged demographics, at the cost of worse relevance for marginalized groups; alienating the latter and causing them to abandon the platform; in turn reducing their representation in search log datasets used for training ranking models making them further underserve marginalized peoples; while online success metrics like clickthrough rate falsely indicate higher average user satisfaction because of the exclusion of underserved users.

https://doi.org/10.5281/zenodo.22920264

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

Capabilities

What functionings are afforded by the act of making the system open is determined by a combination of *what* commodities are made open in the process, *how* they are made open, and the circumstances of *who* they are made open to. These factors interact in complex ways and are highly contextual to the specific functionings we want to achieve.

For example, making a search system available as a runnable application would support critique and copy only if the person critiquing or copying have access to necessary compute and data to run the system. Alternatively, components of the system could be exposed as an application programming interface (API) which may be suitable for someone trying to leverage the component in their own system or critique the system outputs without necessitating significant new compute resources; but this is suboptimal if the goal is to scrutinize how the component functions for a system audit. Availability of corresponding documentations, and the quality thereof, may also be instrumental in determining the utility of open commodities. It is also important to consider the granularity of access being granted. For example, when making the code or runnable binaries for indexing available, do we also make available the same for document parsing? Are there external packages or services that indexing employs that should also be made open?

We must also critically interrogate who the system is being made open to and for what permitted purposes. These constraints can impede achieving desired functionings and in turn the desired goals of open search. *E.g.*, search platforms may provide targeted access to selective auditors for regulatory compliance or to earn social goodwill, but prevent academia and civil society from meaningfully scrutinizing the system. This may create false perceptions of openness that reputationally benefit platform owners without resulting in any meaningful public oversight, *e.g.*, [47]. It is crucial for us to also pay attention to restrictions that may be levied on open commodities. *E.g.*, system components may be made available with non-commercial licensing that allow companies to reputationally benefit from having made their system open without endangering their competitive advantage nor contributing to improving market competition. Similarly, copyleft licensing may deter some from adopting open commodities, but such licensing can also be beneficial for soliciting reciprocal openness [48] and safeguarding open search from exploitation by dominant market players.

The CA lens foregrounds these tensions and encourages the open search community to engage in critical dialog and find basis of solidarity across diverse motivations.

CONCLUSION

We instrumentalize capability theory to provide a different vantage point from which to examine open search. We illustrate how different desired states of well-being demands different functionings from opening the system. The open search path for digital sovereignty, for example, need not run parallel to the path for social justice and emancipation; but where they intersect, we have an opportunity to make intentional choices about our future course. This conceptual move intentionally foregrounds the motivations for open search to highlight points of contentions and negotiations for the community to reconcile with. We invite the open search community to engage in these critical reflections to better prepare for the inevitable tensions that all movements for openness must contend with.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge thoughtful feedback from Sireesh Gururaja in the early stages of this work. No external funding was received in support of this work.

REFERENCES

[1] W. Weissert, J. Keaten, and E. Burrows, "Trump's greenland threats spark outrage from eu and test longtime nato alliance," *Associated Press*, 2026.

[2] R. Mancini, "Canada-trump tensions grow after carney 'rupture' speech," *The Hill*, 2026.

[3] T. Noble and K. Walsh, "President trump's war on "woke ai" is a civil liberties nightmare," *Electronic Frontier Foundation Deeplinks Blog*, 2025.

[4] D. Lu, "We tried out deepseek. it worked well, until we asked it about tiananmen square and taiwan," *The Guardian*, 2025.

[5] D. Kerr, "Musk's ai grok bot rants about 'white genocide' in south africa in unrelated chats," *The Guardian*, 2025.

[6] J. Taylor, "Musk's ai firm forced to delete posts praising hitler from grok chatbot," *The Guardian*, 2025.

[7] B. Mitra, H. Cramer, and O. Gurevich, "Sociotechnical implications of generative artificial intelligence for information access," in *Information Access in the Era of Generative AI*, 2024, pp. 161–200.

[8] B. Mitra, "Search and society: Reimagining information access for radical futures," *Information Retrieval Research*, vol. 1, no. 1, pp. 47–92, 2025.

[9] A. Mager, "European search? how to counter-imagine and counteract hegemonic search with european search engine projects," *Big Data & Society*, vol. 10, no. 1, p. 20 539 517 231 163 173, 2023.

[10] V. R. Grassmuck, "Towards an infrastructure for a democratic digital public sphere," *European Public Spheres, Digitisation and Public Welfare Orientation, iRights*, 2021.

[11] C. Plote and A. Nussbaumer, "A charter for public interest internet search," en, *Open Search Symposium 2025*, 2025. doi:10.5281/ZENODO.17251788

[12] B. Mitra, "Emancipatory information retrieval," *Information Retrieval Research*, vol. 1, no. 2, pp. 313–339, 2025.

[13] B. Mitra, N. Neophytou, and S. Gururaja, "Information access of the oppressed: A problem-posing framework for envisioning emancipatory information access platforms," *arXiv preprint arXiv:2601.09600*, 2026.

[14] R. Stallman, *Why "free software" is better than "open source"*, 1998. https://web.archive.org/web/20240929160154/https://www.gnu.org/philosophy/open-source-misses-the-point.html

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22920264

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

[15] E. S. Raymond, "Why microsoft smears-and fears-open source," *IEEE Spectrum*, vol. 38, no. 8, pp. 14–15, 2002.

[16] *Deadly embrace*, 2000. https://www.economist.com/business/2000/03/30/deadly-embrace

[17] A. Lund and M. Zukerfeld, "Profiting from free and open source software," in *Corporate Capitalism's Use of Openness: Profit for Free?* 2020, pp. 109–147.

[18] S. Curtis, *Beware of cloud 'open-washing' says red hat*, 2012. https://web.archive.org/web/20240927121803/https://www.itbusiness.ca/news/beware-of-cloud-open-washing-says-red-hat/17513

[19] D. G. Widder, S. West, and M. Whittaker, "Open (for business): Big tech, concentrated power, and the political economy of open ai," *appear in Nature*, 2023.

[20] Office of Technology of the US Federal Trade Commission, *On open-weights foundation models*, 2024. https://web.archive.org/web/20250730035615/https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/07/open-weights-foundation-models

[21] D. G. Widder, M. Whittaker, and S. M. West, "Why 'open'ai systems are actually closed, and why this matters," *Nature*, vol. 635, no. 8040, pp. 827–833, 2024.

[22] A. Liesenfeld and M. Dingemanse, "Rethinking open source generative ai: Open-washing and the eu ai act," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 1774–1787.

[23] I. Solaiman, "The gradient of generative ai release: Methods and considerations," in *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, 2023, pp. 111–122.

[24] M. White et al., "The model openness framework: Promoting completeness and openness for reproducibility, transparency, and usability in artificial intelligence," *arXiv preprint arXiv:2403.13784*, 2024.

[25] A. Sen, "Commodities and capabilities," *OUP Catalogue*, 1999.

[26] R. Epstein and R. E. Robertson, "The search engine manipulation effect (seme) and its possible impact on the outcomes of elections," *Proceedings of the National Academy of Sciences*, vol. 112, no. 33, E4512–E4521, 2015. doi:10.1073/pnas.1419828112

[27] C. Schyns, "The lobbying ghost in the machine," 2023.

[28] R. Gorwa, G. Lechowski, and D. Schneiß, "Platform lobbying: Policy influence strategies and the eu's digital services act," *Internet Policy Review*, vol. 13, no. 2, pp. 1–26, 2024.

[29] Corporate Europe Observatory, *Article by article, how big tech shaped the eu's roll-back of digital rights*, 2026. https://corporateeurope.org/en/2026/01/article-article-how-big-tech-shaped-eus-roll-back-digital-rights

[30] M. Djuric, "Canada's digital sovereignty push challenges america's reach," *Politico Pro*, 2025.

[31] M. FitzGerald, "'digital sovereignty': Why the eu may be shifting from internet regulation to building homegrown tech," *The Conversation*, 2025.

[32] CERN, *European project to make web search more open and ethical*, 2025. https://home.cern/news/news/computing/european-project-make-web-search-more-open-and-ethical

[33] Internet Society, *Policy brief: Open internet standards*, 2015. https://wayback.archive-it.org/10101/20251031172008/https://www.internetsociety.org/policybriefs/openstandards/

[34] D. Shepardson and M. Scarcella, *Google has an illegal monopoly on search, us judge finds*, Accessed: 2026-03-01, 2024. https://www.reuters.com/legal/us-judge-rules-google-broke-antitrust-law-search-case-2024-08-05/

[35] US Department of Justice. "Department of justice prevails in landmark antitrust case against google." Press release, https://www.justice.gov/opa/pr/department-justice-prevails-landmark-antitrust-case-against-google

[36] D. Milmo and L. O'Carroll, "Eu investigates apple, meta and google owner alphabet under new tech law," *The Guardian*, 25, 2024. https://www.theguardian.com/business/2024/mar/25/eu-investigates-apple-meta-google-alphabet-digital-markets-act

[37] A. Ulea, "Eu takes on big tech: Here are the top actions regulators have taken in 2025," *Euronews*, 2025. https://www.euronews.com/next/2025/12/17/eu-takes-on-big-tech-here-are-the-top-actions-regulators-have-taken-in-2025

[38] D. G. Widder, M. Whittaker, and S. Myers West, "Why 'open' ai systems are actually closed, and why this matters," *Nature*, vol. 635, pp. 827–833, 2024. doi:https://doi.org/10.1038/s41586-024-08141-1

[39] J. R. Trippas, J. S. Culpepper, et al., "Report from the fourth strategic workshop on information retrieval in lorne (swirl 2025)," in *ACM SIGIR Forum*, vol. 59, 2025, p. 68.

[40] *The facebook files*. https://www.wsj.com/articles/the-facebook-files-11631713039

[41] R. Saith, "Capabilities: The concept and its operationalisation," p. 8, 2001.

[42] J. A. Johnson, "From open data to information justice," *Ethics and Information Technology*, vol. 16, pp. 263–274, 2014.

[43] M. Fricker, *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press, 2007.

[44] S. Lewis, "Vision & justice," *Aperture*, vol. 223, pp. 11–14, 2016.

[45] S. U. Noble, "Algorithms of oppression," in *Algorithms of oppression*, 2018.

[46] A. Urman and M. Makhortykh, ""foreign beauties want to meet you": The sexualization of women in google's organic and sponsored text search results," *new media & society*, vol. 26, no. 5, pp. 2932–2953, 2024.

[47] O. Solon, "While facebook works to create an oversight board, industry experts formed their own," *NBC News*, 2020.

[48] D. A. Frantsvog, "All rights reversed: A study of copyleft, open-source, and open-content licensing.," *Contemporary Issues In Education Research*, vol. 5, no. 1, pp. 15–22, 2012.

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22905394

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

QUO VADIS DIGITAL REGULATION? AN OVERVIEW OF THE DIGITAL REGULATORY LANDSCAPE IN EUROPE AND WHAT IT MEANS FOR THE EUROPEAN WEB DATA INFRASTRUCTURE*

C. Werther, LL.M.†, Open Search Foundation e.V., Munich, Germany

Abstract

Over the last decade, the European Union has established an extensive regulatory framework governing digital technologies, data use, platform governance, cybersecurity, and artificial intelligence. While this framework aims to protect fundamental rights and ensure fair competition, it has also produced a high level of complexity. In response, the European Commission started stress-testing its regulation and has inter alia proposed the so-called Digital Omnibus. The Digital Omnibus is a legislative package designed to simplify, consolidate, and better align existing digital regulations. This paper provides an accessible overview of the European digital policy initiatives (as of February 2026). It explains the rationale behind the omnibus approach, outlines the structure and content of the proposals, and discusses their respective implications for the European Web Data Infrastructure. The paper furthermore gives an overview of other launched and upcoming EU initiatives in the field of digital regulation.

INTRODUCTION

The European Union (EU) has positioned itself as a frontrunner in digital regulation. Through instruments such as the General Data Protection Regulation (Regulation (EU) 2016/679 - GDPR), the Digital Services Act (Regulation (EU) 2022/2065 - DSA), the Digital Markets Act (Regulation (EU) 2022/1925 - DMA), the Data Act (Regulation (EU) 2023/2854 – Data Act) and the Artificial Intelligence Act (Regulation (EU) 2024/1689 - AI Act), the EU seeks to ensure that digital transformation respects fundamental rights, market fairness, and societal values. The EU's approach is based on a human-centric vision [1] which forms the basis of the European digital strategy, the so-called Digital Decade Policy Programme 2030 (DDPP) [2].

The European approach reflects a deliberate choice to regulate digital technologies proactively rather than relying solely on market forces. This approach is a reaction to the evolvement of Big Tech and Europe's dependency on overseas' digital services. However, as the number of digital laws has increased, so has the complexity of the regulatory environment. Companies developing digital services, data-driven applications, or AI systems must often comply simultaneously with multiple legal instruments that overlap in scope, terminology, and reporting requirements. Navigating this landscape has become increasingly challenging.

Pursuant to the so-called Draghi Report "*digital companies are deterred from doing business across the EU via subsidiaries, as they face heterogeneous requirements, a proliferation of regulatory agencies and "gold plating" of EU legislation by national authorities.*" [3, (p. 30)]

This also applies to the European Web Data Infrastructure: Copyright law, data protection and liability aspects complicate the development, operation and use of the European Web Data Infrastructure, especially for commercial purposes. The Web content being collected and shared includes personal information, copyrighted content, and possibly illegal material. [4]

In response, the EU seeks to simplify, clarify and solidify the effectiveness of the rules and their enforcement to bring relief to businesses, public administrations, and citizens alike. [5, 6 (p.7)] This shall be done by stress-testing, reviewing, and revising the digital regulatory landscape. The measures already taken as well as further planned initiatives shall be examined in this paper.

THE EU DIGITAL REGULATORY LANDSCAPE

European digital policy is built around several key legal instruments. The GDPR regulates the processing of personal data and applies horizontally across all sectors. The DSA and DMA govern online platforms, addressing systemic risks, content moderation, and market power. The Data Act and related data legislation aim to facilitate access to and sharing of non-personal data, particularly in business-to-business and business-to-government contexts. The AI Act introduces a risk-based framework for artificial intelligence systems, imposing stricter obligations on applications considered high-risk. Together, these instruments form what is described as the EU digital rulebook. [5]

While the respective regulations target specific policy problems, they frequently apply to the same technical systems and organisational processes. As a consequence, the accumulation of digital laws produces overlaps, fragmentation and inconsistencies. [3, (p. 69)] Various regulations include concepts, such as risk management, transparency, or incident reporting but using different definitions and/or determining divergent responsibilities and regulatory requirements.

* Based on research project Privacy-enhancing digital infrastructures (PriDI), funded by the German Federal Ministry of Research, Technology and Space (BMFTS). The responsibility for the content of this publication lies with the author.

† cw@opensearchfoundation.org

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22905394

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

THE CONCEPT OF THE DIGITAL OMNIBUS

With the aim to give a boost to EU competitiveness by enabling innovation and development in the digital sphere the European Commission (EC) decided to simplify digital rules and proposed the so-called Digital Omnibus on 19 November 2025. [7] The proposal now undergoes legislative process.

The term “Omnibus” possibly derives from the Latin term “Omnibus” meaning “for all”. [8] An Omnibus is a legislative proposal that amends several existing laws. Rather than introducing new policy objectives, the Omnibus is meant to focus on technical adjustments, consolidation, and alignment.

The Digital Omnibus proposed by the EC is seen as a first step to optimise the application of the digital rulebook. Its immediate objective is to reduce cost for compliance with the rules. [9]

The Digital Omnibus is accompanied by the so-called Data Union Strategy and the EU Business Wallets. The Data Union Strategy aims to unlock more high-quality data for AI. The EU Business Wallets have the objective to cut paperwork and ease regulatory burdens. [10] This paper will focus on the Digital Omnibus and the Data Union Strategy.

WHAT IS IN THE DIGITAL OMNIBUS?

The Digital Omnibus’ objective is to simplify rules on data, cybersecurity and AI. This includes (i) better and simpler data access, (ii) streamlined cybersecurity reporting and (iii) innovation-friendly AI rules. [10] It consists of two draft regulations: (i) a general digital omnibus addressing data, cybersecurity, and privacy rules, and (ii) an AI-specific omnibus amending the AI Act.

Both drafts are intended to take the form of EU regulations. As regulations once adopted the Omnibus package would apply directly in all Member states pursuant to Article 288 Treaty on the Functioning of the European Union (TFEU).

Notwithstanding the foregoing, it will certainly take some time until the regulations will apply. The proposals are currently undergoing legislative procedure as they require approval from the European Parliament and the Council of the European Union. Several aspects of the proposed Omnibus package have already sparked considerable debate, and it is to be expected that the Parliament and the Council of the European Union will thrive for substantial changes to the proposal. This particularly applies to privacy rules and AI regulation. [11 - 15] Therefore, it is likely that the present draft will undergo changes before being adopted. At the time of writing, it is therefore not yet possible to definitively assess the implications of the Digital Omnibus for privacy and innovation.

General Digital Omnibus addressing data, cybersecurity, and privacy rules [16]

The general Digital Omnibus covers various topics including data law, cybersecurity, privacy, and cookie rules.

Data Law

One central element of the Digital Omnibus is the consolidation of EU data law with the goal of reducing the number of relevant directives and regulations from four to only one. Relevant provisions included in the Data Governance Act (Regulation (EU) 2022/868), the Free Flow of Non-Personal Data Regulation (Regulation (EU) 2018/1807), and the Open Data Directive (Directive (EU) 2019/1024) shall be integrated into the Data Act. The former acts shall be repealed.

The proposed amendments have the goal to create one coherent legal instrument for (i) data access, (ii) data use, (iii) data intermediation, (iv) data altruism, and (v) reuse of public data. [14]

Cybersecurity

In the context of cybersecurity, the Digital Omnibus proposes streamlined cybersecurity reporting.

The EC wants to establish the European Union Agency for Cybersecurity (ENISA) as the Union-wide central reporting agency. ENISA shall serve as the single-entry point where companies could meet all incident-reporting obligations pursuant to for example GDPR, NIS2 (Directive (EU) 2022/2555), and DORA (Regulation (EU) 2022/2554).

Privacy

Although the EC has repeatedly emphasized its intentions to keep intact the core of the GDPR and to maintain the highest level of personal data protection, [7, 17, 18] probably the most controversial discussions around the Omnibus concern the proposed amendments of the GDPR.

The Digital Omnibus aims to clarify definitions, such as “personal data” and “scientific research”. It also includes proposals regarding the use of personal data in the context of AI and for revising the cookie rules.

Definition of Personal Data

Especially the amendment of the definition “personal data” is subject of intense debates. In Article 4 No. 1 GDPR personal data is currently defined as “a*ny information relating to an identified or identifiable natural person*”. Pursuant to Article 4 No. 1 GDPR an identifiable natural person is one who can be identified, directly or indirectly. Under the current regime uncertainties exist especially regarding the scope of indirect identification.

Taking into account the European Court of Justice’s case law [19, 20] the Digital Omnibus proposes that “*information shall not be personal for a given entity where that entity cannot identify the natural person to whom the information relates, taking into account the means reasonably likely to be used by that entity.*“

https://doi.org/10.5281/zenodo.22905394

Pursuant to the Omnibus proposal only the specific data handling entity and its own possibilities to identify a data subject shall be relevant. Whether a potential recipient or another entity is able to identify that person, shall be irrelevant for determining personal data.

The definition of "personal data" is highly important as it serves as the entry point into the GDPR. For data that is not "personal data" in the sense of the GDPR the GDPR does not apply. Handling non-personal data is way easier and comes with hardly any restrictions.

Pursuant to a leaked presidency compromise text issued by the Council of the European Union it is to be expected that the member states will reject the proposed definition of "personal data". [21] The fear is, that the proposed wording facilitates misuse and circumvention and thereby undermines protection of personal data and privacy.

Definition of Scientific Research

The Digital Omnibus furthermore seeks to include a definition of "scientific research" into the GDPR. This is because the GDPR provides for specific regulations and exemptions for research but does not yet define what exactly falls under "scientific research".

The Omnibus proposes to define "scientific research" as "*any research which can also support innovation, such as technological development and demonstration.*" In recital 28 it furthermore lists criteria which may be used for assessing whether research meets the conditions of scientific research. These criteria include for example the methodological and systematic approach applied. Furthermore, research and technology development should be conducted in "*academic, industry and other settings, including small and medium-sized undertakings, and should be always of a of high quality and should adhere to the principles of reliability, honesty, respect and accountability (verifiability)*".

The wording of the definition including the criteria listed in recital 28 suggests that industrial research would also benefit from the privileges defined in the GDPR for scientific research. That industrial research may not in any case be of scientific character however is not addressed.

Personal Data in the Context of AI

The Digital Omnibus also aims to facilitate the use of personal data in AI training, development and operation. For that purpose, it proposes a legal basis for using personal data in the context of AI.

The Digital Omnibus suggests enabling the use of special categories of personal data in the context of development and operation of an AI system or an AI model. The processing of biometrical data shall be possible where this is necessary for verification purposes.

Furthermore, for developing and operating an AI system or an AI model processing personal data may be pursued for legitimate interest where this is necessary for the development and operation.

In any case, the use of personal data in the context of AI shall be subject to specific safeguarding measures.

With establishing a legal basis for using personal data as well as special categories of personal data for AI development the EC seeks to boost innovation in the field of AI. However, the proposal leaves open follow-up questions, for example what happens in case of an objection to use personal data in the context of AI. It furthermore remains to be seen whether the safeguarding measures suffice to prevent misuse.

Cookie Rules

Another aspect in this part of the Digital Omnibus is the simplification of the rules regarding cookie-banner.

The number of cookie banners shall be reduced in order to counter so-called "cookie-fatigue". Users shall be able to indicate their consent and set preferences with one-click. Their choice shall then be saved through central settings in browsers.

Digital Omnibus on AI [22]

The second part of the Digital Omnibus concerns the AI Act.

The AI-specific Omnibus proposes to link high-risk AI implementation to the availability of harmonised standards, common specifications and guidelines to allow preparation for compliance. According to the Omnibus proposal, the rules shall start applying once such standards and support tools are available. This is not yet the case.

High-risk AI is AI that can pose serious risks to health, safety or fundamental rights. High-risk AI systems are therefore subject to strict obligations before they can be put on the market. These include inter alia (i) risk assessment and mitigation systems, (ii) training with high-quality data sets only, (iii) traceability and documentation, (iv) adequate information, (v) human oversight, and (vi) high level of robustness, cybersecurity, and accuracy.

The AI-specific Omnibus also seeks to extend simplifications granted to SMEs under the current regime to small mid cap companies (SMCs). This would for example include technical documentation requirements.

Regarding the Digital Omnibus on AI, EU legislators are put under significant pressure to reach an agreement. Under the AI Act currently applicable the regime for high-risk AI comes into force in August 2026. To finish the legislative procedure before the regime for high-risk AI comes into force, the European Parliament must finalize its draft report before April and be able to come to a quick deal with the EU member states. [23]

What does the Digital Omnibus mean for the European Web Data Infrastructure?

As said, the definition of "personal data" is highly relevant, also for building, developing and using the Web Data Infrastructure. Depending on whether data crawled, processed, and indexed is deemed personal data or not,

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

https://doi.org/10.5281/zenodo.22905394

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

the GDPR may or may not apply for the respective data handling.

The same applies with regard to the proposal for defining “scientific research”. When building, operating, and/or using the Infrastructure falls under that definition, data handling may benefit from exemptions stipulated in the GDPR for scientific research.

The proposal of the Digital Omnibus to allow the use of personal data in the context of AI could lead to more clarity with regard to building the Web Data Infrastructure.

When using AI for building the Web Data Infrastructure one may rely on legitimate interest when using personal data. The same would possibly apply where the Web Data Infrastructure is used for the development and/or training of AI systems or AI models.

Amendments with regard to cookie rules will likely mainly concern applications built on the Web Data Infrastructure.

However, the proposal of the Digital Omnibus is far from being final and it is foreseeable that the Parliament and the Council resist some changes proposed by the Commission, particularly those impacting the GDPR. Whether or not the Digital Omnibus in the end really brings more clarity and legal certainty to building and using a European Web Data Infrastructure therefore remains to be seen.

DATA UNION STRATEGY

With the aim to scale up access to high quality data for AI, to protect Europe’s data sovereignty, and to streamline and simplify data rules [10], the digital package proposed in November 2025 includes a new Data Union Strategy.

The Data Union Strategy seeks to expand access to high-quality data, such as data labs. The strategy inter alia proposes to make digitised cultural objects available for AI training and to boost the use of synthetic data.

For making the Data Act easier to understand and navigate, the EC plans to publish guiding documents such as (i) model contractual terms for data access and use and (ii) standard contractual clauses for cloud computing contracts. The to be established Data Act Legal Helpdesk shall support the implementation of the Data Act. For safeguarding Europe’s data sovereignty, the EC aims to issue guidelines to assess fair treatment of EU data abroad and to create a toolbox to counter unjustified localisation, exclusion, data leakage and adoption measures to protect sensitive non-personal data. [24]

Access to high quality data is essential. It is therefore a positive sign that the EU seeks to facilitate such access. However, it is important to make sure, that such high-quality data also includes web data, as provided by the European Web Data Infrastructure.

BROADER STRESS TESTING

The Digital Omnibus and the Data Union Strategy are part of a broader stress-testing of the digital rulebook. Besides the Digital Omnibus the EU has inter alia conducted the following initiatives:

Review of Europe’s Digital Decade Policy Programme 2030

The European Commission is required to review the DDPP by 30 June 2026. It shall assess whether the DDPP and its implementation require changes in the light of technical, economic or societal developments and the evolution of EU priorities. To this end, the EC initiated a Call for Evidence, concluding it on 9 January 2026.

The DDPP sets up a monitoring and cooperation mechanism to achieve the common objectives and targets for Europe's digital transformation. Skills, Infrastructure, Business and Government are the four cornerstones of the DDPP that define its targets. Furthermore, the DDPP sets several objectives to ensure the digital transformation respects EU values and benefits all EU citizens. These include among others (i) building a safe and secure digital world, (ii) ensuring everyone can participate in digital opportunities, (iii) ensuring small businesses and industry can access data and adopt digital technologies, (iv) ensuring SMEs can compete in the digital world on fair terms, and (v) promoting the deployment of innovative infrastructures. [2]

After reviewing the feedback received, adoption is planned for Q2 2026. [25]

The DDPP is the basis for the EU’s digital strategy. Having said that, it certainly has an impact on the European Web Data Infrastructure. The EU and its initiatives in the digital sphere will be measured against its digital strategy. Therefore, whether the Infrastructure can and will prevail, also depends on how serious the EU takes its DDPP.

Joint Guidelines on the Interplay between Digital Markets Act and General Data Protection Regulation

On 9 October 2025 the European Data Protection Board (EDPB) and the EC jointly published a first version of Guidelines on the Interplay between the DMA and the GDPR. [26]

The EDPB is an independent European body established in accordance with Article 68 GDPR. It inter alia issues opinions, recommendations, guidelines, and best practices to promote a common understanding of the GDPR. The EDPB furthermore advises the EC on issues related to data protection.

The published guidelines seek to clarify common touchpoints, to facilitate the coherent application of the DMA and GDPR, and to increase legal certainty for gatekeepers, business users, beneficiaries and individuals. [27]

https://doi.org/10.5281/zenodo.22905394

The Guidelines address topics such as end-user choice and consent, Article 5 (2) DMA, the right of data portability of end-users and third parties authorised by end-users, Article 6 (9) DMA, the right to data access of business users and authorised third parties, Article 6 (10) DMA and - with regard to web search probably the most relevant topic – access to anonymised ranking, query, click and view data pursuant to Article 6 (11) DMA. In this regard the guidelines seek to clarify how the data needs to be shared to ensure usability, considering the need for anonymisation and/or pseudonymisation.

The first version of the guidelines is now under review after a public consultation.

In addition to publishing the guidelines, the EC opened so-called specification proceedings pursuant to Article 8 (2) DMA against Google in January 2026. The proceedings shall specify the obligations under Article 6 (11) DMA. The goal is to define the scope of data that needs to be shared, the anonymisation method, the conditions of access, and the eligibility of AI chatbot providers to access the data. [28] In this regard the outcome of the specification proceedings may complement the Joint Guidelines on the Interplay between the DMA and the GDPR.

Access to for example Google's ranking, query, click and view data can be essential for developers of alternative search engines. Therefore, the envisaged guidelines and - probably more importantly - the outcome of the specification proceedings are of interest for web search application developers and operators.

Guidelines on the Interplay between Digital Services Act and the General Data Protection Regulation

In September 2025 the EDPB has adopted Guidelines on the Interplay between the DSA and the GDPR.

As the DSA contains several provisions which relate to the GDPR, the guidelines shall help to understand how the GDPR should be applied in the context of the DSA. This especially applies to notice-and-take down systems, recommender systems, transparency of advertising and the prohibition of profiling-based advertising. [29, 30]

The aforementioned systems may also be relevant in the context of the European Web Data Infrastructure and/or applications built on such Infrastructure. Therefore, the guidelines may also be of use for the infrastructure and application developers and operators.

Amendment EuroHPC Regulation

On 20 January 2026 the amendment to the EuroHPC Joint Undertaking regulation (Regulation (EU) 2026/150) entered into force. This regulation amends Regulation (EU) 2021/1173 on establishing the European High Performance Computing Joint Undertaking (EuroHPC JU). The revised framework provides a legal and operational basis for the establishment of AI Gigafactories across the EU. The amendment also allows EuroHPC JU to further develop Europe's quantum technology capabilities. [31]

In light of the envisaged implementation of the European Web Data Infrastructure into the EuroHPC ecosystem, the establishment of the AI Gigafactories is a positive sign.

Resolution on European technological sovereignty and digital infrastructure

The European Parliament on 22 January 2026 adopted a resolution on European technological sovereignty and digital infrastructure. [32, 33]

The European Parliament uses resolutions to express its opinion on a topic it considers important. Resolutions are mainly issued in areas where the Parliament has limited rights of participation.

The resolution inter alia underlines that European sovereignty is "*the ability to build capacity, resilience and security by reducing strategic dependencies, preventing reliance on foreign actors and single service providers, and safeguarding critical technologies and infrastructure; (…)*". Pursuant to the resolution, technological sovereignty is defined as "*the capacity to design, develop and scale up digital technologies needed for the competitiveness of our economy, the welfare of our citizens and the EU's open strategic autonomy in a globalised world; (…)*".

With regard to regulation, the resolution emphasizes "t*he need to foster a supportive regulatory environment that encourages innovation, investment and the development of cutting-edge technologies in Europe, while protecting EU end users from the consequences of extraterritoriality; (…)*".

Furthermore, the resolution contains sections outlining the Parliament's positions and understandings concerning inter alia (i) digital public infrastructure, (ii) digital infrastructure, (iii) high-speed connectivity, (iv) high-performance computing (HPC) systems, (v) AI systems, (vi) data centres, (vii) cybersecurity, (viii) simplification, (ix) research and innovation.

It is encouraging that the Parliament deems European technological sovereignty and digital infrastructure important. However, whether the resolution will indeed foster technological and digital sovereignty and for example facilitates building, operating, and using a European Web Data Infrastructure as a cornerstone of sovereignty remains to be seen. This mainly depends on the actions taken by the EU and its member states.

Other initiatives to make regulation better

Besides simplifying, clarifying, and streamlining existing rules with inter alia the Digital Omnibus, the EU has further topics on its agenda that shall contribute to making regulation in Europe better. These measures for example include a so-called Digital Fitness Check, a review of the Copyright Directive, and the European Open Digital Ecosystem Strategy.

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

https://doi.org/10.5281/zenodo.22905394

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

Digital Fitness Check

With its Digital Fitness Check the EC aims to evaluate whether the overall regulatory framework remains coherent, proportionate and fit for purpose in a fast-evolving digital environment. The Digital Fitness Check serves as a response to the competitiveness concerns highlighted for example in the Draghi Report on the future of European competitiveness. [3]

Whereas the Digital Omnibus proposes targeted amendments and consolidation measures, the Digital Fitness Check shall focus on broader systemic questions and evaluate how the digital rulebook functions as an integrated whole.

The EC has opened a Call for Evidence asking for feedback until 11 March 2026. Adoption of the feedback received is planned for Q1 2027. [34]

If done correctly, a Fitness Check can contribute to a clearer and more coherent legal framework, which would be to the benefit of market players, public administrations and citizens alike.

Copyright

Pursuant to Article 30 Directive on copyright and related rights in the Digital Single Market (Directive (EU) 2019/790 – Copyright Directive) the EC shall carry out a review of the directive and present a report on the main findings no sooner than June 2026. Pursuant to the 2025 Annual Progress Report the Copyright Directive shall undergo review in 2026. [35 (p. 4)]

The Legal Affairs Committee of the European Parliament has already addressed copyright issues. It adopted a series of proposals to ensure full transparency and fair remuneration of rightsholders for the use of copyrighted work by generative artificial intelligence. With the proposals, the Committee seeks to protect the news and media sector as well as individual rights. The proposal includes a call for the establishment of voluntary collective licensing agreements as well as tools allowing rightsholders to prevent their work from being used by general-purpose AI systems. [36, 37] The Proposal is currently awaiting Parliament's vote.

The envisaged review of the Copyright Directive may be an option to clarify the use of copyrighted work in the context of building, operating, and using a European Web Data Infrastructure. It will hopefully also contribute to clarifying the use of copyrighted work in the context of AI.

European Open Digital Ecosystem Strategy

With the European Open Digital Ecosystem Strategy, the EC aims to set out a strategic approach to the open source sector in the EU, addressing the importance of open-source inter alia for technical sovereignty. Furthermore, the EC seeks to establish a comprehensive framework supporting the entire open-source lifecycle from development to market integration.

For that purpose, the EC conducted a Call for Evidence on Open-Source Digital Ecosystems with the goal to identify barriers to open-source adoption, demonstrate the added value of open source and share suggestions for concrete EU level measures to strengthen the ecosystem.

After reviewing the feedback received the EC plans adoption in Q1 2026. [38]

It is a positive sign that Europe is working on an Open Digital Ecosystem Strategy. The open-source sector is an important building block for technological and digital sovereignty. Strengthening open source in the EU would also help the European Web Data Infrastructure as it is inter alia built on the principles of openness and transparency.

Review of the Digital Markets Act

By May 2026 the EC has to conduct a review of the DMA. The review shall assess how the rules are contributing to the objective of ensuring contestable and fair markets and their impact on business and end users. The review should also assess whether there is a need to take further measures to ensure the effectiveness of the DMA.

For that purpose, the EC conducted a Call for Evidence. A summary to the consultation was published on 8 January 2026. Besides a broad support for the DMA and its goals, the feedback highlights the need for effective, well-resourced, predictable, rigorous, transparent and timely enforcement. Additionally, the consultation brought to light a call to expand and future-proof the DMA's scope, especially regarding cloud and AI. [39]

On 3 May 2026 the EC will present the EC's review report. [40] Adoption is planned for Q2 2026. [41]

Effective and timely enforcement of the DMA should be ensured after reviewing the DMA. Enforcement of the DMA is essential for enabling alternative digital applications and services to enter the market and to foster digital and technological sovereignty in Europe. This would also be for the benefit of the European Web Data Infrastructure and services built on it, for example with regard to accessing anonymised ranking, query, click and view data.

Joint Guidelines on the Interplay between the AI Act and EU data protection laws

The EDPB is working on Joint Guidelines on the Interplay between the AI Act and the EU data protection laws.

In the context of AI, it is necessary to clarify the use of personal data. A balance must be struck between the need for innovation and the protection of fundamental rights and privacy. Hopefully, the envisaged guidelines will be able to shed light into the darkness. With regard to using AI for building the European Web Data Infrastructure such guidelines could be helpful.

https://doi.org/10.5281/zenodo.22905394

CONCLUSION

The regulatory landscape in Europe remains dynamic and complex.

The digital rulebook is under pressure and the EC, the European Parliament, the Council, industry leaders and civil society are wrangling over the right level of regulation. The numerous initiatives underline that the EU strives for simplification and clarification to boost innovation and to foster digital and technological sovereignty.

It seems clear that European digital regulation needs an update to facilitate digital innovation and development. The complexity of European regulations poses a barrier for market entry for innovative European digital services and business models, including a European Web Data Infrastructure and services built upon. At the same time, Europe needs to ensure protection of fundamental rights and privacy and must prevent misuse of simplifications and clarifications to the detriment of personal data subjects. Fundamental digital rights and European values must be strengthened and protected.

This is a balancing act, that is certainly not easy.

With regard to the European Web Data Infrastructure, key legal aspects that still need clarification are related to publicly available Web content being systematically downloaded, processed and shared through the European Web Data Infrastructure and later on being used, especially for commercial purposes. [4]

Copyright and data protection issues may be - at least partially - clarified with the Digital Omnibus and the review of the Copyright Directive. A clarification of liability issues arising from building and using a European Web Data Infrastructure however does not seem to be on the table at this point. A solution would most likely require a clarification in the DSA that a European Web Data Infrastructure falls under the liability exemptions for intermediary services. Furthermore, making crawled data available to third parties should be treated as a neutral data intermediary activity, excluding liability for content. [4]

REFERENCES

[1] European Commission https://digital-strategy.ec.europa.eu/en/policies/europes-digital-decade

[2] European Commission https://digital-strategy.ec.europa.eu/en/policies/digital-decade-policy-programme

[3] M. Draghi, “The future of European competitiveness Part A | A competitiveness strategy for Europe” (“Draghi Report”), September 2024, https://commission.europa.eu/document/download/97e481fd-2dc3-412d-be4c-f152a8232961_en?filename=The%20future%20of%20European%20competitiveness%20_%20A%20competitiveness%20strategy%20for%20Europe.pdf

[4] European Web Search Initiative Policy Paper “Boosting digital sovereignty and competitiveness through a European Web Data Infrastructure”, November 2025 https://openwebsearch.eu/wp-content/uploads/2025/12/251130_WebDataInfrastructure-Policy-Paper.pdf

[5] European Commission https://digital-strategy.ec.europa.eu/en/policies/digital-rulebook

[6] European Data Protection Board “EDPB-EDPS Joint Opinion 2/2026 On the Proposal for a Regulation as regards the simplification of the digital legislative framework (Digital Omnibus), 10 February 2026 https://www.edpb.europa.eu/system/files/2026-02/edpb_edps_jointopinion_202602_digitalomnibus_en.pdf

[7] European Commission https://ec.europa.eu/commission/presscorner/detail/en/speech_25_2732

[8] S. Wettach, “Das verbirgt sich hinter den Omnibussen der EU” published in Das Parlament, 02.02.2026, Berlin, https://www.das-parlament.de/aussen/europa/das-verbirgt-sich-hinter-den-omnibussen-der-eu

[9] European Commission, https://digital-strategy.ec.europa.eu/en/library/digital-omnibus-regulation-proposal

[10] European Commission, Factsheet: Digital Package, 19 November 2025 https://ec.europa.eu/commission/presscorner/api/files/attachment/881887/Factsheet%20Digital%20Package.pdf

[11] D. Leisegang, “EU-Kommission will Datenschutzgrundverordnung und KI-Regulierung schleifen”, 07 November 2025, https://netzpolitik.org/2025/digitaler-omnibus-eu-kommission-will-datenschutzgrundverordnung-und-ki-regulierung-schleifen/

[12] European Digital Rights (EDRi), “Reopening GDPR and ePrivacy through the Digital Omnibus: a risky path for EU digital rights“, 11 February 2026 https://edri.org/our-work/reopening-gdpr-and-eprivacy-through-the-digital-omnibus-a-risky-path-for-eu-digital-rights/

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

https://doi.org/10.5281/zenodo.22905394

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

[13] European Data Protection Board “EDPB-EDPS Joint Opinion 2/2026 On the Proposal for a Regulation as regards the simplification of the digital legislative framework, (Digital Omnibus), 10.02.2026 https://www.edpb.europa.eu/system/files/2026-02/edpb_edps_jointopinion_202602_digitalomnibus_en.pdf

[14] T. Hoeren, “Digital-Omnibus der EU-Kommission” published in MMR-Aktuell 2025, 01501

[15] P. Johannes, C. Geminn, “Digital-Omnibus – Ein Überblick” published in ZD-Aktuell 2025, 01491.

[16] European Commission Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL amending Regulations (EU) 2016/679, (EU) 2018/1724, (EU) 2018/1725, (EU) 2023/2854 and Directives 2002/58/EC, (EU) 2022/2555 and (EU) 2022/2557 as regards the simplification of the digital legislative framework, and repealing Regulations (EU) 2018/1807, (EU) 2019/1150, (EU) 2022/868, and Directive (EU) 2019/1024 (Digital Omnibus) {SWD(2025) 836 final}, COM(2025) 837 final 2025/0360 (COD), 19.11.2025

[17] European Commission https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2718

[18] European Commission https://digital-strategy.ec.europa.eu/en/faqs/digital-package

[19] ECJ, judgement of 19 October 2016 - C-582/14 – (Breyer)

[20] ECJ, judgement of 27 October 2025 - C-413/23 P - (SRB/EDPS)

[21] Council of the European Union Presidency compromise text on Omnibus VII – Digital (GDPR/P2B), 6406 /26, 20 February 2026, Interinstitutional File: 2025/0360 (COD) https://www.euractiv.com/content/uploads/sites/2/2026/02/EURACTIV_OMNIBUS-1.pdf

[22] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI) {SWD(2025) 836 final} 2025/0359 (COD), 19 November 2025

[23] Kai Zenner, “The AI Act Omnibus: Timeline, Key Players, and Documents (February Update)”, 24 February 2026 https://www.kaizenner.eu/post/ai-act-omnibus

[24] European Commission https://digital-strategy.ec.europa.eu/en/policies/data-union

[25] European Commission https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/15514-Review-of-the-Digital-Decade-policy-programme_en

[26] European Data Protection Board “Joint Guidelines on the Interplay between the Digital Markets Act and the General Data Protection Regulation” https://www.edpb.europa.eu/system/files/2025-10/joint_com-edpb_gls_interplay_dma_gdpr_for_public_consultation_en.pdf

[27] European Data Protection Board https://www.edpb.europa.eu/news/news/2025/dma-and-gdpr-edpb-and-european-commission-endorse-joint-guidelines-clarify-common_en

[28] European Commission https://ec.europa.eu/commission/presscorner/detail/en/ip_26_202

[29] European Data Protection Board https://www.edpb.europa.eu/news/news/2025/interplay-between-dsa-and-gdpr-edpb-adopts-guidelines_en

[30] European Data Protection Board “Guidelines 3/2025 on the interplay between the DSA and the GDPR Version 1.1” https://www.edpb.europa.eu/system/files/2025-09/edpb_guidelines_202503_interplay-dsa-gdpr_v1_en.pdf

[31] European Commission https://digital-strategy.ec.europa.eu/en/news/eurohpc-regulation-amended-strengthen-europes-ai-and-quantum-capabilities

[32] European Parliament https://oeil.europarl.europa.eu/oeil/en/document-summary?id=1884418

[33] European Parliament resolution of 22 January 2026 on European technological sovereignty and digital infrastructure (2025/2007(INI)) P10_TA(2026)022 https://www.europarl.europa.eu/doceo/document/TA-10-2026-0022_EN.pdf

[34] European Commission https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/15554-Digital-fitness-check-testing-the-cumulative-impact-of-the-EUs-digital-rules_en

[35] Henna Virkkunen, 2025 Annual Progress Report Simplification, Implementation & Enforcement, September 2025 https://commission.europa.eu/document/download/68237940-a9e3-460c-bf49-932d432a6bba_en?filename=VIRKKUNEN_APR_SPI_2025_70_EN.pdf

[36] European Parliament Legal Affairs Committee, https://www.europarl.europa.eu/news/en/press-room/20260126IPR32636/protect-copyrighted-work-used-by-generative-ai-say-legal-affairs-meps

[37] European Parliament Committee on Legal Affairs, Rapporteur Axel Voss “REPORT on copyright and generative artificial intelligence – opportunities and challenges”, Procedure 2025/2058(INI) https://www.europarl.europa.eu/doceo/document/A-10-2026-0019_EN.html#_section1

[38] European Commission https://digital-strategy.ec.europa.eu/en/news/commission-opens-call-evidence-open-source-digital-ecosystems

[39] DMA Review Summary of the contributions to the targeted consultation, call for evidence and AI consultation https://digital-markets-act.ec.europa.eu/document/download/244d8f93-e969-41af-bdcc-23e791863449_en?filename=Public%20summary%20of%20DMA%20Review%20consultation_0.pdf

[40] European Commission, Daily News 08 / 01 / 2026 https://ec.europa.eu/commission/presscorner/detail/en/mex_26_54

[41] European Commission https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14831-Review-of-the-Digital-Markets-Act_en

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22905603

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

Data protection responsibility of online platforms for user-generated content – recognising sensitive data, preventing scraping and identifying users in light of the ECJ's 'Russmedia' judgement*

P. C. Johannes†, Research Centre for Information Systems Design (ITeG), University of Kassel, Kassel, Germany

Abstract

In its judgment of 2 December 2025, – C-492/23 (X v Russmedia Digital SRL and Inform Media Press SRL) [2], the European Court of Justice (ECJ)) clarified the data protection responsibilities of online platforms. The decision is part of a string of rulings [3-6] that question the formal neutrality of digital intermediaries and focus on their actual role in the dissemination of personal data. This Russmedia-ruling contains four key points, that are explained and contextualized in this paper: Firstly, operators of online marketplaces are generally at least jointly responsible for the advertisements posted by their users in terms of data protection law (see 2). Secondly, operators of online marketplaces cannot invoke the exemptions from liability under Articles 12 to 15 of the e-commerce Directive (now Articles 4 to 6 of the Digital Services Act) (see 3). Thirdly, Article 32 of the General Data Protection Regulation (GDPR) imposes an obligation on operators of online marketplaces to prevent the copying and unlawful redistribution of special categories of data within the meaning of Article 9(1) of the GDPR (see 4). Fourthly, online platforms have an obligation to identify users who wish to publish content (e.g. advertisements on an advertising marketplace) that contains sensitive data within the meaning of Article 9(1) GDPR (see 5). In conclusion, this paper gives an outlook on implementation and design options of the ECJ's findings (see 6) and how the judgement might impact the operation of an Open Web Index (OWI) and search engines (see 7).

1 INTRODUCTION

The initial proceedings of the Russmedia decision were based on a dispute between a victim of defamation and the operators of two Romanian online marketplaces on which classified ads can be published. Specifically, the case concerned the Romanian online marketplace website "publi24.ro". In August 2018, an advertisement was published on this website containing photos and a telephone number of the plaintiff without her consent. The advertisement implied that she was offering sexual services. Although the advertisement was removed after a complaint was lodged, it remained available on other websites run by third parties due to automated copying (i.e. 'scraping'), which led to legal action against the operators. The plaintiff claimed that her personal data had been automatically scraped from these platforms without her consent and disseminated on third-party sites. She demanded that the platform operators take effective measures against the scraping of her data and verify the identity of the advertisers on its site, i.e. those that want to place personal marketplace adds ('classifieds'). The national court in Romania referred several questions to the ECJ regarding the interpretation of Articles 12–15 of Directive 2000/31/EC (now repealed) [7] and the scope of the obligations of platform operators under the GDPR.

2 JOINT RESPONSIBILITY FOR PLATFORM OPERATORS

In its ruling, the ECJ refers to online marketplaces. It clearly does not use the term in a commercial or economic sense, but rather in a functional sense. Based on the initial proceedings, it refers to services that enable users to post advertisements containing, among other things, personal data, which are then systematically structured and made available to the general public (see [2], para. 72). The decisive factor here is not the conclusion of a transaction, but the public disclosure of personal information via a technical environment controlled by the operator. The ECJ emphasises that the platform operator determines the terms and conditions of publication, in particular by specifying requirements for data entry, the presentation of advertisements, their categorisation and the duration of their visibility (see [2] para 55 et. seqq.). According to the ECJ, these controlling elements are sufficient to establish a relevant influence on the means of data processing within the meaning of Article 4 No. 7 GDPR. The operator is thus not merely a technical intermediary, but a structural actor in the dissemination of data. This makes it a (joint) controller within the meaning of the GDPR ([2], para. 55-75). The decision is consistent with the ECJ's previous case law, which has established a broad, function-based interpretation of the term "controller" [3-6]. The fact that the data originates from a third party does not preclude the platform operator from being a controller, provided that the operator has a say in the purposes and means of processing.

Although the case concerned a classified ads portal, the Court of Justice's reasoning can be generalised. The ECJ does not focus on the type of service, but on how it works. The decisive factor is whether the provider, through the design of its platform, creates the conditions under which personal data is publicly disseminated ([2], para 31 et. seqq.). This means that the scope of the ruling extends to virtually all digital services that host and publish user-generated content for their own commercial or institutional purposes. This includes, in particular, job portals, real estate

* This article was written as part of the PriDI project funded by the German BMFTR. It is based on a previously published German-language article by the author [1], presented here in a condensed and revised form.
† paul.johannes@uni-kassel.de

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22905603

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

platforms, review platforms, social networks, online forums, and business directories.

The ruling can apply to all cases, where the operator publishes relevant personal data for commercial or advertising purposes that go beyond the mere provision of a service to the actual user ([2], para. 66). A platform operator's joint or sole responsibility can only be ruled out if they do not use the publication of personal data in user-generated content for their own economic or functional interests but process it exclusively as a necessary by-product of a user-related service. Examples of such services might include 'traditional' hosting service providers that provide internet addresses (domains) and storage space for a fee. This may also include cloud storage or backup services that do not access the data stored with them for their own interest. If a platform operator does not want to be qualified as a controller, it should not structure or monetise the content for themselves. For example, they should not evaluate content in order to place matching advertisements or other content.

3 NO LIABILITY PRIVILEGE FOR DATA PROTECTION

A central part of the judgment is the distinction between responsibility under data protection law and privileges under liability law. The ECJ clarifies that operators of online marketplaces cannot invoke the liability provisions of Articles 12–15 of the e-Commerce Directive to evade their obligations under the GDPR ([2], para. 127 et. seq). According to the ECJ, these provisions concern only civil and criminal liability for third-party content, but not the classification under data protection law and the obligations deriving from it. This interpretation is consistent with the structure and rationale of the Directive itself since recital 14 of the e-Commerce Directive already emphasises that its application is without prejudice to Union law on the protection of personal data. The ECJ takes this up and draws the methodologically consistent conclusion that data protection obligations exist independently of liability privileges under the Directive.

The ruling is also significant in light of the recent change from the e-Commerce Directive to the Digital Services Act (DSA)[8], even though the ECJ does not address the DSA directly at all in Russmedia. The DSA has repealed the e-Commerce Directive. The liability privileges for various types of intermediary services are now set out in Articles 4 to 6 DSA (see [9] for further reference). This is not merely a formal replacement, but a substantive development. Article 2(4)(g) of the DSA makes in no uncertain terms clear that the Regulation applies 'without prejudice' to, among other things, data protection law. Accordingly, recital 10 of the DSA states that the DSA should not affect the GDPR. It further states that the protection of individuals with regard to the processing of personal data is governed 'solely' by the provisions of Union law in this area, in particular by the GDPR. The ECJ's finding that the liability privileges do not apply to data protection obligations will therefore have to apply to the DSA as well (see [10] for more details on the relationship between the GDPR and the DSA).

This is all the more true given that the DSA now imposes obligations on intermediary services. The e-commerce directive was reactive in nature. Recital 47 of the DSA makes it clear that providers are not limited to responding to specific legal violations, but must actively mitigate risks. The DSA establishes a preventive and risk-based regulatory model. See, for example, Article 26 DSA on risk management systems, Article 28 DSA on preventive measures for systemic risks and Article 34 DSA on reporting and accountability obligations. The regulation obliges online platforms and very large online platforms in particular to identify systemic risks to fundamental rights and to take appropriate remedial measures.

This systemic decision also covers risks to the protection of personal data. Article 34(1)(b) DSA explicitly mentions threats to fundamental rights, which include Article 8 'Data protection' of the Charter of Fundamental Rights of the European Union (CFREU). Article 35 DSA then requires the implementation of proportionate risk mitigation measures. The DSA thus reinforces the requirements developed by the ECJ's Russmedia judgement. They should be integrated into a comprehensive platform responsibility concept.

These obligations under DSA correspond to the controller's obligation under Article 25(1) GDPR to promote data protection through technology design and the obligation under Article 32 GDPR to ensure the security of processing. See also [2] para. 89; for an assessment of the proportionality of the obligations under Article 25, 32 GDPR. The ECJ has already ruled in another case that the reference in Article 32(1) and (2) GDPR to a "level of protection appropriate to the risk" and an "appropriate level of protection" shows that the GDPR introduces a risk management system and in no way claims to eliminate the risk of personal data breaches completely [11].

4 PREVENTION OF SCRAPING

4.1 Findings of the ECJ on scraping-prevention

Against this background, the ECJ has found that marketplace operators have an obligation to prevent or limit the copying and redistribution of ads containing special categories of data within the meaning of Article 9(1) of the GDPR (see [2], para 107 et. seqq.). The ECJ's finding that marketplace operators must take measures against copying and unlawful publishing on other websites of their content is particularly significant. The ECJ correctly assigns this obligation to the operator's area of responsibility, as copying (or 'scraping') is a foreseeable consequence of the public provision of personal data ([2], para 120). Dogmatically, this obligation can be based on the principles of processing security under Article 5(1)(f) and Article 32 GDPR as well as Article 25 GDPR. When sensitive data is published online, the controller is obliged under Article 32 GDPR to take all technical and organisational measures to ensure a level of security appropriate to effectively prevent loss of control over this data ([2], para 121).

However, the requirement emphasised by the ECJ to take effective measures against the copying and

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22905603

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

reproduction of online content of (sensitive) personal data is in structural tension with the legitimate business interests of platform operators and the communication interests of their users as well as the public. Online marketplaces, opinion and discussion platforms, and social networks in particular are designed to make content as visible as possible, generate reach, and ensure easy discoverability. This regularly requires open accessibility of content, which is automatically collected, copied and scraped not only by human users but also by search engines and aggregation services. Indexing by search engines is an integral part of the platform ecosystem and often a key prerequisite for economic success and social relevance.[12]. A part of indexing is the copying or scraping of the website in order to analyse its content. A blanket or technically overly restrictive ban on copying or scraping would therefore run the risk of counteracting the purpose of the platform and the intention of the users itself by significantly limiting the visibility, findability and communicative impact of the content.

Against this background, the ECJ cannot be understood to mean that any automated collection of personal data is to be prohibited. Rather, a distinction must be made between legitimate copying acts, like for indexing, and copying for abusive or unlawful secondary redistribution of personal data outside the context of use intended by the platform operator and their users.

4.2 Implementation

The ECJ requires risk-adequate technical and organisational measures in accordance with Articles 25 and 32 of the GDPR, which are intended to prevent or at least hinder unlawful copying and/or the unlawful redistribution of user-generated content on online platforms. It formulates this requirement in a technology-neutral manner. The operator of an online platform cannot completely rid itself of this obligation by just informing the user about the possibility of scraping or redistribution in its terms of use, general terms and conditions or consent information, as the ECJ derives the obligation directly from Articles 25 and 32 GDPR. Nevertheless, normative control mechanisms in the relationship between platform operators and data subjects and third parties are also important for a risk-adequate overall system. It is therefore still advisable to clarify in terms of use or similar documentations that the unauthorized copying and redistribution of personal data is prohibited and that users may only post personal data of third parties that they themselves are authorised to publish. If a user or third party violates such clauses, the platform operator can provide for graduated responses, such as removing the content, blocking the account or access or taking civil action. Such ToS-based mechanisms are regarded in the literature as an integral part of platform governance and also serve to provide legal protection for technical access restrictions [13].

From a technical perspective, platform operators could various methods to make copying/scraping and the unauthorised dissemination or processing of personal data more difficult. None of the measures offer absolute security. But the ECJ does not require absolute security. What is required is an appropriate, risk-oriented level of technical and organisational security within the meaning of Articles 24, 25 and 32 GDPR.

Technical measures for content-based detection of replicated content can form a first level security. Through the use of cryptographic hashing and fingerprinting methods, content can be digitally marked during upload so that identical or highly similar copies on other sites can be automatically detected later more easily. These methods are well known from copyright content control as well as from the fight against non-consensual uploads. If nearly identical content is redistributed on other websites, hash matching enables automated detection and, depending on the platform architecture, the initiation of internal blocking or external takedown processes. In parallel, techniques could be used that are functionally reminiscent of digital rights management systems (DRM) without being DRM in the strict sense. This refers in particular to digital watermarks that are embedded in image or text material to enable subsequent assignment to the original upload. Such markers make mass copying considerably more difficult, as automated extraction techniques are often unable to remove these markings cleanly.

Even more effective is the defence against automated copying/scraping attempts using monitoring and bot detection methods. Rate limiting and behaviour-based crawler detection, IP reputation systems, dynamic captcha mechanisms and heuristic traffic analysis are also established web security measures that are specifically designed to identify, prevent or at least slow down scraping [14, 15].

Finally, effective notification and deletion mechanisms are necessary to limit the continued dissemination of sensitive content on other sites and platforms after it has been removed from one's own platform. This includes marking removed advertisements as deleted to crawlers and search engines – for example, by updating robot metadata or sending targeted takedown requests to third-party services. Although the effectiveness of such measures is limited because completely decentralised data is often no longer fully retrievable, the ECJ expects the establishment of consistent processes that are harmonised with the requirements of Article 17 GDPR.

5 IDENTIFICATION OF USERS AND ACCOUNTABILITY

The ECJ states that it follows from Article 5(2) in conjunction with Articles 24 to 26 GDPR that the operator of an online marketplace must verify whether the advertising user who is about to place an classified ad containing sensitive data within the meaning of Article 9(1) GDPR, is the person whose sensitive data is contained in that ad and, if this is not the case, to refuse to publish it. Unless of course, the user can prove that the data subject has given their consent or can invoke another authorisation pursuant to Article 9(2)(b) to (j) GDPR for the processing/publication of sensitive data of third parties. A prerequisite to this is the obligation to check whether advertisements contain sensitive personal data.

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22905603

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

5.1 Identification of advertisements containing sensitive data

The decision obliges platform operators to adequately identify advertisements that contain sensitive personal data as such ([2], para. 97]. Operators know or should know that advertisements containing sensitive data within the meaning of Article 9(1) GDPR may be published by users on their platforms. Therefore, as the controller, the operator is obliged to take appropriate technical and organisational measures when designing its service to identify such content before it is published. Only in this way is it able to verify whether the sensitive data contained in them is published in accordance with the GDPR. This applies despite the clarification in Article 8 DSA that providers of intermediary services are not subject to a general obligation to monitor the information they transmit or store or to actively investigate circumstances that indicate illegal activity. The content of the provision corresponds to Article 15 of the e-commerce Directive, which the ECJ also considered inapplicable in the Russmedia decision ([2], para. 137 et. seqq.). According to the ECJ, the obligation to comply with the requirements of the GDPR cannot be classified as a general monitoring obligation. The decisive factor in this obligation is, not the identity of the advertiser or content creator, but the content itself, as this is the only way to preventively control the risk of abusive publication. In practice, this requires platforms to try to detect which content contain potentially sensitive information before further review or verification measures take effect.

A risk-adapted compliance system can combine different mechanisms. For example, texts, metadata or uploaded documents can be examined for indications of sensitive data using keyword recognition, semantic analysis or machine learning-based classification methods. Images can also be identified as containing sensitive information through automated pattern recognition or hashing. In addition, categories, keywords or other metadata from the advertisement can be used to flag advertisements with increased risk and place them in a separate audit trail. Voluntary or profile-related information provided by users can also provide clues. With such a flagging system, only advertisements that potentially contain sensitive personal data could be specifically reviewed, while the majority of the remaining content could be published without additional hurdles.

5.2 Identification requirement

The ECJ considers the obligation to identify and verify advertisers to be a consequence of the accountability requirement under Article 5(2) of the GDPR. The platform operator must be able to demonstrate that the publication of personal data is lawful. The ECJ requires a level of verification that is appropriate to the risk ([2], para 97). The greater the reach of the service and the more sensitive the data processed, the higher the identification requirements. Platforms can tailor the effort required for identification and verification to the specific risk situation.

Given that scraping cannot be completely prevented, the requirement for appropriate identification and verification of advertisers and users takes on particular importance. Effective legal protection begins not only with downstream control of distribution channels, but also at the source of data processing itself. The primary risk that the ECJ wants to address with its ruling is the unlawful publication of sensitive personal data by malicious or anonymous users. Identifying users in a manner appropriate to the risk has a twofold effect here: it increases the inhibition threshold for abusive uploads and at the same time enables effective enforcement of (data protection) claims retrospectively. Compared to technically complex and potentially reach-limiting anti-scraping measures, identification therefore appears to be the milder and at the same time more effective means of ensuring the protection of the persons concerned without undermining the fundamental openness and visibility of the platform content. The ruling thus suggests that the focus of platform operators' obligations is less on complete control of the dissemination of publicly accessible content and more on the preventive avoidance of illegal publications at the point of entry itself.

5.3 Design of the identification obligation

However, the ECJ does not require a comprehensive know-your-customer obligation ("KYC obligation") for operators of online marketplaces. Rather, it is crucial that advertisers (or users respectively) are adequately identifiable, with the required level of identification depending in particular on the type of platform, its reach, the sensitivity of the personal data processed and the potential for abuse. For this reason, it also limits the identification requirement to advertisements or content that contain sensitive data within the meaning of Article 9(1) GDPR.

Extending the identification requirement to all personal data, regardless of sensitivity or category, would circumvent the wording of the judgment and no longer be covered by the meaning and purpose of the GDPR's protection. Platforms and users in general have a legitimate interest in the widest possible dissemination of non-sensitive data. An obligation to identify all users could significantly limit the reach and usability of the platform, which would be contrary to the principle of proportionality. However, a more important argument is that the potential for misuse is typically lower for non-sensitive data. The ECJ expressly emphasises that the intensity of the measures must be proportionate to the risk. Blanket identification of all users would contradict this principle.

Rather, the ECJ links the obligation to screen and identify to a qualified risk situation resulting specifically from the processing of particularly sensitive data. In such cases, however, according to the logic of the ruling, neither a purely formal user registration nor a mere self-declaration by the advertiser or user is sufficient. The identification requirement does not serve general platform governance here, but rather the preventive avoidance of serious violations of fundamental rights and the ensuring of effective law enforcement. This also shifts the standard of appropriateness: measures that would be sufficient for non-

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22905603

sensitive data are generally no longer adequate for sensitive data. Against this background, simple identification mechanisms – such as email verification on account registrations – are no longer sufficient for advertisements or content containing sensitive data. They neither have a sufficient deterrent effect on malicious users nor do they allow for reliable ex-post attribution. Even purely intermediate-level procedures such as reputation systems or IP plausibility checks are regularly insufficient because they do not secure the identity of the advertiser with the necessary certainty and only address abuse reactively, if at all. However, such measures have a complementary risk-mitigating effect as protective measures.

Authentication via a telephone number using an SMS code could also have this effect. Particularly in the case of telephone numbers that are only assigned after identity verification in accordance with national legal requirements (e.g. Section 172 TKG in Germany) this measure can already establish an implicit link between the user and their identity and thus have a preventative effect. In cases such as the initial case, such verification could have made it more difficult for malicious users to publish personal data, or only possible with considerable additional effort, thereby reducing the damage to the person concerned.

Strictly speaking, however, this type of identification is also unlikely to be sufficient for advertisements or content containing sensitive data, as it cannot prevent malicious third parties from publishing sensitive data illegally. In the original case, it might not have been possible for the perpetrator to use the plaintiff's telephone number. However, he would still have been able to publish false factual claims. For this reason, more extensive procedures should also be considered that enable reliable assignment of the advertiser or user to the sensitive data to be published. This includes, in particular, temporary ID checks, video identification procedures or the use of state-certified identity services, in particular within the meaning of eIDAS Regulation 910/2014 and Regulation 2024/1183 (see [16] for further references). In future, the European Digital Identity Wallet (EUDI wallet or eID wallet) could be used for user-friendly and system-integrated identification on online platforms ([16], para 18-42). The decisive factor in choosing the type of identification is that the measure has a real deterrent effect depending on the context and at the same time should enable effective law enforcement. Overall, this results in a risk-adaptive standard: platforms must align the effort required for identification and verification with the specific risk situation.

6 IMPLEMENTATION, MONITORING AND HOW THINGS COULD PROCEED

Platform operators are required to implement identification and protection measures not in isolation, but as part of a risk-based platform design. A key factor is an upstream assessment of the risks arising from the publication of certain categories of personal data. On this basis, a decision must be made as to what level of identity verification is necessary and adequate to plausibly ensure the lawfulness of the processing.

At the same time, scraping risks must be systematically addressed. It is therefore necessary to continuously review the effectiveness of existing protection mechanisms and adapt them to technical developments. Documentation and traceability of the measures taken are essential in order to meet the accountability requirements of the GDPR and the supervisory requirements of the DSA.

The ECJ makes it clear that the public dissemination of personal data via digital infrastructures necessarily gives rise to responsibility. With the DSA, this is now even enshrined in applicable law in a preventive, fundamental rights-oriented regulatory framework at Union level. Platform operators will therefore have to be measured against an integrated understanding of data protection and platform responsibility.

However, this is neither automatic nor inevitable, but rather the result of the stringent, fundamental rights-promoting application of the secondary legal acts GDPR and e-commerce Directive (or now DSA) by the ECJ. With regards to the further development of the law, it should not be forgotten that platform operators, both small and large, are also bearers of fundamental rights. With this in mind, the EU legislator should try to facilitate the practical implementation of the requirements formulated in the ECJ ruling and at the same time create legal certainty for platform operators and data subjects by carefully adapting DSA and/or GDPR. Legal clarification regarding responsibility in relation to third-party platforms on which content is redistributed after scraping would be desirable. More precise regulation of the exemption from liability for hosting providers, defining the limits of responsibility and the conditions for appropriate security and control measures, would reduce legal uncertainty and at the same time strengthen the protective rights of those affected.

In the meanwhile, standardised, interoperable identification procedures could be used to uniformly define which verification methods are recognised as being compliant with data protection and legally secure. This could be achieved, for example, by expanding national eID systems or by establishing certified identity providers within the meaning of Articles 25 and 26 of the DSA (see also Recitals 66-68 DSA). Such an approach would allow platforms to rely on reliable, tested procedures instead of developing their own risky procedures.

On the other hand, clear minimum technical standards for scraping prevention, for example in the form of mandatory security mechanisms or guidelines for monitoring and rate limiting procedures, analogous to the requirements of Article 24 DSA and recitals 66–68. Such harmonisation would enable platform operators to verify in a comprehensible manner which measures are sufficient to meet the requirements of the ECJ without each individual case requiring a risky interpretation.

In summary, it is clear that although the Russmedia ruling significantly tightens the obligations of platform operators, they still have room for manoeuvre and discretion. At the same time, legislators also have opportunity to shape

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

https://doi.org/10.5281/zenodo.22905603

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

the implementation of these obligations. The establishment of standardised procedures for identification and scraping prevention, coupled with clear minimum requirements, would significantly improve the balance between data protection, platform operation and innovation. Until then, the supervisory authorities for data protection and intermediary services are now called upon to adapt their coordinated guidelines (see [18]) and add instructions for the implementation of identification and scraping prevention obligations within the framework of their legal mandate.

7 IMPLICATIONS FOR WEB INDEXING AND SEARCH ENGINES

The Russmedia judgment carries implications for web indexing infrastructures and search engines, even though these actors are not directly at issue in the underlying dispute.

By reaffirming a functional and purpose-oriented interpretation of the concept of controller under Article 4(7) GDPR, the ECJ confirms that responsibility does not depend on the formal role of a service provider but on its effective influence over the dissemination and accessibility of personal data. Operators of web indexes and search engines typically determine the means of processing through crawling, scraping, indexing, ranking and presentation mechanisms. To the extent that such design choices pursue autonomous economic objectives—most notably optimisation of discoverability, advertising revenues or user engagement—they may constitute an independent determination of processing purposes, thereby grounding controller responsibility under GDPR.

The decision further sharpens the legal relevance of indexing as an act that amplifies publicity rather than merely reflects existing publication. Search engines transform isolated user-generated content into globally searchable information structures. This amplification function links indexing directly to the fundamental rights framework of Arts. 7 and 8 CFREU, while simultaneously engaging the freedom to conduct a business under Article 16 CFREU and the freedom of information under Article 11 CFREU. The Court's reasoning therefore reinforces the established balancing paradigm known from other cases such as Google Spain [3] and GC v CNIL [17], according to which search engines bear autonomous responsibilities precisely because their processing and services, including indexing, changes the qualitative impact of personal data dissemination.

Indirect regulatory consequences also arise from the interaction with the DSA. While web index operators are not necessarily classified as online platforms (but can and should be, see [18]), the judgment aligns with the DSA's preventive and risk-based governance logic. Very large search engines are expressly subject to systemic risk assessment obligations under Arts. 34 and 35 DSA, requiring them to evaluate risks to fundamental rights resulting from information dissemination at scale. The Court's emphasis on preventing unlawful publication of sensitive personal data strengthens the expectation that indexing systems incorporate ex ante safeguards—such as risk-sensitive crawling policies, rapid de-indexing procedures and detection mechanisms for manifestly unlawful content—in order to mitigate systemic risks linked to automated discoverability.

Finally, the decision highlights a structural tension between openness and compliance inherent to web indexing. The economic and societal function of search engines presupposes maximum accessibility and wide dissemination of information; excessive restrictions on copying, hindering indexing or scraping could undermine both legitimate user interests in visibility and the internal market freedoms protected by the common market in EU-Treaties. The Court's reasoning therefore suggests a regulatory shift away from limiting copying online content as such toward strengthening upstream safeguards against the unlawful upload of sensitive personal data. For search engines and open web index initiatives, the practical implication is not a general duty to restrict crawling, but an increased obligation to integrate fundamental-rights-aware governance into indexing architectures, thereby reconciling large-scale information accessibility with data protection compliance.

REFERENCES

[1] P.C. Johannes, "Datenschutzrechtliche Verantwortlichkeit von Online-Plattformen für nutzergenerierten Content – Sensible Daten erkennen, Scraping verhindern und Nutzer identifizieren", *Multimedia und Recht*, vol. 28, no. 4, 2026.

[2] ECJ, judgment of 2 December 2025 – C-492/23, ECLI:EU:C:2025:935.

[3] ECJ, judgment of 13 May 2014 – C-131/12 (Google Spain).

[4] ECJ, judgment of 5 June 2018 – C-210/16 (Witschaftsakademie).

[5] ECJ, judgment of 29 July 2019 – C-40/17 (Fashion ID).

[6] ECJ, judgment of 3 October 2019 – C-18/18 (Glawischnig-Piesczek).

[7] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ("Directive on electronic commerce").

[8] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act).

[9] K. Erenli, "§ 16 Haftungsregelungen für Vermittlungsdienste", in *Europäisches Datenrecht*, C. Geminn and P. C. Johannes, Ed., Baden-Baden, Germany: Nomos, 2025, pp. 285-293, para. 2 et. seqq.Geminn/Johannes, EU-DatenR-HdB/Erenli §16 para. 2 et seq.

[10] M. Nebel, "§ 32 Datenschutzrecht", in *Europäisches Datenrecht*, C. Geminn and P. C. Johannes, Ed., Baden-Baden, Germany: Nomos, 2025, pp. 605-635, para. 32 et. seqq.

[11] ECJ, judgment of 14 December 2023 – C 340/21 (Natsionalna agentsia za prihodite).

[12] L. Beer *et. al.*, "Ein offener Webindex: Chancen und Risiken für Grundrechte und Grundfreiheiten", in Tagungsband Plattform Privatheit "Freiheit digitaler Infrastrukturen", 2024, Berlin, Germany, Oct. 2024, pp. 77-105.

https://doi.org/10.5281/zenodo.22905603

[13] G. de Gregorio, *Digital Constitutionalism in Europe*, Cambridge, UK: CU Press, 2022.

[14] M. A. Brown *et. al.*, "Web scraping for research: Legal, ethical, institutional, and scientific considerations", *Big Data & Society*, vol. 12, no. 4, Dec. 2025, https://doi.org/10.1177/20539517251381686.

[15] R. Arboretti *et al.*, "An Integrated Framework for Automated Web Scraping and Sentiment Analysis of Product Reviews", *Journal of Machine Intelligence and Data Science (JMIDS)*, vol. 5, 2024, pp. 46-53, https://doi.org/10.11159/jmids.2024.

[16] P. C. Johannes, "§ 29 Elektronische Identifizierung", in *Europäisches Datenrecht*, C. Geminn and P. C. Johannes, Ed., Baden-Baden, Germany: Nomos, 2025, pp. 563-574.

[17] ECJ, judgment of 24 September 2019 – C-136/17 (GC and Others v Commission nationale de l'informatique et des libertés (CNIL)).

[18] EDPD, "Guidelines 3/2025 on the interplay between the DSA and the GDPR", Version 1.1, Adopted on 11 September 2025, https://www.edpb.europa.eu/our-work-tools/documents/public-consultations/2025/guidelines-32025-interplay-between-dsa-and-gdpr_en.

[19] M. Nebel and P. C. Johannes, "Open Web Index im Lichte des Digital Services Act – Voraussetzungen – Grenzen – Rechtsfolge", *Multimedia und Recht*, vol. 26, no. 12, 2024, pp. 1010-1016.

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22903744

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

OUR^{2}S & YOARS: An Open Retrieval Platform and a Concept for Agentic Information Access

M. Granitzer[A,B], S. Zerhoudi[A], J. Mitrović[A]
K. Ghosh Dastidar[A], L. Caspari[A], M. Dinzinger[A]
Mohammed Al-Mamaari[A], D. H. Dang[A], D. Ciz[C]

[A]University of Passau, Germany
[B]IT:U Interdisciplinary Transformation University Austria, Linz
[C]IT4Innovations National Supercomputing Center, Ostrava, Czech Republic
Correspondence: michael.granitzer@uni-passau.de

Abstract

The dominance of proprietary search engines has created a well-known structural gap for information retrieval (IR) research: web-scale retrieval infrastructure is controlled by a handful of technology companies and remains largely unavailable for independent scientific investigation. As AI agents increasingly mediate information access, an analogous gap is opening in agentic retrieval—and the window to address it before the same lock-in occurs is narrow. Open, transparent, and deployable infrastructure for agentic IR is urgently needed, both for research and for society at large.

We present OUR^{2}S (OpenWebIndex **U**niversal **R**etrieval **R**esearch **S**ystem), an open, multi-corpus retrieval platform supporting sparse, dense, and hybrid search with quality indicators. We then develop YOARS (Your Open Agentic Retrieval System) as a conceptual extension: a proposed collaborative knowledge layer in which people and AI agents could create, review, and maintain structured knowledge derived from the web. One design hypothesis is that hypermedia-guided navigation, following HATEOAS [1], could make valid next actions explicit to an agent while the server independently enforces authorization. Individual YOARS APIs and data structures exist, but the integrated platform and these proposed benefits have not been completed or evaluated. The paper therefore presents an operational retrieval foundation, a conceptual architecture, and questions for future study.

Introduction

The web has become the primary medium through which society accesses information, yet the infrastructure for indexing and searching it is almost entirely controlled by a small number of commercial actors. Google, Bing, and a handful of others determine what is indexed, how it is ranked, and what information is surfaced to users—at a scale and with a level of opacity that makes scientific replication, independent auditing, or community-driven improvement essentially impossible [2, 3].

This concentration creates a fundamental tension in IR research. The field nominally studies web-scale retrieval, but in practice most of its work is conducted on benchmark datasets that are static, limited in scope, and increasingly distant from the live, dynamic reality of the web [4]. Researchers cannot experiment with real-time crawl pipelines, cannot study how index freshness affects result quality, or audit ranking behaviour at web scale without licensing proprietary APIs. A significant portion of retrieval research is therefore not generalisable to the web as it actually exists.

The structural deficit extends beyond academia. Search engines are public-knowledge infrastructure: their design choices—what to index, how to rank, what to suppress—directly affect what citizens, journalists, and policymakers can discover. When this infrastructure is proprietary and opaque it is not subject to public accountability [3, 5]. Open alternatives matter not because proprietary systems are necessarily malicious, but because their concentration creates structural risks: commercial misalignment, political influence, vendor lock-in, and the inability of anyone outside these organisations to verify or improve how the web's knowledge is organised.

The OpenWebSearch.eu project [6, 7] was established to address this deficit. Its central artefact is the Open Web Index (OWI) [8], a large-scale, openly accessible index of the web built from original European web crawls conducted by the OpenWebSearch.eu consortium deployed across federated European HPC infrastructure [9], and freely available to researchers and developers without commercial restrictions. But an index alone is not a research platform. To be useful it must be queryable, extensible, manageable, and deployable—and it must support the full spectrum of modern retrieval paradigms, from classical term-based ranking to dense neural retrieval and, increasingly, agentic workflows.

This paper presents an open retrieval system and a proposed research platform built around OWI. **OUR^{2}S** is a modular search and retrieval engine for managing multiple corpora with support for sparse, dense, and hybrid retrieval, integrated quality indicators, and a clean programmatic interface. **YOARS** is the design of a collaborative knowledge layer on that foundation, where agents and people would jointly create, validate, maintain, and share structured knowledge. Early YOARS components are implemented in the open-source codebase,[1] but the envisioned end-to-end col-

[1] `https://opencode.it4i.eu/openwebsearcheu-public/ours`

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22903744

laboration platform is unfinished and its research claims remain untested.

Crucially, we do not propose a single replacement search engine. We propose a *platform on which multiple search systems—and multiple research communities—can build*. Running high-quality open infrastructure is not a task for one group alone, and the goal is not to replicate big tech but to provide an open foundation that is collectively better. This requires three communities working together: *researchers* from information retrieval, AI, NLP, HCI, ethics, and adjacent disciplines who study algorithms, evaluation, human interaction, and the societal implications of information access; the *open-source community* who ensure code quality, long-term development, and sustainability; and *society at large*—institutions who fund the infrastructure and individuals who choose to use it as an alternative to proprietary services. The contribution of this paper is to describe the operational OUR^{2}S foundation, motivate the YOARS design, and identify what must be built and evaluated before the broader collaboration claim can be made.

Background: The Open Web Search Ecosystem

The Open Web Index

The Open Web Index (OWI) is a large-scale web index produced by the OpenWebSearch.eu consortium [7, 8] through original European web crawls complemented by Common Crawl data [10], deployed across three federated data centres on EuroHPC JU infrastructure [9]. Unlike proprietary search indexes, OWI is openly accessible, continuously updated, and accompanied by documented crawl methodology, provenance information, and quality assessments. It provides researchers with a realistic, large-scale crawl corpus without requiring independent crawling infrastructure.

By the end of the project the OWI infrastructure had crawled approximately 39.6 billion web pages, yielding 12.6 billion unique pages and 1.25 PiB of raw data distributed across 516 million hosts. Preprocessing and indexing has been conducted on EuroHPC JU Infrastructure [9] and efforts for crawling, preprocessing and indexing are ongoing. This scale is sufficient to study the long-tail phenomena—rare queries, low-resource languages, niche topics—that are systematically underrepresented in standard benchmark datasets but critically important for understanding real-world retrieval behaviour.

The Research Gap

Despite the availability of OWI, using it for rigorous retrieval research requires significant engineering effort. Researchers need a scalable indexing and query engine; support for diverse retrieval paradigms (sparse, dense, hybrid); tools for managing multiple corpora side by side; reproducible benchmark infrastructure; and, increasingly, interfaces compatible with LLM-based agents and retrieval-augmented generation (RAG) pipelines [11, 12], including agents that require scope- and freshness-aware access to live web content [13]. Until now these components had to be assembled from scratch for each project, creating redundant effort and high barriers to entry.

Closing this gap is not a task for any single group. It calls for collaboration across the three communities introduced above: researchers who push the algorithmic frontier, developers who maintain and improve the shared codebase, and institutions and individuals who deploy, fund, and use the resulting services. OUR^{2}S supplies an existing retrieval foundation; YOARS proposes a possible collaboration layer above it.

OUR^{2}S: OpenWebIndex Universal Retrieval Research System

Design Philosophy

OUR^{2}S is a research-quality retrieval platform built on Vespa [14], an open-source serving engine for large-scale real-time query evaluation. It is actively developed and open for community contributions; the design reflects four core principles.

Single application, multiple corpora. All corpora are managed within one Vespa application, with isolation achieved through schema-level suffixing so that term statistics remain independent per corpus while infrastructure is shared.

Model-driven architecture. Data models defined in Pydantic [15] drive schema generation, API validation, and CLI serialisation—the model is the single source of truth across all interfaces.

Unified interface. Indexing, querying, corpus management, and RAG orchestration are exposed through a REST API (FastAPI), a CLI, and a browser dashboard—three entry points, one consistent contract.

Deployment independence. OUR^{2}S runs on Kubernetes, a single server, or a local machine, with no dependency on any proprietary cloud service or commercial API.

Retrieval Capabilities

The retrieval capabilities of OUR^{2}S are provided by Vespa [14]; our contribution is to configure, expose, and orchestrate them through a unified, corpus-aware interface. Three paradigms are supported.

Sparse retrieval via BM25 [16] operates over per-corpus term statistics, ensuring ranking is calibrated to each corpus's vocabulary rather than a shared global index.

Dense retrieval via HNSW-indexed embedding vectors [17] is configured per corpus and document type, allowing researchers to plug in any embedding model—including those from recent dense retrieval work [18].

Hybrid retrieval combines BM25 and vector scores through configurable fusion, enabling study of the complementarity of sparse and dense signals.

Quality-Aware Retrieval

OUR^{2}S includes an initial set of document quality indicators—a starting point we invite the community to extend.

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22903744

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

The current signals cover text quality (language, readability, content completeness), link quality (inbound/outbound ratios, anchor density), structural quality (markup richness, schema.org coverage), and entity quality (entity density, disambiguation confidence).

Signals are normalised using distribution models suited to each type (logit-normal for bounded fractions, log-normal for unbounded counts) and stored compactly in Vespa's native field types. At query time they can feed ranking profiles, filters, or document metadata. This enables quality-aware retrieval research that standard benchmark datasets—where all documents are treated as equally reliable—do not support. The indicator framework is intentionally modular: adding a new signal requires a model field, a normalisation entry, and a ranking expression.

Corpora and Data Model

OUR^{2}S organises content into a hierarchy of document types: full web documents, extracted text chunks, named entities, geographic places, structured products and events, and conversation fragments.

Standard corpora available out of the box include the OWI itself, MS MARCO [19], and Wikipedia, but planned to be expanded. This makes OUR^{2}S suitable both as a platform for standard benchmark evaluation and as a workbench for novel corpus construction and exploration.

YOARS: A Concept for Agentic Knowledge Collaboration

Vision

YOARS begins from the observation that web-scale knowledge is not static: it must be maintained, updated, and validated over time. The proposed platform would use the web index as a source for a living knowledge base curated jointly by people and AI agents, in a spirit analogous to Wikipedia [20] but designed for both human and machine participation. Whether this collaboration can produce reliable knowledge at useful scale remains an empirical question.

The proposed core abstraction is the *nugget*: a compact, structured unit of knowledge with temporal, provenance, and governance metadata. OUR^{2}S supplies retrieval over source material; the envisioned YOARS layer would support review and maintenance of knowledge derived from those sources. Prototype data models and routes explore parts of this design, but their presence does not establish the full workflow.

Nugget-Centred Architecture

In the proposed model, a nugget has a stable key, a compact description usable by people and language models, and temporal metadata. Validity would be established through evidence and review rather than inferred from a timestamp alone.

Typed *aspects* would extend this core with domain-specific information without fragmenting the index into incompatible object types. Examples include licence information, geographic context, source provenance, confidence and validation metadata, procedural steps, and usage observations. The nugget would remain the retrieval anchor while aspects add filtering and domain context.

The design distinguishes epistemic status from search rank. Inspired by formal treatments of knowledge and argument [21–23], a status annotation could mark a claim as candidate, confirmed, disputed, or retracted. Topic annotations would supply retrieval context, while provenance records would connect claims to evidence, support, and challenges. Competence annotations could identify procedural knowledge without treating it as evidence that an agent is trustworthy. The semantics, review rules, and effect of these annotations on retrieval still require specification and evaluation.

Human–Agent Co-Creation

The central design goal of YOARS is *co-creation*: people and agents would both be able to propose knowledge, revisions, and evidence. This differs from a workflow in which an agent only assists a human or generates text for human approval, and from agent-only networks such as Moltbook [24]. The design would allow either kind of participant to initiate a nugget, challenge a claim, or request review. Whether such participation improves knowledge quality is untested.

The proposed workflow would apply the same evidence and review criteria to human- and agent-authored claims, while keeping their authorship visible. A human could challenge an agent's nugget, and an agent could flag a human-authored claim for review. Participants could hand off retrieval, synthesis, or adjudication tasks when the next step requires different expertise. These interactions need explicit responsibility rules and an end-to-end trial before they can be described as a working collaboration model.

One governance option would give the original author a default role in reviewing proposed changes, with escalation when authors disagree or are unavailable. Temporal metadata could flag stale claims for renewed review. The authority model, treatment of disputes, and any influence of these signals on ranking are design questions rather than validated platform behavior.

The proposed pipeline would treat LLM-extracted nuggets as candidates, with uncertainty and source evidence attached. Human and agent review could confirm, revise, dispute, or reject a candidate. An evaluation must determine whether reviewers detect errors, how disagreements are resolved, and whether uncertainty can be represented usefully in search.

Purposeful Social Discovery

The YOARS concept includes a social discovery layer intended to help agents find useful capabilities, sources, and workflows. A feature would be justified by improved task outcomes or coordination, rather than engagement alone; that criterion still needs experimental testing.

One proposed representation is the *capability card*: a discovery record for an endpoint, dataset, prompt, or workflow,

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22903744

linked to its source and observed outcomes. The research question is whether sharing such records helps other agents select effective resources without amplifying unverified recommendations.

Hypermedia-Driven Agent Control

A recurring challenge for agents using APIs is deciding which action to take next from a large catalogue of endpoints. An agent may choose a call that is irrelevant, stale, or unauthorized. Server-side access control remains necessary, but it does not by itself help the agent discover a useful workflow.

The YOARS design explores **Hypermedia as the Engine of Application State (HATEOAS)** [1] as a way to expose relevant next actions. A response could carry typed *affordances* such as links to inspect a search result, refine a query, or propose a nugget. In a mature design, the server would vary these links by workflow state and audience. An early API already has audience-separated routes and some affordance metadata, but a complete state-dependent navigation model has neither been implemented nor evaluated.

The proposed workflow model would define server-validated transitions and record both offered links and attempted actions. A link would guide navigation; it would not grant authority. Every request would still need an authorization check, including requests that guess a URL or reuse an old link. Logs could then show whether an agent followed an offered transition and whether the server accepted it.

Such navigation might reduce the need for agents to memorize endpoint paths and might help them adapt as an API evolves. Those are hypotheses, not measured properties of YOARS. Agents may ignore links, misread their meaning, or call remembered endpoints anyway.

An evaluation should compare linked navigation with a flat tool catalogue under the same agents and tasks. It should measure task completion, invalid and unauthorized calls, token cost, and behavior when links or documentation change. This would test whether hypermedia improves predictability in practice.

Agent-to-agent (A2A) coordination is a further research direction. Delegation across agents could amplify errors, including those caused by contaminated knowledge or unclear authority. A future YOARS testbed could study these risks through controlled tasks, provenance records, and audit logs. This paper does not report such an experiment.

Agent Interface

Early YOARS components expose separate API routes for public, agent, human, and admin audiences. Route metadata and links provide an initial navigation vocabulary. This implementation is a starting point for the conceptual interface above; it is not evidence that agents can complete the proposed workflows or that link selection enforces a full state machine.

A machine-readable skill catalogue could package recurring interaction patterns with typed parameters and expected outputs. Prototype skill records and discovery routes exist, but the catalogue's usefulness, role-specific discovery, and effect on task completion require end-to-end evaluation. A changing count of configured skills would not establish those properties.

Sovereignty and Research Scope

Full Deployability as a Design Goal

A core goal of OUR^{2}S is deployment on infrastructure controlled by its operators. The same independence is a requirement for the proposed YOARS platform, but a complete YOARS deployment path and its operational costs have not yet been demonstrated.

For YOARS, the design target is an open-source stack with no mandatory proprietary cloud API, documented deployment, and control of identities, data, and policy by the operating institution. The existing codebase contains components toward that target; it does not yet show that a third party can install and operate the full collaboration workflow. Independent deployment remains an acceptance criterion for future work.

However, full deployability is only one side of the problem. Many research groups do not have the infrastructure, nor the operational capacity, to run and maintain a web-scale retrieval platform at all. The IR community has responded to this gap partly through simulation [25, 26]: if a live platform is unavailable, one can model user behaviour, index access, and retrieval dynamics computationally. Simulation is a valuable methodology, but it is inherently limited by the fidelity of the model—and the agentic systems now entering the field are complex enough that simulating their behaviour faithfully may be more difficult than running them against a real platform.

OUR^{2}S can supply an open retrieval base for shared experiments. A completed YOARS could add a common environment for studying knowledge review and agent collaboration on top of it. To establish that role, independent groups would need to deploy the same version, use documented datasets and tasks, and reproduce the results. Community ownership remains an objective, not an outcome shown by this paper.

This goal matters as agents become a more common interface to retrieval. An open testbed would let researchers inspect agent actions, compare policies, and reproduce failures under shared conditions. OUR^{2}S is a starting point for such a testbed; YOARS describes the collaboration functions that would still need implementation and evaluation.

The Research Platform Case

For IR research, OUR^{2}S offers an inspectable retrieval implementation and a route toward experiments over open corpora. Comparable studies still require shared corpus snapshots, configurations, and evaluation protocols. YOARS proposes an additional layer for studying how people and agents create and maintain knowledge; this paper supplies its design rationale rather than results from such a study.

Prior open-infrastructure efforts include controlled evaluation: TIRA, for instance, provides a sandbox for reproducible system comparison [27]. OUR^{2}S targets a complementary

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22903744

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

need for inspectable retrieval services that can be deployed and extended. The proposed YOARS layer would make agentic knowledge maintenance a subject of similar controlled study if a complete testbed is built.

Reproducibility matters because large model weights, private training data, and proprietary serving systems can obstruct independent verification of neural retrieval results [18]. An open platform can reduce some of these barriers when its datasets, configurations, and measurements are also published.

Impact Beyond Academia

Open retrieval infrastructure can support work outside academia, such as search over collections controlled by journalists or public-interest organisations. A completed YOARS could additionally support collaborative claim review and maintenance. Those applications are possible use cases, not deployments or impact demonstrated here.

Wikipedia illustrates the potential of shared knowledge maintenance [20]. The long-term ambition is an analogous commons for open retrieval and agentic knowledge work. OUR^{2}S contributes retrieval infrastructure to that ambition; the governance and public value of the proposed YOARS layer remain to be established.

Current State and Outlook

OUR^{2}S is operational with full support for OWI, MS MARCO, and Wikipedia corpora. The REST API and CLI are available alongside an HTMX admin UI. Import pipelines are benchmarked at 75 documents per second with full quality indicator computation per lane and scaling up to 10 lanes. Kubernetes deployment is documented and tested.

The YOARS codebase contains partial implementations of domain-scoped APIs, knowledge records, agent profiles, skill discovery, and logging. These components have not yet been brought together as the complete collaboration platform described above or tested through an integrated research evaluation. In particular, this paper reports no measured effect of human–agent review, hypermedia-guided navigation, social discovery, or A2A coordination.

The next step is to implement and exercise a complete path from source retrieval through candidate creation, evidence review, dispute, and subsequent retrieval. Independent deployment and controlled comparisons with simpler agent interfaces and review workflows would then test the design hypotheses and expose operational limits.

Source code for OUR^{2}S and the developing YOARS components is available at the project repository.[2] The public search service at `https://ourrs.eu` demonstrates the retrieval side; it should not be interpreted as a validated deployment of the full YOARS concept.

Conclusion

OUR^{2}S provides an open retrieval foundation for studying information access beyond proprietary search services. YOARS is a proposed collaboration layer on that foundation, supported by early components but not yet complete or evaluated as an integrated system. Its central question is whether people and agents can maintain useful, accountable knowledge together using explicit evidence, review, and governance.

For IR research, the combination suggests a path from an inspectable retrieval engine to controlled studies of agentic knowledge work. For society, it sets a goal of independently governed information access. Neither the proposed collaboration model nor its broader effects can be claimed from the implementation status reported here.

A research agenda follows from the design: implement a complete knowledge lifecycle, test whether HATEOAS-guided agents navigate more reliably than agents using a flat tool catalogue, and measure the quality and cost of human–agent review. Server-side authorization and audit remain necessary regardless of the navigation interface.

We invite the research community to use and challenge OUR^{2}S, and to develop and test the YOARS concept against these questions.

[2] `https://opencode.it4i.eu/openwebsearcheu-public/ours`

Acknowledgements

This work was supported by the OpenWebSearch.eu project, funded by the European Union under the Horizon Europe programme (grant agreement No. 101070014) and by the B2B-Konnekt project (28DI110A23) funded by Federal Ministry of Agriculture, Food and Regional Identity. Computational resources were provided by EuroHPC JU.

The authors used AI-assisted tools (Claude by Anthropic) during the preparation of this manuscript to support writing improvement, code development, and literature organisation. All content was reviewed, verified by the authors, who take full responsibility.

REFERENCES

[1] R. T. Fielding, "Architectural styles and the design of network-based software architectures," Ph.D. dissertation, University of California, Irvine, 2000, chapter 5 introduces REST and the HATEOAS constraint. [Online]. Available: https://ics.uci.edu/~fielding/pubs/dissertation/top.htm

[2] D. Metzler, Y. Tay, D. Bahri, and M. Najork, "Rethinking search: Making domain experts out of dilettantes," *ACM SIGIR Forum*, vol. 55, no. 1, pp. 1–27, 2021.

[3] L. D. Introna and H. Nissenbaum, "Shaping the web: Why the politics of search engines matters," *The Information Society*, vol. 16, no. 3, pp. 169–185, 2000.

[4] E. M. Voorhees and D. K. Harman, Eds., *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, MA: MIT Press, 2005.

[5] A. Mowshowitz and A. Kawaguchi, "Bias on the web," *Communications of the ACM*, vol. 45, no. 9, pp. 56–60, 2002.

[6] OpenWebSearch.eu Consortium, "OpenWebSearch.eu: Towards a european open web search infrastructure," https://openwebsearch.eu, 2022, accessed March 2026.

https://doi.org/10.5281/zenodo.22903744

[7] M. Granitzer, S. Voigt, N. A. Fathima, M. Golasowski, C. Guetl, T. Hecking, G. Hendriksen, D. Hiemstra, J. Martinović, J. Mitrović, I. Mlakar, S. Moiras, A. Nussbaumer, P. Öster, M. Potthast, M. S. Srdić, M. Sharikadze, K. Slaninová, B. Stein, A. P. de Vries, V. Vondák, A. Wagner, and S. Zerhoudi, "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.

[8] G. Hendriksen, M. Dinzinger, S. M. Farzana, N. A. Fathima, M. Fröbe, S. Schmidt, S. Zerhoudi, M. Granitzer, M. Hagen, D. Hiemstra, M. Potthast, and B. Stein, "The open web index: Crawling and indexing the web for public use," in *Advances in Information Retrieval – ECIR 2024*, ser. Lecture Notes in Computer Science. Springer Nature Switzerland, 2024, pp. 130–143.

[9] M. Granitzer, M. Hayek, S. Heineking, G. Hendriksen, M. Golasowski, M. Dinzinger, and S. Zerhoudi, "OpenWebSearch.eu — building an open web index on EuroHPC JU infrastructures," *Procedia Computer Science*, vol. 255, pp. 43–52, 2025.

[10] Common Crawl Foundation, "Common crawl: Open repository of web crawl data," https://commoncrawl.org, accessed March 2026.

[11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, vol. 33, pp. 9459–9474.

[12] S. Zerhoudi and M. Granitzer, "PersonaRAG: Enhancing retrieval-augmented generation systems with user-centric agents," in *Proceedings of the Workshop on Information Retrieval's Role in RAG Systems (IR-RAG) at SIGIR*, Washington D.C., USA, 2024.

[13] S. Zerhoudi, M. Dinzinger, M. Granitzer, and J. Mitrović, "OwlerLite: Scope- and freshness-aware web retrieval for LLM assistants," in *Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26)*. ACM, 2026, pp. 216–220.

[14] Vespa.ai, "Vespa: The open big data serving engine," https://vespa.ai, accessed March 2026.

[15] S. Colvin *et al.*, "Pydantic: Data validation using Python type annotations," https://docs.pydantic.dev, accessed March 2026.

[16] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[17] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 824–836, 2020.

[18] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. tau Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6769–6781.

[19] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang, "MS MARCO: A human generated machine reading comprehension dataset," 2016, arXiv:1611.09268.

[20] J. Wales and L. Sanger, "Wikipedia: The free encyclopedia," https://wikipedia.org, 2001, accessed March 2026.

[21] S. A. Kripke, "Semantical considerations on modal logic," *Acta Philosophica Fennica*, vol. 16, pp. 83–94, 1963.

[22] P. M. Dung, "On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and *n*-person games," *Artificial Intelligence*, vol. 77, no. 2, pp. 321–357, 1995.

[23] D. Walton, C. Reed, and F. Macagno, *Argumentation Schemes*. Cambridge University Press, 2008.

[24] Y. Jiang, Y. Zhang, X. Shen, M. Backes, and Y. Zhang, ""Humans welcome to observe": A first look at the agent social network Moltbook," 2026.

[25] L. Azzopardi, T. Breuer, B. Engelmann, C. Kreutz, S. MacAvaney, D. Maxwell, A. Parry, A. Roegiest, X. Wang, and S. Zerhoudi, "SimIIR 3: A framework for the simulation of interactive and conversational information retrieval," in *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP)*, 2024, pp. 197–202.

[26] S. Zerhoudi and M. Granitzer, "UXSim: Towards a hybrid user search simulation," in *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM)*, 2025, pp. 6570–6574.

[27] M. Potthast, T. Gollub, M. Wiegmann, and B. Stein, "TIRA integrated research architecture," in *Information Retrieval Evaluation in a Changing World*, ser. The Information Retrieval Series. Springer, 2019, pp. 123–160.

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.23070340

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

AN OPEN-SOURCE RETRIEVAL LAYER FOR INSTITUTIONAL RAG: FROM AUTHENTICATED CRAWL TO SEARCHABLE INDEX

Noor A. Fathima*, A. Wagner[1], M. Guijarro[1], F. Rehm[1]

[1] CERN, Geneva, Switzerland

Abstract

Large research organisations maintain extensive internal web estates behind Single Sign-On authentication, making institutional knowledge invisible to standard web crawlers and RAG systems. We present an open-source retrieval layer, developed within the OpenWebSearch.EU project and deployed at CERN, that bridges authenticated web content to downstream AI applications. The system comprises two crawlers — one for HTML (WARC archiving via StormCrawler/Playwright) and one for PDF (text extraction via Apache Tika) — both authenticating through browser automation against CERN's Keycloak SSO. Deployed in production, the system crawled 25,293 URLs across 182 domains, producing 3.1 GB of WARC archives and 30 MB of extracted PDF text. We report architectural findings including the use of WARC resource records (rather than request/response pairs) as an inherent consequence of browser-level fetching, and the necessity of auth-redirect detection to prevent SSO login pages from polluting the content corpus.

INTRODUCTION

Large research organisations maintain extensive internal web estates behind Single Sign-On (SSO) authentication. At CERN, over 180 semi-independent web applications — from Jira issue trackers and EDMS engineering document repositories to Confluence knowledge bases and Indico event systems — hold institutional knowledge that is invisible to public search engines and inaccessible to standard web crawlers.

Retrieval-Augmented Generation (RAG) systems offer a path to making this knowledge queryable through natural language. However, the retrieval layer — the component responsible for acquiring, processing, and indexing content — is often treated as a solved problem when the target corpus is public. For authenticated institutional content, it is not: standard crawlers cannot navigate SSO flows, and the content mix of HTML pages and PDF documents requires fundamentally different processing pipelines.

This paper presents an open-source retrieval layer developed within the OpenWebSearch.EU project [1] and deployed at CERN for AccGPT, an internal RAG-based AI assistant. The system comprises two crawlers that authenticate via browser automation, process HTML and PDF content through separate pipelines, and deliver structured output to a shared object store. We describe the architecture, report production results from three crawl campaigns covering 25,293 URLs across 182 domains, and discuss the design decisions that emerge when building retrieval infrastructure for authenticated institutional content.

The system is released as open source under the MIT licence, with repositories available on CERN GitLab.

* afshan.shokath@gmail.com

SYSTEM ARCHITECTURE

The retrieval layer connects four stages: seed management, authenticated crawling, output processing, and content delivery (Fig. 1). Content type — determined during seed classification — routes each URL to the appropriate crawler.

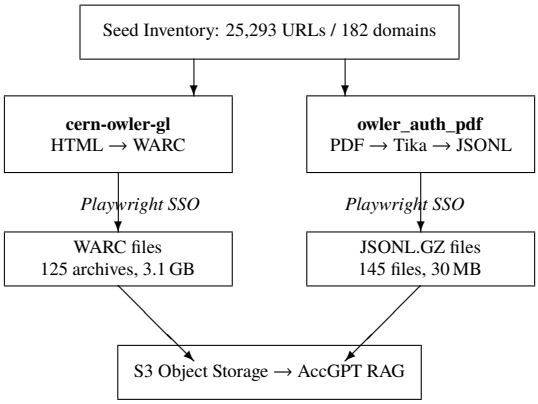


Figure 1: Two-crawler architecture. Both crawlers authenticate via Playwright browser automation and deliver to a shared S3 endpoint.

cern-owler-gl is a Java-based crawler built on the StormCrawler/OWLer framework [2]. It extends the standard OWLer pipeline with a `PlaywrightFetcher` that renders pages in headless Chromium after navigating CERN's Keycloak SSO flow. Output is stored as WARC 1.1 records [3] — the same archival format used by the Internet Archive — preserving the full rendered DOM for each page.

owler_auth_pdf is a Python microservice stack (FastAPI, Celery, Redis, Apache Tika) purpose-built for PDF content. It authenticates via its own Playwright session manager, downloads PDFs through authenticated HTTP, extracts text and metadata via Tika, and outputs structured JSONL records. A 10-container deployment provides clean service boundaries, independent failure domains, and integrated monitoring (Prometheus, Grafana, Loki).

Both crawlers run on a shared server, authenticate against the same SSO infrastructure, and deliver to an S3 endpoint at CERN's Prévessin data centre.

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.23070340

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

AUTHENTICATED CRAWLING

SSO Navigation

CERN uses Keycloak-based SSO with multi-step redirects: the target application redirects to `auth.cern.ch`, which presents a login form requiring username, password, and optionally a second factor. Standard HTTP clients cannot navigate this flow because the login page requires JavaScript execution and the redirect chain involves client-side state.

Both crawlers use Playwright [4] to automate a headless Chromium instance. The browser navigates to a protected URL, detects the SSO redirect, fills credentials, completes the login, and captures the resulting session cookies. These cookies are then reused for subsequent HTTP requests, avoiding repeated browser launches.

A key design decision was *auth-redirect detection*: if a response URL contains `auth.cern.ch` instead of the expected content domain, the session has expired. The crawler detects this condition, triggers re-authentication, and retries. Without this guard, SSO redirect pages — containing only JavaScript login redirects — would be archived as valid content, polluting the downstream RAG pipeline.

HTML Pipeline: Resource Records

An unexpected architectural consequence of browser-based fetching is the WARC record type. Standard HTTP crawlers (Heritrix, wget) capture paired `request` and `response` WARC records, preserving the full HTTP exchange including status codes, cache headers, and ETags. The `PlaywrightFetcher` operates at the browser level, not the HTTP level: it renders the page in Chromium, extracts the final DOM, and stores it as a WARC `resource` record. No HTTP headers are preserved.

Listing 1: WARC resource record (truncated). Note the absence of HTTP headers — content starts immediately with the HTML document.

```
WARC/1.0
WARC-Type: resource
WARC-Target-URI: https://its.cern.ch/jira/...
Content-Type: application/octet-stream
OWS.TAGS: HTML
OWS.GENAI: TRUE

<!DOCTYPE html><html>...
```

This trade-off is acceptable for RAG (which needs text content, not HTTP metadata) but would need to be addressed for change detection or deduplication, where cache headers and ETags are essential.

PDF Pipeline: Text Extraction

PDFs require active processing: a WARC-archived PDF binary is useless for a language model. The PDF crawler downloads each file via authenticated HTTP, passes it to Apache Tika for text and metadata extraction, and outputs structured JSONL with 11 fields per record including the full extracted text, 40+ Tika metadata keys, processing duration, and a status indicator.

Output files are segregated by status: `SUCCESS` (text extracted), `FAILED-DOWNLOAD` (auth redirect, 404, timeout), `FAILED-TIKA` (corrupted or unsupported PDF), and `ERROR` (unexpected exception). This segregation ensures the RAG pipeline ingests only validated content.

SEED MANAGEMENT AND CONTENT CLASSIFICATION

The crawl targets 25,293 URLs across 182 domains, with 98.6% CERN-internal. URLs were classified into HTML and PDF categories using URL-pattern heuristics: file extensions (`.pdf`), known PDF-serving domains (EDMS, CDS, CAS), and URL path patterns (`/file/`, `/record/`).

The seed inventory reveals that CERN's internal web is not a monolithic intranet but a federation of semi-independent applications. Table 1 shows the distribution across the largest domains.

Table 1: Top Seed Domains by URL Count

Domain	URLs	Type
its.cern.ch (Jira)	14,187	HTML
edms.cern.ch	8,473	PDF
indico.cern.ch	802	HTML
home.cern	350	HTML
cas.web.cern.ch	234	PDF
177 other domains	1,247	mixed
Total	**25,293**	

Each domain is processed independently with separate seed files, output directories, and status tracking. This per-domain isolation allows selective re-crawling and simplifies failure diagnosis.

PRODUCTION RESULTS

The system was deployed on a dedicated server (64 vCPU, 251 GB RAM, AlmaLinux 9.7) and ran three HTML crawl campaigns (June–October 2025) and one PDF extraction campaign (June 2025). Table 2 summarises the output.

Table 2: Production Output Summary

Metric	HTML	PDF
Crawl campaigns	3	1
Output files	125 WARCs	145 JSONLs
Total output size	3.1 GB	30 MB
Throughput	~1 pg/s	0.3–4.5 s/PDF
Fetcher threads	300	10 workers

HTML Crawling

The `PlaywrightFetcher` achieved approximately one authenticated page per second per thread, with 300 fetcher threads operating concurrently. A sample WARC archive from September 2025 contains 36 Jira issue pages crawled in 14 seconds, each stored as a `resource` record ranging from 8 to 45 KB.

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.23070340

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

Three campaigns ran sequentially as the authentication mechanism evolved. The first (June 2025) used cookie-based authentication, which failed against Keycloak — captured pages were SSO redirect pages rather than content. The `PlaywrightFetcher` was introduced on June 5, successfully navigating the SSO flow, and subsequent campaigns (September, October 2025) produced valid content.

PDF Extraction

The PDF pipeline processed content from 16 domains. Public sources achieved 100% success: 23 arxiv.org physics papers (607 pages, 1.89 M characters extracted) and 5 CERN Courier magazine editions (1 M+ characters). The bulk of processing targeted EDMS engineering documents (8,473 PDFs), where auth-redirect failures and corrupted engineering drawings caused the majority of `FAILED-DOWNLOAD` and `FAILED-TIKA` outcomes.

Extraction quality varied with PDF provenance. LaTeX-generated papers (arxiv) produced clean text with occasional watermark artifacts. Professionally published magazines (Adobe InDesign) yielded high-quality text with explicit text layers. Scanned engineering drawings produced poor or no extractable text, a known limitation of Tika's OCR-free pipeline.

DISCUSSION

Why Two Crawlers?

The most common question about this architecture. HTML archiving and PDF text extraction are fundamentally different operations: an HTML page is archived as-is in a WARC container, while a PDF must be actively processed through Tika to extract the text that makes it useful for a language model. Combining both into a single tool would create a monolith that does neither task well. The two-crawler approach allows each to be optimised for its content type while sharing infrastructure (server, SSO target, S3 endpoint).

Resource Records vs. Request/Response Pairs

The `PlaywrightFetcher`'s use of WARC `resource` records rather than the conventional `request`/`response` pair is an inherent consequence of browser-level fetching. This choice captures JavaScript-rendered content — essential for Single Page Applications and Jira's client-side rendering — but sacrifices HTTP metadata. For RAG applications this is acceptable; for web archiving or change-detection workflows, a hybrid approach that captures both the rendered DOM and the HTTP exchange would be preferable.

Credential Management

Crawler credentials (SSO passwords, API tokens, S3 keys) must be managed carefully. Both repositories use `.gitignore` to exclude credential files from version control, with `.env.example` templates providing the expected configuration shape. The git histories have been cleaned using `git-filter-repo` [6] to remove any credential traces before public release. We recommend rotating all credentials before publishing any crawler repository, regardless of `.gitignore` coverage.

FUTURE WORK

The current system uses static seed files with no change detection or re-crawl scheduling. Integration with the OpenWebSearch.EU URL Frontier [5] would add prioritised scheduling, distributed coordination, and metadata-driven re-crawl policies. We estimate 6–10 developer-days for this integration.

Generalisation beyond CERN's Keycloak SSO to other identity providers (SAML, OAuth2, institutional Shibboleth federations) would broaden the system's applicability to other research organisations with similar authenticated content challenges.

CONCLUSION

We have presented an open-source retrieval layer that bridges authenticated institutional web content to RAG pipelines. The system uses browser automation (Playwright) to navigate SSO flows, processes HTML and PDF content through separate optimised pipelines, and delivers structured output to downstream AI applications. Deployed at CERN, it crawled 25,293 URLs across 182 domains, producing 3.1 GB of WARC archives and 30 MB of extracted PDF text across three campaigns.

The key architectural insights are: (1) browser-level fetching produces WARC `resource` records rather than HTTP `request`/`response` pairs, an acceptable trade-off for RAG but not for web archiving; (2) HTML archiving and PDF text extraction require separate pipelines with different output formats; and (3) auth-redirect detection is essential to prevent SSO login pages from polluting the content corpus.

Source code is available under the MIT licence at:
`https://gitlab.cern.ch/openwebsearch_eu/cern-owler-gl`
`https://gitlab.cern.ch/openwebsearch_eu/owler_auth_pdf`

ACKNOWLEDGEMENTS

This work was supported by the European Union's Horizon Europe programme under Grant Agreement No. 101070014 (OpenWebSearch.EU). The author thanks the AccGPT team at CERN for defining the retrieval requirements that shaped this system, and the OpenWebSearch.EU consortium for the OWLer crawler framework and URL Frontier infrastructure.

REFERENCES

[1] OpenWebSearch.EU, "Open Web Search – Collaborative Open Web Search," `https://openwebsearch.eu/`, 2022–2025.

[2] J. Nière *et al.*, "StormCrawler: An Open Source Framework for Building Web Crawlers," `https://stormcrawler.net/`, 2014.

https://doi.org/10.5281/zenodo.23070340

[3] IIPC, "WARC File Format (ISO 28500)," https://iipc.github.io/warc-specifications/, 2017.

[4] Microsoft, "Playwright: Reliable end-to-end testing for modern web apps," https://playwright.dev/, 2020.

[5] S. Guertl *et al.*, "URL Frontier: A Scalable URL Management Service for Web Crawling," in *Proc. OSSYM 2024*, Munich, Germany, 2024.

[6] E. S. Newren, "git-filter-repo: Quickly rewrite git repository history," https://github.com/newren/git-filter-repo, 2019.

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22904930

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

PROPOSED METHODOLOGY FOR OPEN SEARCH INFRASTRUCTURE LEVERAGING FEDERATED RAG-SYSTEMS AND OPEN SOURCE SPEECH INTELLIGENCE FOR CROSS-DOMAIN COLLABORATION FEATURING DISASTER MANAGEMENT, TRAFFIC FLOW PREDICTION AND LITERATURE STUDIES

J. Schuster *, M. Stoffel, W. Kühn, J. Rapp, E. Niehaus

RPTU University Kaiserslautern-Landau, Landau, Germany

Abstract

Given the current research trajectory, the use of natural language for communication with machines for searching purposes will be an integral part in everyday life in the near future, featuring promising outlooks, but also raising concerns regarding voice data sovereignty, which stem from the exceptionally high sensitivity of voice data. For research purposes, the lack of sovereignty over search index data is problematic, as data curation is beyond ones control, potentially leading to suboptimal quality of the index data. This work proposes a methodology for integrating open source Speech Intelligence technologies into an open search infrastructure based on federated Retrieval Augmented Generation (RAG) systems containing domain knowledge aiming to provide data sovereignty over both RAG and speech data as well as improving output transparency, reliability and quality for small size cross-domain collaboration. Based on the considered use-cases of disaster management, traffic flow prediction and literature studies this paper further demonstrates how the system can drastically enhance small size cross-domain collaboration by overcoming communication barriers while retaining sovereignty over both voice and search index data.

INTRODUCTION

The research field of speech intelligence recently sees rapid advances by using the capabilities of Large Language Models (LLM), including speech recognition [Fathullah et al., 2024] and speech synthesis. Given the speed of this development it is most likely that speech intelligence soon will play a pivotal role in human-machine interaction on all levels. The exceptionally high sensitivity of voice data [Ritter, 2021] makes sovereignty of users over them crucial, otherwise putting them at risk of facing discrimination, including areas of healthcare and employment [Ritter, 2021].

Combining speech intelligence, including speech recognition and synthesis and open search infrastructure has the potential to drastically enhance user interaction with Retrieval Augmented Generation (RAG) systems, especially in situations where traditional input methods are not feasible or inefficient. In a RAG-system, existing LLMs are combined with an information retrieval system [Oche et al., 2025] that contains external information. This paper proposes a methodology to use open source technologies for speech recognition and speech synthesis to interact with RAG-systems and LLMs as an open search infrastructure. The paper discusses the approach from a small size application of RAG-systems, but retains the possibility to scale this view larger towards group specific aggregated use of RAG-systems or integrated open search ecosystem [Granitzer et al., 2024].

The proposed framework pursues three core objectives: protection against profiling of voice data through sovereignty over voice data, ensuring sovereignty over search index data by providing transparency and autonomy by using individually developed RAG-containers for managing index data, and overcoming barriers of interdisciplinary collaboration, particularly by breaking down communication barriers and enabling domain-external researchers to easily adapt index boundaries.

Such a system can be applied in various domains and use-case scenarios, like in disaster management, traffic flow prediction or literature studies. For example, the information retrieval system could be enriched with relevant materials like risk and resource maps, traffic flow data or literature related materials. By combining the information retrieval system with speech recognition and synthesis tools a voice controlled support bot for domain specific and interdisciplinary use of user-centric information systems is obtained [Zerhoudi and Granitzer, 2024].

Literature studies and classroom situations have less time constraints, but could benefit from similar interactive voice command driven approaches to access and explore knowledge from a specific domain with an exploratory approach to find relevant information for a current focus of interest [Jain et al., 2025]. Still, this cannot replace reading the source documents.

The paper also introduces the bottom-up scaling approach from a single-user application of a RAG-system to group and institutional applications of open search retrieval in the context of a support bot [Neumann et al., 2024]. In this context, a voice driven access to user-controlled RAG-container is considered as support for human interaction, creativity and decision making and not for replacing human decision making and creativity by a machine-based approach [Lu et al., 2024].

* jonah.schuster@rptu.de

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22904930

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

METHODOLOGY

In today's fast-paced world, the ability to access and provide critical information rapidly and accurately is essential for real-time use-cases. Yet, even in less time-constrained contexts like literature studies, timely, reliable and accurate information access remains crucial. Traditional methods of information retrieval and communication often fall short in these situations due to their reliance on visual interaction, which may not always be feasible or safe, especially under stress or in hands-free environments.

This paper proposes a methodology that leverages open source speech recognition (e.g., for German language [Milde and Köhn, 2018]) and speech synthesis technologies to enable natural, hands-free interaction with an open search infrastructure. This framework offers a powerful solution for an intuitive, multi-modal access to transparent, reliable and accurate knowledge — without requiring manual device operation — and thereby supporting real-time decision-making and interdisciplinary collaboration across diverse domains. The framework consists of three key components: First, speech recognition to convert spoken queries into text. Secondly, language generation as core component. And lastly speech synthesis to reconvert the responses into audibly receivable information. The first and last components are based on open source implementations, such as Kaldi, Mozilla TTS, and MaryTTS.

The mentioned second core component consists of a LLM with an integrated RAG-system to optimize prompt strings [Nagano et al., 2025] as well as generate the linguist response. Crucially, RAG-systems, powered by LLMs and trained on domain-specific corpora, can serve as shared knowledge containers [Oche et al., 2025]. The true power of RAG therefore lies in its ability to integrate expert knowledge from multiple disciplines into a single, accessible system. For instance, a linguist's analysis of warning language, a traffic engineer's model of evacuation routes or a geoscientist's risk assessment can all be encoded into a shared knowledge base [Odubola et al., 2025]. With regard to a specific application, these models develop a nuanced understanding of use-case-related terminology and domain-specific complexity through pretraining and grounding in curated knowledge resources. Therefore, the system facilitates collaboration between different research domains and supports by overcoming barriers of varying nomenclature or conventions. Thus, through RAGs, interdisciplinary knowledge becomes dynamically retrievable: Stakeholders from different backgrounds can query it in natural language, the system retrieves relevant information and generates clear spoken or text-based responses. This makes it possible for users with little domain knowledge to understand complex data, outside their domain. The same logic applies beyond immediate crisis response, where interdisciplinary collaboration depends not only on technical interoperability but also on the meaningful integration of different fields of knowledge. "Silo knowledge" where each domain carries expertise on its own and is not brought to the necessary exchange, can thus be broken up.

In addition, LLMs can be used to improve speech recognition in low-resource settings [Nagano et al., 2025, Granitzer et al., 2024]. The LLM interprets the query, retrieves relevant information via RAG, and generates a clear, spoken or text-based answer.

Data sovereignty and detection of bias are critical challenges for society [Spinde et al., 2021]. This includes competence in index data topics and control, output transparency and reproducibility, data security, autonomy and compliance. Otherwise, risks such as exposure of sensitive data, compliance violations, query leakage or non-validation of index data may arise. Sovereignty over the content of RAG-containers is essential to ensure autonomous, reliable and transparent delivery of critical information — particularly in sensitive, time-critical contexts such as disaster management and traffic flow prediction. Data mining with LLMs and its performance analysis provide a foundation for domain-specific management of resources [Al Zubaer et al., 2023].

This paper therefore discusses this methodology, the integration of a RAG-system, and its potential use-cases in disaster management and traffic flow prediction. In addition, the methodology also applies to non-real-time domains such as literature studies. In this case, the focus shifts from immediate decision support to exploratory, evidence-based interaction with curated corpora and contextual resources. The same technical components — speech recognition, retrieval, LLM-supported response generation, and speech synthesis — can support access, navigation, and dialogue-based analysis, while domain governance, transparent provenance, and verification procedures ensure that interpretive responsibility remains with the human user. By the integration of speech recognition and synthesis the proposed framework allows experts to interact with information naturally, without manual device operation [Sharma et al., 2003].

Furthermore, it enables necessary interdisciplinary collaboration in every thinkable use-case and supports an approach to ensure the accessibility and interoperability of required knowledge. By sharing RAGs and platforms, enabling interoperable knowledge systems, and providing tools that reduce cognitive load especially in time-constrained scenarios, this approach transforms RAG-systems into useful interdisciplinary knowledge platforms as they democratize access to expert knowledge. Domain experts contribute their insights once, and the system enables all others to benefit from them instantly in a simple, intuitive way. It ensures that decisions are not only fast but also supported by the best available, cross-domain evidence and that expert knowledge is not locked in silos, but made universally and low-threshold accessible. The idea becomes particularly exciting when one expands the framework and considers RAGs not just from the perspective of individuals, but also from the standpoint of working groups or entire subject-specific universities towards an open-search infrastructure.

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22904930

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

Software Design

The software implementing the described framework should adhere to key design principles and provide key capabilities to ensure its effectiveness, usability, and reliability. These principles and capabilities are crucial for developing a system that can reliably provide critical information in real-time. Such principles and capabilities are presented in this section.

User-Centric Design: The system should be designed with the user in mind, focusing on simplicity, clarity, and ease of use. This includes intuitive navigation, clear and concise information presentation, and minimal cognitive load [Gajos and Chauncey, 2017] for the user. Since conversational and avatar-like interfaces can increase overtrust, grounding, disclosure, and accountability should be treated as design requirements rather than optional safeguards.

Real-Time Data Processing: Given the nature of disaster management and traffic flow prediction, the system must be capable of processing and responding to real-time data. This involves efficient data retrieval, processing, and synthesis to provide users with the most current and relevant information. If users need to process visual information in situations where that's not feasible or unsafe (e.g., traffic) then additional information can be provided via audio (speech synthesis).

Privacy and Security: The extreme sensitivity of voice data motivates a privacy-preserving approach retaining user data sovereignty. Furthermore, sovereignty over search index data is crucial as well to prevent loss of data control, autonomy and transparency.
Thus, the system should incorporate robust security measures to protect both user and index data, including encryption of voice data, secure and sovereign data storage, incorporating the concept of data minimalism in processing, transmitting, storing and sharing of data and adherence to privacy regulations (e.g., General Data Protection Regulation [European Union, 2016], California Consumer Privacy Act [California Legislature, 2018]). This includes local processing whenever possible. For necessary data transfers additional measures should be taken, such as implementing differential privacy and homomorphic encryption techniques.

Error Handling and Recovery: The system should be designed to handle errors gracefully, providing clear and helpful feedback to users when errors occur. Quality of information (e.g., spatial data quality) is part of the response, as described in [Rapp, 2025]. This includes handling speech recognition errors, network connectivity issues, and other potential failures if the quality of the response does not find appropriate match. Weak responses with low confidence in quality can be excluded from being an output to users (critical response) or a warning is attached to the message, that the data quality is poor.

Scalability and Performance: The system should be designed to be able to scale easily with increasing user demand, ensuring responsivity and function even under heavy usage. This involves leveraging edge computing, optimizing database queries, using efficient algorithms for speech recognition and synthesis and implementing adaptive load balancing techniques. The RAG-system should be operational even if network connectivity fails by storing relevant parts of the RAG-system locally. Furthermore, the RAG-systems themselves can take the shape of federated systems depending on collaboration scale and audience.

Accessibility and Language: The system should be designed to be accessible to users with disabilities, should recognize and answer in the user language. All functionalities can be accessed via voice commands, retaining usability in situations when visual or haptic interaction is not possible, including high-stress or hands-free environments.

Continuous Updates and Feedback: The system should be designed with the ability to receive continuous updates (e.g., new risk maps, updated traffic data) and user feedback. This allows improvement of the system over time based on experience from real-world usage, sensor network data and collaborative user input (e.g., collaborative mapping as context information for retrieval).

Context-Aware Information Provision: The system should be able to provide contextually relevant information based on the user's location, situation, and preferences. This involves integrating with location services, understanding user context through speech or other input, and using this context to tailor the information provided.

Testing and Validation: Rigorous testing and validation should be conducted across various scenarios, including different languages, accents, and environmental conditions (e.g., noisy environments). This ensures the system's reliability and effectiveness in real-world situations.

ILLUSTRATIVE USE-CASES

As mentioned above real-time (disaster management and traffic flow prediction) and non-real-time (literature studies and literary analysis) use-case-scenarios will be considered according the application of the proposed framework.

Disaster Management & Traffic Flow Prediction: Both disaster management and traffic flow prediction are inherently complex topics that demand coordinated, interdisciplinary efforts across technical, social, and organizational domains [Twigg et al., 2015]. This requires overcoming entrenched disciplinary silos and fostering mutual understanding to enable effective interdisciplinary collaboration, as isolated knowledge is often not sufficient to face upcoming

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22904930

challenges [Twigg et al., 2015]. This is of special importance in stressful and time-critical scenarios which frequently occur in disasters [Twigg et al., 2015, Alexander, 2002]. For disaster management this implies, that the ability of diverse stakeholders — such as engineers, emergency responders, linguists, urban planners, and affected communities — to collaborate effectively in real time is crucial [Alexander, 2002]. A fast access of every involved individual to large amount of data in a time-critical situation is required [Sharma et al., 2003, Alexander, 2002]. This could be achieved by the proposed methodology by providing a RAG-container including for example historical disaster reports, humanitarian response manuals, emergency protocols, and crisis communication records for disaster management. Additionally it could corporate historic and real-time traffic and weather data, real-time user data, geospatial or infrastructure data, Internet of Things data, event and calendar data.
As mentioned above, by the integration of domain-specific RAG-systems, it is possible to overcome thresholds in accessing necessary information and knowledge as well as enhancing of interdisciplinary interaction. This may be especially effective in disaster scenarios, where decision-makers and civilians are dependent on data provided by experts. These possibly benefit from time savings by removing the necessity to simplifying their findings into a comprehensible form for a broad non-domain and non-academic audience.
An example for interdisciplinary collaboration is given by the illustrative interaction of disaster management, traffic flow prediction and linguistics: Critical infrastructure such as roads, airports, and rail lines — so-called "lifelines" — must remain functional during crises to enable evacuation and aid delivery [Alexander, 2002]. However, their efficiency depends not only on physical resilience but also on dynamic traffic flow prediction, what shows that effective response thus requires integration of traffic flow modeling, urban planning, and real-time coordination.
In parallel, linguistics is essential for clear, actionable communication. Risk warnings and emergency instructions must be linguistically precise to avoid misinterpretation. Conversely, many expert assessments rely on qualitative "fuzzy" judgments (e.g., "high likelihood," "severe impact") that must be translated into quantifiable parameters for risk models [Romeike and Hager, 2020].
The sovereignty over search index data, which is provided by the system, is of great importance with regard to these real-time use-cases: In both domains the processing of sensitive data is often necessary. In disaster management this data includes medical, political, personal and disaster victim related data, in traffic prediction personal user data, especially location data [Alexander, 2002, Twigg et al., 2015]. Given the extreme sensitivity of input data in disaster events [Alexander, 2002], sovereignty of users over this data is crucial.

Furthermore, for real-time applications frequently found in both domains, reliability of the system is paramount, as there's often no possibility for intervention if an error occurs. Sovereignty over search index data directly aids this goal by limiting above risks.

RAG-systems can enable enhanced model output transparency for machine learning models, aiding training and improvement of systems in both domains, as errors can be traced more easily. This transparency can lead to better user acceptance by increasing credibility through reasoned suggestions, e.g., for routing in traffic flow prediction [Vaziri and Lam, 1983] or decision making in disaster management. Furthermore, in case of disasters, provision of risk map data for routing purposes can be highly beneficial. For example, this could be implemented by adding route risk as an objective to user navigation that should be minimized [Zhong et al., 2020]. Reliability and acceptance of this data are key, highlighting the suitability of RAG-systems for this purpose.
Voice interaction with machines provides a more natural, efficient and less error-prone way than text inputs and has the potential to retain access to information in scenarios, where only hands-free interaction with a device is feasible or safe [Sharma et al., 2003]. This frequently occurs in driving and disaster scenarios, highlighting the benefit of voice interaction. It enables the use of the system as a driver assistant, for information retrieval and possibly in dual use as a safety component [Huang et al., 2024] for example for detection of driving fatigue.

Literature Studies & Chatbot-Driven Literary Analysis: A literary use-case can begin with a private, curated RAG-container maintained by an individual scholar and later shared with a seminar or research group. By combining primary texts with contextual and research materials, such a container creates a controlled environment for evidence-based interaction that supports scholarly engagement without replacing reading or interpretation.

Within this retrieval-grounded setup, chatbot-based dialogue becomes productive only if it is tied to explicit evidence-based moves such as quoting, locating, justifying, and revising textual observations. Literary interaction with artificial intelligence (AI) should therefore be linked to reflective "step-back" routines and to a form of perspectival mistrust that keeps users attentive to textual detail, alternative readings, and the need for verification rather than relying on conversational plausibility alone [Lösener and Magirius, 2024]. Prompt design is thus not merely a technical interface feature, but a pedagogically and analytically relevant element of interaction design, since AI-supported literary reading and writing depend on how tasks, expectations, and forms of response are structured [Führer and Gerjets, 2024]. In this way, the design of interaction becomes part of the didactic arrangement rather than an external layer added to model output.

Core functions include passage retrieval, quotation, read-aloud and navigation support, as well as analytic prompting for hypothesis formation, comparison, and revision on the basis of textual evidence. Visual and auditory information in conjunction can also support literary studies, for example when a passage from E. T. A. Hoffmann's *Der Sandmann* is read aloud while users focus on a visual sketch that con-

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

https://doi.org/10.5281/zenodo.22904930

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

centrates attention, supports text-based imagination, and sharpens perception of mood, ambiguity, and projection. Voice-based access is thus not a replacement for reading, but a low-threshold means of navigating a curated corpus and structuring text-based interaction with complex literary materials. Retrieval and quotation thereby form part of a didactic regime that preserves interpretive authority through evidence, revision, and reflective checks rather than delegating it to the machine [Kühn, 2025]. In teaching, this can scaffold close reading; in research, it can support orientation in larger corpora while still requiring textual verification.

The same setup can also contribute to interdisciplinary exchange: Romantic nature poetry, for example, may be read as a historically specific mode of perceiving and valuing the non-human world, while environmental knowledge helps to situate and critically assess such literary constructions without reducing them to factual illustrations.

To remain reliable, the system should make source provenance transparent, distinguish retrieved material from inferred output, and organize quality control through explicit verification checkpoints. From an ethics and governance perspective, this also implies a form of digital text sovereignty: Users must remain aware of model limitations, source status, and the difference between corpus-grounded support and generated inference [Frederking, 2023].

CONCLUSION

The proposed methodology leverages open source speech recognition and speech synthesis technologies to enable natural language interaction with an open search infrastructure across different domains. The open search infrastructure consists of a LLM in combination with a RAG-system. The integration of a RAG-system containing curated knowledge across multiple domains, including disaster management, traffic flow prediction and literature studies allows contextually relevant, accurate, controlled, and transparent information provision. The use-cases of disaster management, traffic flow prediction, and literature studies demonstrate the potential of this approach in both real-time and non-real-time scenarios. While the first two highlight the value of voice-based interaction for mobile spatial decision support and hands-free access to critical information, the literary use-case shows how curated corpora and retrieval-grounded dialogue can support exploratory, evidence-based analysis. The combination of the domain specific knowledge systems into an interdisciplinary RAG-system, its application and potential benefits were described. The domain specific RAG-systems are created by experts, who preselect and concentrate the knowledge base of their respective system on relevant information with high quality. This can help to increase the relevance, transparency and reliability of the obtained model outputs, e.g., by minimizing hallucinated results of the LLM based on the RAG-system.
Additionally, the RAG-system is connected with locally installed open source speech recognition and speech synthesis systems. By linking the collaborative and interdisciplinary RAG-system with local speech recognition and synthesis systems, the proposed system allows the users to retain sovereignty over both their voice data as well as the search index data, leading to a privacy-friendly system. Furthermore, privacy-friendly in this scenario means, that speech data does not leave the local user computer infrastructure and the RAG-container is under the control of the individual or the collaborative team.
The described methodology can be used on a larger scale to develop an index for interdisciplinary search in a bottom-up-approach, whereby the final index is derived from several different domain-specific search indexes.
Future work may include further refinement of the system, testing in real-world settings, and the systematic evaluation across different domains. In addition to disaster management and traffic flow prediction, future research should further explore non-real-time applications such as literature studies, where curated corpora, retrieval-grounded dialogue, and evidence-based interaction raise distinct methodological and design questions. Beyond the use-cases discussed in this work, the proposed open search ecosystem may also be extended to arbitrary other fields, such as education or healthcare.
The open source character of the used technologies enables community contribution, which may further strengthen the system's capabilities and applicability. Overall, the methodology is designed to be user-centric, responsive, secure, and accessible, while remaining adaptable to the differing requirements of domain-specific and interdisciplinary open search ecosystems.

Key Takeaways

- Domain-Adaptive Information Access: The proposed system enables voice-based access to domain-specific knowledge in both real-time and non-real-time scenarios, ranging from disaster management and traffic flow prediction to literature studies.
- Integration with Controlled RAG-Systems: The integration with domain-specific RAG-systems ensures that retrieved information remains contextually relevant, controlled, and adaptable to different use-cases.
- Governance, Privacy, and Transparency: The methodology emphasizes privacy-focused design, domain governance, and transparent handling of retrieved versus inferred information in order to support trustworthy interaction.

REFERENCES

[Al Zubaer et al., 2023] Al Zubaer, A., Granitzer, M., and Mitrović, J. (2023). Performance analysis of large language models in the domain of legal argument mining. *Frontiers in artificial intelligence*, 6:1278796.

[Alexander, 2002] Alexander, D. (2002). *Principles of emergency planning and management*. Terra Publishing, Edinburgh, Scotland.

https://doi.org/10.5281/zenodo.22904930

[California Legislature, 2018] California Legislature (2018). California consumer privacy act of 2018. California Civil Code § 1798.100 et seq., https://california-ccpa.org/.

[European Union, 2016] European Union (2016). General data protection regulation (gdpr) (regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (data protection directive)). Official Journal of the European Union, L 119, 4.5.2016, pp. 1–88, https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[Fathullah et al., 2024] Fathullah, Y., Wu, C., Lakomkin, E., Jia, J., Shangguan, Y., Li, K., Guo, J., Xiong, W., Mahadeokar, J., Kalinli, O., et al. (2024). Prompting large language models with speech recognition abilities. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13351–13355. IEEE.

[Frederking, 2023] Frederking, V. (2023). Von Fake News bis ChatGPT: Digitale Textsouveränität als ethisch-politische Bildungsaufgabe für Deutschdidaktik und Deutschunterricht in der digitalen Welt. *MiDU – Medien im Deutschunterricht*, 5(2).

[Führer and Gerjets, 2024] Führer, C. and Gerjets, P. (2024). How to Understand & Write Literature with AI? Potenziale und Risiken von KI-Tools für literarisches Lesen und Schreiben. *MiDU – Medien im Deutschunterricht*, 6(1):1–18.

[Gajos and Chauncey, 2017] Gajos, K. Z. and Chauncey, K. (2017). The influence of personality traits and cognitive load on the use of adaptive user interfaces. In *Proceedings of the 22nd international conference on intelligent user interfaces*, pages 301–306.

[Granitzer et al., 2024] Granitzer, M., Voigt, S., Fathima, N. A., Golasowski, M., Guetl, C., Hecking, T., Hendriksen, G., Hiemstra, D., Martinovič, J., Mitrović, J., 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):512–520.

[Huang et al., 2024] Huang, S., Zhao, X., Wei, D., Song, X., and Sun, Y. (2024). Chatbot and fatigued driver: Exploring the use of llm-based voice assistants for driving fatigue. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8.

[Jain et al., 2025] Jain, A., Cui, L., and Chen, S. (2025). Aligning llms for the classroom with knowledge-based retrieval: A comparative rag study. In *2025 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, pages 1–8. IEEE.

[Kühn, 2025] Kühn, W. (2025). "Poesie-Programmierer aller Länder, vereinigt euch!" Zur Gestaltung interaktiver Lernumgebungen im Literaturunterricht. *Wirkendes Wort*, 75(1):125–135.

[Lösener and Magirius, 2024] Lösener, H. and Magirius, M. (2024). Perspektivisches Misstrauen beim literarischen Verstehen: KI-gestützte Gespräche mit Figuren aus Kafkas "Ein altes Blatt". In Bernhardt, S., editor, *Unzuverlässiges Erzählen in Literatur und Medien*, pages 385–405. Frank & Timme, Berlin.

[Lu et al., 2024] Lu, L.-C., Chen, S.-J., Pai, T.-M., Yu, C.-H., Lee, H.-y., and Sun, S.-H. (2024). Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *arXiv preprint arXiv:2405.06373*.

[Milde and Köhn, 2018] Milde, B. and Köhn, A. (2018). Open source automatic speech recognition for german. In *Speech communication; 13th ITG-symposium*, pages 1–5. VDE.

[Nagano et al., 2025] Nagano, T., Kurata, G., Thomas, S., Kuo, H.-K. J., Bolanos, D., Jung, H., and Saon, G. (2025). Llm based text generation for improved low-resource speech recognition models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[Neumann et al., 2024] Neumann, A. T., Yin, Y., Sowe, S., Decker, S., and Jarke, M. (2024). An llm-driven chatbot in higher education for databases and information systems. *IEEE Transactions on Education*, 68(1):103–116.

[Oche et al., 2025] Oche, A. J., Folashade, A. G., Ghosal, T., and Biswas, A. (2025). A systematic review of key retrieval-augmented generation (rag) systems: Progress, gaps, and future directions. *arXiv preprint arXiv:2507.18910*.

[Odubola et al., 2025] Odubola, O., Adeyemi, T., Olajuwon, O., Iduwet, N., Aniekan, A., and Odubola, T. (2025). Ai in social good: Llm powered interventions in crisis management and disaster response. *Journal of Artificial Intelligence, Machine Learning and Data Science*, 3(1):3353–3360.

[Rapp, 2025] Rapp, J. (2025). *Human diseases and exposure to pesticides in developing countries: temporal spatial analysis, risk mitigation strategies and spatial decision support system*. PhD thesis, Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau.

[Ritter, 2021] Ritter, E. (2021). Your Voice Gave You Away: The Privacy Risks of Voice-Inferred Information. *Duke University School of Law*.

[Romeike and Hager, 2020] Romeike, F. and Hager, P. (2020). Methoden zur Identifikation und Bewertung von Risiken. In *Erfolgsfaktor Risiko-Management 4.0*, pages 143–199. Springer Fachmedien Wiesbaden, Wiesbaden.

[Sharma et al., 2003] Sharma, R., Yeasin, M., Krahnstoever, N., Rauschert, I., Cai, G., Brewer, I., MacEachren, A. M., and Sengupta, K. (2003). Speech-gesture driven multimodal interfaces for crisis management. *Proceedings of the IEEE*, 91(9):1327–1354.

[Spinde et al., 2021] Spinde, T., Rudnitckaia, L., Mitrović, J., Hamborg, F., Granitzer, M., Gipp, B., and Donnay, K. (2021). Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management*, 58(3):102505.

[Twigg et al., 2015] Twigg, J. et al. (2015). Disaster risk reduction. *Good Practice Review*, 9.

[Vaziri and Lam, 1983] Vaziri, M. and Lam, T. N. (1983). Perceived Factors Affecting Driver Route Decisions. *Journal of Transportation Engineering*, 109(2):297–311.

[Zerhoudi and Granitzer, 2024] Zerhoudi, S. and Granitzer, M. (2024). Personarag: Enhancing retrieval-augmented generation systems with user-centric agents. *arXiv preprint arXiv:2407.09394*.

[Zhong et al., 2020] Zhong, S., Cheng, R., Jiang, Y., Wang, Z., Larsen, A., and Nielsen, O. A. (2020). Risk-averse optimization of disaster relief facility location and vehicle routing under stochastic demand. *Transportation Research Part E: Logistics and Transportation Review*, 141:102015.

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22916430

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

"HYBRID SEARCH ECOSYSTEMS: AN EMPIRICAL ANALYSIS OF ONLINE SEARCH BEHAVIOR OF GENERATION Z AND ALPHA IN GERMANY"

Marie Einsiedler, Technische Hochschule Ingolstadt, 85049 Ingolstadt, Germany
Alexander Decker, Technische Hochschule Ingolstadt, 85049 Ingolstadt, Germany

Abstract

The dominance of traditional web search engines is increasingly challenged by social media platforms, marketplaces, and AI-based search systems, particularly among younger generations. This phenomenon, often referred to as the "search shift," reflects a structural transformation in online information seeking. However, existing research remains fragmented, focuses primarily on individual platforms, and lacks integrative models and empirical evidence, especially for Generation Alpha.

This study addresses this gap by developing and empirically validating an adapted model of the online search process that incorporates emerging platforms such as social media and AI applications. The model integrates influencing factors including credibility, experience, habits, preferences, and search intention. To validate the model and examine platform choice and search behavior, a quantitative online survey was conducted with 403 participants from Generation Z and Generation Alpha in Germany.

The results show that traditional search engines remain the most frequently used platform, particularly for fact-based and credibility-sensitive queries. However, social media platforms such as TikTok and YouTube play a central role in exploratory and inspiration-driven searches, while AI-based search systems are increasingly used for dialog-based and complex information queries. Platform choice is strongly influenced by perceived credibility, prior experience, and usage habits. Furthermore, younger generations demonstrate a preference for visual, personalized, and efficient search processes, indicating a shift toward a hybrid search ecosystem.

This study contributes to information behavior research by refining the online search process model to reflect contemporary multi-platform search environments and by providing empirical evidence on the evolving search behavior of Generation Z and Alpha.

BACKGROUND AND METHODOLOGY

For more than two decades, traditional web search engines have served as the primary gateway to online information. However, recent technological and societal developments have significantly transformed the digital information environment. Social media platforms, marketplaces, and AI-based search systems increasingly offer alternative pathways to information access. This shift is particularly evident among younger generations, who have grown up in highly digital and platform-integrated environments.

This development is often described as a "search shift," referring to a transition from exclusive reliance on traditional search engines toward a more diverse and platform-integrated search ecosystem [10]. Younger users increasingly use social media platforms such as TikTok or YouTube for exploratory searches, marketplaces for product-related queries, and AI-based systems such as ChatGPT for dialog-based information retrieval and learning [23]. These platforms do not merely complement traditional search engines, but fundamentally reshape how users discover, evaluate, and interact with information.

Despite growing academic and practical interest in this phenomenon, existing research remains fragmented. Many studies focus on individual platforms or specific aspects of search behavior, such as social media search or AI-based search systems, without providing an integrative perspective on the overall search process. In addition, the role of modern search queries has hardly been considered in research [11, p. 71]. Hence, empirical research on this topic –especially in the German context–remains limited. In particular, Generation Alpha has received little attention so far, despite being the first generation to grow up fully immersed in AI-mediated and platform-integrated digital environments [14].

To address this research gap, this study develops and empirically validates an adapted model of the online search process that reflects contemporary hybrid search environments. To summarize, the study addresses the following research question:

How do Generation Z and Generation Alpha in Germany search for information online, and which factors influence their platform choice and search behavior? [1]

To answer this question, a systematic literature analysis was conducted to develop a theoretical model of the online search process. This model was subsequently validated using empirical data from a quantitative online survey with 403 participants from Generation Z and Generation Alpha in Germany.

[1] A detailed description of the two generations has been omitted here. It is assumed that readers are already familiar with them.

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22916430

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

THEORETICAL BACKGROUND AND MODEL DEVELOPMENT

In order to obtain findings for the research area mentioned, a systematic literature analysis was carried out based on the PRISMA method [see 19]. The PRISMA flow chart describes the specific procedure for selecting literature, including inclusion and exclusion criteria.

In a first step, the PRISMA process was carried out to create the relevant sources for the theoretical basis. Building on this, a targeted search for studies on current online search behavior was carried out in a second step. The complete flow chart, including the databases used, the keywords, and the inclusion and exclusion criteria, can be found in Appendix A. As a result of the process, ten studies were pre-selected and four studies were finally selected.

The results of the systematic literature analysis form the basis for the more in-depth aspects of the research project described below. These are, on the one hand, the (further) development of the model of the online search process and, on the other hand, the analysis of the identified studies.

Model of the online search process

Various models of the information process have been developed in the literature in the last decades, covering different focal points and process phases. Due to the changed search behavior of younger generations, the aim of this paper is to outline an adapted online search process for this target group.

A frequently used ideal-typical model is the online research process according to Riemer and Brüggemann [22, similar: 11, p. 68]. This serves as the basis for the present work, as it contains the central process steps that also occur in other models.

What is missing in most models is an explicit examination of possible factors influencing the online search process. In this context, Wilson's general model of information behavior [25] seems to be of particular interest, as it explicitly takes various influencing factors into account.

Process steps in the model

Once a need for information arises, the first step in the model for the user is to select a platform. In addition to traditional web search engines as presented by Riemer and Brüggemann [22, p.151], younger generations are now increasingly using social media, AI applications, and marketplaces such as Amazon, or comparison portals. Next, the search query is formulated, typically in the form of text, voice, or image input [11, p. 66]. In the original model the information system then performs the search and delivers various information objects (presentation). Since the adapted model shall take in the perspective of the user and in analogy to Lewandowski's model [11, p. 68,], these two steps were swapped with another step–the selection of information provided by the search engine. In the end, the searcher reviews and evaluates according to content type (e.g., document, image or video) and source.

Riemer and Brüggemann's model [22, p.151] is linear and ends after one search, whereas the real search process is dynamic and iterative, as results can lead to adjustments or repetitions of individual steps [11, p. 68; 21, pp. 85–120]. Hence, if the searcher is satisfied, the process ends. Otherwise, based on the experience, they can adjust, repeat, or cancel individual steps or the entire search process.

Influencing factors

The online search process is individual and influenced by various factors. In his general model of information behavior [25], Wilson includes several influencing factors. Most important are demographic (e.g., age), psychological (e.g., interest), and cognitive factors (e.g., prior knowledge and expertise) as well as the characteristics of the information source (e.g., credibility) and search intentions. Since previous research on online search has made little reference to these aspects (see chapter Background), it is important to take them into account in our model as well.

First of all, personal charakteristics of the searcher have a significant influence on the search process. In addition to sociodemographics, these include experience and skills [21, pp. 85–120] and the resulting search habits [11, pp. 208–209].

In addition to these internal factors, factors of the search intent also play a role. In particular, platform choice and search behavior depend on the topic and search intention [4], the later is divided into informational, navigational, and transactional searches [5]. Besides that, it was found, that the perceived credibility of the medium, the source, and the message plays a central role in online search [15, pp. 309–311]. Figure 1 provides an overview of the model described here.

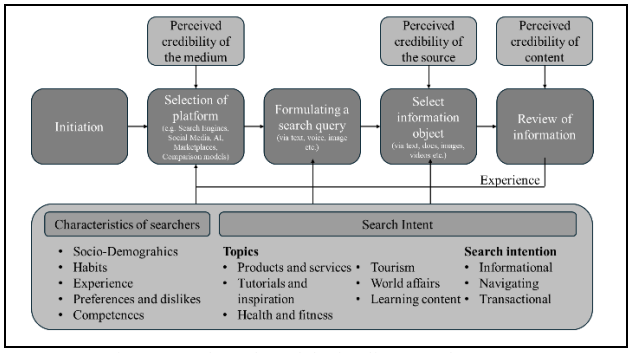


Figure 1: Adapted model of online search process

As a first conclusion, compared to traditional models of the online search process, the adapted model developed in this study introduces three key extensions:

First, platform selection is explicitly modeled as an independent and central step, reflecting the increasing diversity of search platforms such as social media and AI systems. Second, the model explicitly integrates influencing factors such as credibility, experience, habits, and user preferences, based on Wilson's general model of information behavior. Third, the model reflects the iterative and multi-platform nature of contemporary search behavior, allowing users to switch between platforms during the search process. These extensions enable a more accurate representation of search behavior in hybrid search ecosystems.

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22916430

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

Study comparison

The current state of research on the online search behavior of younger generations is characterized by high dynamism and fragmentation. While traditional information systems dominated for a long time, recent studies show an increasing diversity of search platforms in use. The study comparison in this paper is based on several quantitative surveys from 2020 to 2024, including State of Search [4], the Digital 2023 Global Overview Report [8], and studies on the use of social media as search engines [1; 6].

The comparative analysis reveals several key trends that shape the search behavior of Generations Z and Alpha. These have been assigned to the online search process shown in Figure 1 and are summarized in Figure 2.

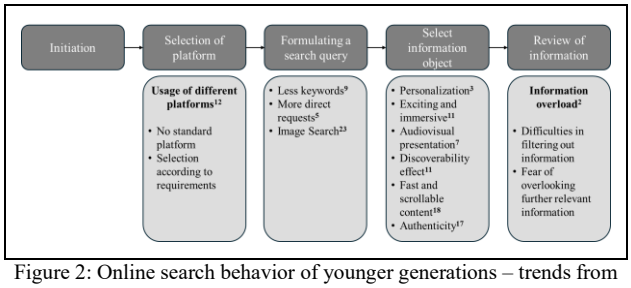


Figure 2: Online search behavior of younger generations – trends from recent studies [sources: 2/3/5/7/9/11/12/17/18/23]

One key trend is the "**search shift**" already described above. Traditional search engines are losing their role as the sole starting point, as search is increasingly being integrated into existing media usage routines. The studies confirm, that young users are increasingly searching on social media, in AI applications, or on marketplaces, depending on their information needs and where they are.

The choice of platform is highly **context-dependent**: web search engines are mainly used for fact-based and trust-sensitive topics such as news or health, while social media dominates exploratory and inspiration-driven topics such as lifestyle or travel. Marketplaces and comparison portals play a central role in transactional search queries in e-commerce.

In addition, the use of **AI-based search systems** is increasing, especially for quick and dialog-based information queries. However, these mostly complement traditional search engines rather than completely replacing them. At the same time, the **formulation of search queries** is changing, especially in comparison to older users [20]: Young users use **fewer keywords** and increasingly **direct requests**, especially in the context of AI applications [5; 9; 18]. **Image search** is also gaining in importance and leading to a more visual search process [22].

In addition, younger generations prefer an **entertaining and exploratory search process** that enables inspiration and new content ("**discoverability**") [13, p. 87; 16]. At the same time, they expect **fast, simple, and mobile-optimized** answers in short, scrollable formats [18]. Finally, young users want **personalized results** that make it easier to find information and reduce cognitive effort.

One negative trend is what is known as "**information overload**." The large amount of information makes it difficult for users to identify relevant and credible content and often leads to prolonged search processes due to concerns about overlooking important information [2, p. 4].

EMPIRICAL STUDY

The aim of the empirical survey was to validate and expand on the findings of the literature review.

Study design

The quantitative online survey is based on the model of the online search process shown in Figure 1. Accordingly, the empirical study focuses on **platform selection** and its **influencing factors**, and the **formulation of search queries**. In addition, the study examines whether younger generations experience **information overload**. For all of these points, the aim is to investigate whether and what differences exist between Gen Z and Gen Alpha.

The population comprises Gen Z and Gen Alpha individuals living in Germany, which numbered around 23.51 million in 2023. People born between 1997 and 2012 were surveyed in order to cover the entire Gen Z and the older part of Gen Alpha. This age limit was chosen because young people aged 12 and above use the internet predominantly independently and are able to answer the questionnaire.

To approximate the population distribution, the sample was weighted based on birth rates. This resulted in a ratio of 84% Gen Z and 16% Gen Alpha, with target values of at least 325 and 60 participants, respectively. These were achieved with n = 337 (Gen Z) and n = 66 (Gen Alpha). The survey was conducted in January 2025.

The questionnaire was based on the model of the online search process and the research focus areas described above. The questionnaire included Likert-scale items measuring perceived credibility, experience, and platform preferences, based–where applicable–on established operationalizations from Metzger et al. [15] or Wang and Lind [24]. However, the sample is not fully representative, as the majority of respondents come from the same region and have an above-average level of education, and there is a high proportion of students. These limitations must be taken into account when interpreting the results.

Results from the empirical study

Platform selection in search context

To examine platform usage, only the responses of respondents who use the platform at least once a year were considered. Table 1 shows the distribution of platform usage in a search context in percent.

The results in Table 1 show that web search engines are used most intensively. With 91% of weekly users, it is clear that despite all the criticism, web search engines are still the number one online search platform. AI search engines or AI chatbots are the second most frequently used search tools. TikTok is also used regularly by many respondents

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22916430

to search for information. There was also a significant difference between Gen Z and Alpha in terms of web search engines, Instagram, marketplaces, and comparison sites. The respective platforms were used significantly more often by Gen Z. Gen Alpha uses TikTok more often, but the differences here were not significant.

in %, rounded	weekly	monthly	Diff. Gen Z/α
Search Engines (SE)	91	98	p<0,001
Instagram (IG)	31	57	p<0,001
TikTok (TT)	47	65	n.s.
YouTube (YT)	28	61	n.s.
AI-Applications (AI)	52	62	n.s.
Marketplaces (MP)	20	60	p<0,001
Comparison portal (CP)	7	29	p<0,001

Table 1: Results "Usage of platform in search context"[2]

Influencing factors on platform selection

The following section examines the factors identified in the model in Figure 1 that influence the choice of platform.

Habits

Habits can influence platform choice. This includes the assumption that platforms that are frequently used in general are also used more often for search queries [11, p. 209]. The following hypothesis was derived from this:

H1: The more frequently the platform is used in general, the more often it is also used in a search context.

The null hypothesis (*no correlation between general use and use in a search context*) was tested using Spearman's Rho rank correlation. For all platforms, the probability of error was below the significance level of 5%, so the null hypothesis was rejected. There is therefore a significant positive correlation between general use and use in a search context, with correlation values between 0.414 (comparison portals) and 0.799 (TikTok). The results confirm that habits influence platform choice (results in Appendix B1).

Experience

According to the literature, experience with platforms influences platform choice in the search context [21, pp. 85–120]. This leads to the following hypothesis:

H2: The better the experience with the platform, the more often it is used in the search context.

Web search engines are predominantly rated as "very good" to "good" and continue to offer a suitable search experience for younger generations. AI search engines and chatbots also receive positive ratings. YouTube, TikTok, marketplaces, and comparison portals are rated "good" on average, while Instagram scores only "mediocre." There are no significant differences between Gen Z and Gen Alpha. The results are shown in Table 2.

The correlation hypothesis H2 was tested for each platform using Spearman's Rho rank correlation. For all platforms, the probability of error was less than 5%, so the null hypothesis (*no correlation between experience and use in the search context*) was rejected. There is a positive correlation (correlation values between 0.180 for web search engines and 0.871 for TikTok). This shows that platforms with better experiences are used more frequently for online searches (detailed results in Appendix B2).

	∅All	∅Gen Z	∅Gen A
Search Engines	1,58	1,54	1,79
Instagram	2,67	2,68	2,61
TikTok	2,27	2,31	2,08
YouTube	2,09	2,10	2,02
AI-Applications	1,83	1,84	1,77
Marketplaces	2,14	2,12	2,24
Comparison Portals	2,15	2,13	2,37

Table 2: Results"Experiences"
Scale from 1 = very good to 4 = bad experiences

Credibility

The online search process model (see figure 1) distinguishes between three types of credibility. However, the survey only considers the credibility of the medium, as this is the only factor that directly influences the choice of platform [15, p. 309]. This leads to the following hypothesis:

H3: The higher the perceived credibility of a platform, the more frequently it is used in the search context.

Table 3 shows the average credibility ratings on a 4-point scale. Web search engines achieve the highest scores. Around 60% rate them as "credible" or "very credible." Among social media platforms, YouTube is rated most credible, while TikTok and Instagram are predominantly perceived critically. AI search engines are considered credible by the majority. Marketplaces and comparison portals are also rated credible by around 60%, with comparison portals more often classified as "very credible."

	∅All	∅Gen Z	∅Gen A
Search Engines	3,26	3,26	3,17
Instagram	2,12	2,12	2,09
TikTok	1,77	1,77	2,09
YouTube	2,62	2,62	2,53
AI-Applications	2,64	2,26	2,53
Marketplace	2,69	2,69	2,74
Comparison Portal	2,88	2,88	2,65

Table 3: Results" Credibility of the medium"
Scale from 1 = not credible to 4 = very credible

Spearman's Rho rank correlation coefficient shows an error probability of less than 5% for hypothesis H3 on every platform, which is why the null hypothesis is rejected here as well. Due to the positive correlation coefficients, there is thus a significant positive correlation between the variables (values between 0.116 for web search engines and 0.515 for TikTok). The higher the perceived credibility, the more frequently the platforms are used in a search context (detailed results in Appendix B3).

[2] Since Facebook played virtually no role in this question or any other questions, it was not included in this paper. The data is listed in the appendix, though.

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22916430

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

Search intention

In order to determine the preferred platforms depending on the search context, respondents were presented with various scenarios with different search intentions. In most cases, web search engines (in green) were chosen as the preferred platform (see Table 4).

	All	Gen Z	Gen A
Almost all search intentions	SE	SE	SE
Exeptions			
Travel-Trip to London	SE	SE	TT
Booking accomodation	CP	CP	n.a.
Training Plan	YT	YT	n.a.
Eras Tour	TT	SE	TT
Baking Recipe	SE	SE	YT
Programming Language	YT	YT	YT

Table 4: Results "Preferred platform according to search intent" Abbreviations correspond to those in Table 1

Exceptions include scenarios such as creating a training plan or learning a programming language, where YouTube is preferred. When it comes to the topic of Eras Tour (searches related to pop culture), TikTok is slightly ahead of web search engines. When all social media platforms are combined, they are the most widely used alternative in many scenarios and are sometimes more popular than web search engines.

Neither AI search engines nor AI chatbots are the most popular platforms for any of the search intentions mentioned. This underscores that they have not yet achieved the relevance of social media or even web search engines in online searches. Nevertheless, they sometimes come in second place, which shows that these platforms are definitely used for certain topics. A comparison of the generations shows that Gen Alpha uses social media more frequently than Gen Z.

The survey design makes it also possible to determine the most popular platforms for each search intention. Since web search engines were the most frequently selected option across all search intentions, the second most popular choice is particularly interesting for analysis. Table 5 summarizes the results (values rounded, in %).

	Informational			Navigating			Transactional		
In %	All	Z	A	All	Z	A	All	Z	A
SE	44	45	35	67	66	63	56	54	71
IG	11	10	12	5	5	11	1	1	0
TT	17	16	24	1	1	2	2	2	0
YT	12	10	17	8	8	11	9	9	0
AI-SE.	2	2	1	4	4	5	3	3	0
AI-Bot	9	9	8	9	9	7	9	9	0
MP.	2	3	0,5	3	4	1	5	4	26
CP	1	1	0	0,5	1	0	13	14	0

Table 5: Results"Platform selection based on search intent" Abbreviations correspond to those in Table 1

The following key differences can be observed with regard to the second most frequently used platforms:

- **Informational search**: Social media, especially TikTok, is frequently used. The platform values are close to each other, while web search engines, despite ranking first, show comparatively weaker values.
- **Navigational search**: Among Gen Z, AI applications, especially chatbots, play the second most important role because they provide quick and personalized answers. Gen Alpha, on the other hand, uses Instagram more frequently and prefers visual content.
- **Transactional search**: Marketplaces and comparison sites are the second most important source. Gen Z prefers marketplaces, while Gen Alpha uses comparison sites more frequently.

Search intent topics (subject areas)

The choice of platform also varies depending on the subject area. Although web search engines are the most popular in all areas, they only achieve shares of between 39% (learning content Gen Z) and 65% (all, services).

Marketplaces are often used for product searches, while Gen Alpha also uses Instagram with similar frequency. For services, TikTok usually ranks second, with comparison portals leading the way in tourism and Instagram and TikTok also popular among Gen Alpha. YouTube is the second most important platform for health and tutorials, with Gen Alpha searching for tutorials more often on YouTube than on web search engines. AI chatbots also play an important role in learning content. A complete overview is provided in Appendix B4.

Preferences

Personal preferences are another factor influencing platform choice. The results show clear differences between the generations: Gen Alpha cites visual presentation as its most important preference and attaches greater importance to up-to-date information and trends. Gen Z, on the other hand, tends to prefer simple search queries and clear, concise presentation of results.

When preferences are determined for each search intention, simple search queries remain the most important platform feature across all search intentions. Informational search queries often focus on orientation or inspiration, which is why visual content such as images and videos are particularly valued. Due to these visual preferences, social media platforms are preferred for this purpose. The detailed results can be found in Appendix B5.

Search queries

In addition to platform selection, the study examined how younger generations formulate their search queries and use voice and image search.

The results show an average length of 4.3 terms per search query across all scenarios. There are significant differences between the platforms: for classic search engines, the average is 3.01 terms, while AI applications have significantly longer queries with an average of 12.33 terms. This confirms the trend toward prompt queries, in which users formulate specific tasks.

There are only minor differences between Gen Z and Gen Alpha, with Gen Alpha using slightly shorter and less frequent prompt-based queries.

https://doi.org/10.5281/zenodo.22916430

Image search is used occasionally to rarely by the majority, while voice search is predominantly not used at all. Image search is slightly more prevalent among Gen Z than among Gen Alpha.

Information Overoad

Around 85% of respondents expect all their questions to be answered on the internet, which highlights the high expectations placed on online searches. At the same time, 58% say they find it difficult to filter relevant content from the wealth of information available. Nevertheless, 60% feel only slightly overwhelmed or not overwhelmed at all.

There is a difference between the generations: Gen Alpha feels significantly less overwhelmed by the large amount of information than Gen Z. In addition, more than half of Gen Z say they often continue searching even after finding relevant information, out of concern that they may overlook important or correct content. Gen Alpha exhibits this behavior less frequently. This suggests that Gen Z evaluates search results more critically and conducts longer search processes.

DISCUSSION

The results of this study provide important insights into the evolving nature of online search behavior among Generation Z and Generation Alpha and highlight the emergence of a hybrid search ecosystem in which multiple platforms coexist and fulfill complementary roles.

First, traditional web search engines remain the most frequently used and most trusted platforms across nearly all search contexts. This finding confirms the continued relevance of search engines as central information infrastructures, particularly for fact-based, credibility-sensitive, and goal-oriented queries. The high credibility ratings and usage frequencies demonstrate that traditional search engines still represent a primary and reliable entry point for structured information seeking.

At the same time, the results clearly show that search engines are no longer the exclusive starting point of the search process. Social media platforms such as TikTok, YouTube, and Instagram play an increasingly important role, particularly for exploratory, experiential, and inspiration-driven searches. These platforms offer visual, socially validated, and algorithmically personalized content, which aligns with the preferences of younger users for efficient and engaging information access. Search is thus increasingly embedded within broader media consumption practices rather than being a distinct and isolated activity.

Furthermore, AI-based search systems and chatbots represent a fundamentally new type of search interaction. Unlike traditional search engines, which primarily provide lists of information sources, AI systems offer synthesized, context-aware, and dialog-based responses. Although AI systems are not yet the dominant platform for most search intentions, their frequent use and positive experience ratings indicate their growing importance. This suggests a transition from information retrieval toward information synthesis and interactive assistance.

The findings also confirm the importance of key influencing factors identified in the theoretical model. Perceived credibility, prior experience, and habitual platform use significantly influence platform choice. This highlights that search behavior is not solely driven by functional efficiency, but also by subjective trust, familiarity, and user expectations. In addition, preferences for visual presentation, personalization, and efficiency play an increasingly important role, particularly among Generation Alpha.

The comparison between Generation Z and Generation Alpha suggests a gradual generational transition rather than a disruptive shift. Both generations use similar platforms, but Generation Alpha shows stronger preferences for visual and socially integrated search environments. This supports the assumption that hybrid, multi-platform search behavior will become even more dominant in the future.

From a theoretical perspective, these findings extend existing models of the online search process. Traditional models typically assume a single information system, usually a search engine. However, the results of this study demonstrate that contemporary search behavior involves strategic platform selection among multiple alternatives. Platform choice has thus become a central and dynamic component of the search process, influenced by contextual, psychological, and experiential factors.

Overall, the findings indicate that online search behavior is evolving from a search engine-centric model toward a hybrid, platform-integrated, and user-centric search ecosystem.

RESEARCH CONTRIBUTIONS

This study makes several important contributions to research on online information behavior. First, it provides a theoretical contribution by refining the traditional online search process model to reflect contemporary hybrid search environments. The adapted model explicitly incorporates social media platforms, marketplaces, comparison portals, and AI-based search systems as integral components of the search process.

Second, the study contributes empirically by providing quantitative evidence on platform choice and search behavior among Generation Z and Generation Alpha in Germany. The inclusion of Generation Alpha represents a particularly relevant contribution, as this generation has been largely underrepresented in previous research.

Third, the study contributes to the understanding of AI-based search systems by demonstrating that these systems currently complement rather than replace traditional search engines. This provides important insights into the evolving role of generative AI in information seeking.

Finally, the study contributes methodologically by empirically validating the adapted search process model and demonstrating the relevance of influencing factors such as credibility, experience, habits, preferences, and search intention. Thus, the study contributes to the broader theoretical understanding of online information behavior.

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

https://doi.org/10.5281/zenodo.22916430

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

LIMITATIONS AND FUTURE RESEARCH

This study has several limitations that should be considered. First, the sample is not fully representative of the entire population of Generation Z and Generation Alpha in Germany. The relatively high level of education and the high proportion of students may influence search behavior and platform usage patterns.

Second, the study relies on self-reported data, which may differ from actual observed behavior. Future research could use behavioral data such as log file analysis or experimental methods.

Third, the study focuses on the German context, which may limit generalizability to other cultural environments.

Future research should investigate the role of AI-based search systems in greater depth, particularly as these technologies continue to evolve rapidly. Longitudinal studies could provide valuable insights into how search behavior changes over time.

CONCLUSIONS

This study demonstrates that online search behavior among Generation Z and Generation Alpha is undergoing a structural transformation toward a hybrid search ecosystem. While traditional search engines remain central, alternative platforms such as social media and AI-based systems play an increasingly important complementary role.

The adapted search process model and empirical findings highlight that platform choice has become a central component of contemporary search behavior. Search is no longer confined to a single system, but distributed across multiple platforms depending on search intention, trust, experience, and user preferences.

These findings contribute to a better understanding of contemporary information behavior and provide a foundation for future researchdemonstrates that search must be understood as a multi-platform, context-dependent, and dynamic process.

REFERENCES

[1] Adobe Express. (2024). Using TikTok as a Search Engine. Electronically published: https://www.adobe.com/express/learn/blog/using-tiktok-as-a-search-engine.

[2] Blocksidge, K./Primeau, H. (2023). Adapting and evolving: Generation Z's information beliefs. The Journal of Academic Librarianship, 49(3). DOI:10.1016/j.acalib.2023.102686.

[3] Criteo. (2018). Generation Z - der Report. Basierend auf der Criteo Shopper Story Electronically published: https://www.criteo.com/de/wp-content/uploads/sites/3/2018/06/GenZ_Report_DE.pdf.

[4] Claneo & Appinio. (2024). State of Search Studie 2024: Einblicke in das Suchverhalten der Deutschen. Electronically published: https://www.claneo.com/de/state-of-search-studie-2024/.

[5] Dean, B. (2024). Search Intent. Electronicall published: https://backlinko.com/hub/seo/search-intent.

[6] Haan, K. (2024). Is Social Media The New Google? Gen Z Turn To Google 25% Less Than Gen X When Searching. Electronically published: https://www.forbes.com/advisor/business/software/social-media-new-google/.

[7] Johnson, K. (2022). Google Borrows From TikTok to Keep Gen Z Searching Electronically published: https://www.wired.com/story/google-borrows-from-tiktok-to-keep-gen-z-searching/.

[8] Kemp, S. (2023). Digital 2023 Deep-Dive: The ongoing evolution of search behaviours. Electronically published: https://datareportal.com/reports/digital-2023-deep-dive-the-ongoing-evolution-of-search-behaviours.

[9] Lewanczik, N. (2022). Konkurrenz für die Google-Suche: Gen Z sucht vermehrt via TikTok und Instagram. Electrolically published: https://onlinemarketing.de/technologie/konkurrenz-google-suche-gen-z-tiktok-instagram.

[10] Lewanczik, N. (2024). Googeln verliert an Bedeutung. Electronically published: https://onlinemarketing.de/branding/googeln-verliert-an-bedeutung.

[11] Lewandowski, D. (2023). Understanding Search Engines. Cham: Springer International Publishing.

[12] Lewandowski, D./Womser-Hacker, C. (2023). D6 Information Seeking Behaviour. In: R. Kuhlen, D. Lewandowski, W. Semar und C. Womser-Hacker, Hg. Grundlagen der Informationswissenschaft. 7. Auflage: De Gruyter Saur, pp. 553-566.

[13] Lidal, L.N. (2023). TikTok – bare tidsfordriv eller nyttig scrolling? - En kvalitativ studie av unge voksnes bruk, motivasjon og holdninger til det sosiale mediet TikTok. Electronically published: https://www.duo.uio.no/handle/10852/104642.

[14] mdr. (2023). Generation Alpha. Electronically published: https://www.mdr.de/next/userverse-2035/profile/profile-generation-alpha-100.html.

[15] Metzger, M.J./Flanagin, A.J./Eyal, K./Lemus, D.R./ McCann, R.M. (2003). Credibility for the 21st Century: Integrating Perspectives on Source, Message, and Media Credibility in the Contemporary Media Environment. *Annals of the International Communication Association*, 27(1), S. 293-335. DOI: 10.1080/23808985.2003.11679029.

[16] Moran, S. (2023). Consumer trends dominating 2023. Electronically published: https://www.gwi.com/webinars/connecting-the-dots-2023.

[17] Muhammad, A.S./ Adeshola, I./ Isiaku, L. (2024). A mixed study on the "wow" of impulse purchase on Instagram: insights from Gen-Z in a collectivistic environment. Young Consumers, 25(1). DOI: 10.1108/YC-04-2023-1728.

[18] Newberry, C. (2024). Social media search engines: What they are & why they matter. Electronically publisher: https://blog.hootsuite.com/social-media-search/.

[19] Page, M.J. et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ (Clinical research ed.)*, 372. DOI: 10.1136/bmj.n71.

[20] Prabhakar, R. (2022). Organizing the world's information. In: FORTUNE, Hg. *Brainstorm Tech 2022*.

[21] Riedel, A./Michelis, D. (2023). Informations- und Entscheidungsprozesse. In: Riedel, A./Michelis, D., Hg. Die vernetzten Konsument*innen. Grundlagen des Marketing im Zeitalter partizipativer Unternehmensfuhrung. 2. Auflage. Berlin und Heidelberg: Springer, S. 85-120.

[22] Riemer, K./Brüggemann, F. (2007). Personalisierung der Internetsuche. *Wirtschaftsinformatik*, 49(2). DOI: 10.1007/s11576-007-0033-6.

[23] Rogelberg, S. (2024). Gen Z moves on from 'Googling'—TikTok emerges as the new search engine. Electrolically published: https://fortune.com/2024/09/10/gen-z-google-verb-social-media-instagram-tiktok-search-engine/.

[24] Wang, M./Lind, S. (2023). Exploring the Search Behavior of Teenagers: A Comparative Study of Social Media and Browser Usage. *Journal of Student Research*, 12(4). DOI: 10.47611/jsrhs.v12i4.5172.

[25] Wilson, T.D. (1999). Models in information behaviour research. Journal of Documentation, Vol. 55, No. 3, June 1999, pp. 249-270.

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22916430

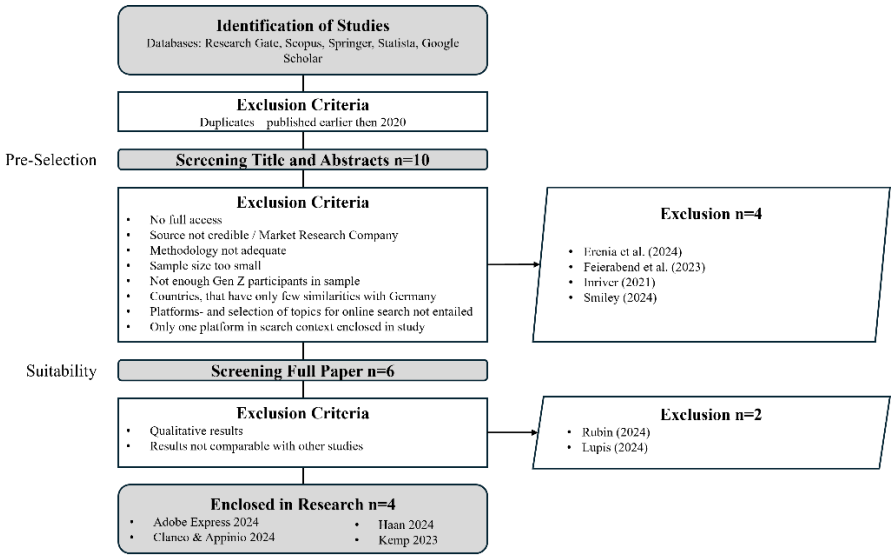
APPENDIX A

Flow chart of the PRSMA analysis as the theoretical basis for the study, based on Page et al., 2021

The databases mentioned are searched using the following keywords to select the studies: Study Search Engine, Study Search Behavior, Study Search Habits, Study Gen Z Search Behavior, Study Search Trends, Study Social Media as Search Engine, Study Social Search.

To be included in the preliminary selection, studies must be empirical studies from 2020 to 2024 with a clear reference to online information search.

Final PRISMA-Flow-Chart



Preselected studies

The following 10 studies were identified. Finally selected studies are marked in light grey.

Source	Survey Period	Country	Sample	Age	Title and Source
Adobe Express (2024)	2023/2024	USA	808	/	Using TikTok as a Search Engine. Electronically published: https://www.adobe.com/express/learn/blog/using-tiktok-as-a-search-engine.
Claneo & Appinio (2024)	April 2024	Germany	1001	16-65	State of Search Studie 2024: Einblicke in das Suchverhalten der Deutschen. Electronically published: https://www.claneo.com/de/state-of-search-studie-2024/.
Haan (2024)	April2024	USA	2000	/	Is Social Media The New Google? Gen Z Turn To Google 25% Less Than Gen X When Searching. Electronically published: https://www.forbes.com/advisor/business/software/social-media-new-google/.
Kemp (2023)	2023	worldwide	/	16-64	Digital 2023 Deep-Dive: The ongoing evolution of search behaviours. Electronically published: https://datareportal.com/reports/digital-2023-deep-dive-the-ongoing-evolution-of-search-behaviours.
Erenia et a. (2024)	2024	/	/	/	Search behavior of Gen Z on YouTube and TikTok. International Journal of Science and Research Archive, 12(1), S. 1023-1032. DOI:10.30574/ijsra.2024.12.1.0888.
Feierabend et al. (2023)	May-Juni 2023	Germany	200	12-19	IM-Studie 2023 - Jugend, Information, Medien. Medienpädagogischer Forschungsverbund Südwest. Electronically published: https://www.mpfs.de/fileadmin/files/Studien/JIM/2022/JIM_2023_web_final_kor.pdf.
Inriver (2021)	2021	Germany	2000	/	Inside the mind of an online shopper. Electronically published: https://www.inriver.com/resources/inside-the-mind-of-an-online-shopper-infographic-us-statistics/.
Lupis (2024)	August 2023	12 countries	15.469	16-64	Search Remains Tops, but Brand Discovery Channels Are Evolving. Marketing Charts. Electronically published: https://www.marketingcharts.com/customer-centric/analytics-automated-and-martech-232699.
Rubin (2024)	2024	/	/	/	Google who? Gen Z is searching on TikTok, YouTube instead . AXIOS Electronically published: https://www.axios.com/2024/04/11/google-gen-z-search-engines-tiktok-youtube.
Smiley (2024)	July 2023	USA	1005	/	Gen Z Social Search Habits - Redefining Search Across YouTube, TikTok, and Instagram. Electronically published: https://www.emarketer.com/content/gen-zsocial-search-habits.

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22916430

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

APPENDIX B

Detailed tables on the results of the empirical study

B1: Influencing factor: habits

H1: *The more frequently the platform is used in general, the more often it is also used in a search context.*

Null hypothesis: *On platform X, there is no correlation between general use and use in a search context.*

Alternative hypothesis: *On platform X, there is a correlation between general use and use in a search context.*

Platform	SE	IG	TT	FB	YT	AI	MP	CP
Correlation coefficient	0.566	0.444	0.799	0.662	0.440	0,746	0.761	0.414
Probability of error	<0,001	<0,001	<0,001	<0,001	<0,001	<0,001	<0,001	<0,001
H0	rejected	rejected	rejected	rejected	rejected	rejected	rejected	rejected
Significance of relation	positive	positive	positive	positive	positive	positive	positive	positive

Please note: SE=Search Engines; IG=Instagram; TT=TikTok; FB=Facebook; YT=YouTube; AI=Artificial Intelligence; CP=Comparison Portals. The abbrevaiations are valid for the following tables as well.

B2: Influencing factor: experiences

H2: *The better the experience with the platform, the more often it is used in a search context*.

Null hypothesis: *There is no correlation between experience with platform X and use of the platform in a search context*.

Alternative hypothesis: *There is a correlation between experience with platform X and use of the platform in a search context*.

Platform	SE	IG	TT	FB	YT	AI	MP	CP
Correlation coefficient	0.180	0.617	0.871	0.750	0.486	0.626	0.465	0.732
Probability of error	<0,001	<0,001	<0,001	<0,001	<0,001	<0,001	<0,001	<0,001
H0	rejected	rejected	rejected	rejected	rejected	rejected	rejected	rejected
Significance of relation	positive	positive	positive	positive	positive	positive	positive	positive

B3: Influencing factor: credibility of the medium

H3: *The higher the perceived credibility of the platform, the more often it is used in a search context.*

Null hypothesis: *There is no correlation between the perceived credibility of platform X and the use of the platform in a search context*.

Alternative hypothesis: *There is a correlation between the perceived credibility of platform X and the use of the platform in a search context.*

Platform	SE	IG	TT	FB	YT	AI	MP	CP
Correlation coefficient	0.116	0.422	0.515	0.267	0.284	0.295	0.219	0.414
Probability of error	<0,001	<0,001	<0,001	<0,001	<0,001	<0,001	<0,001	<0,001
H0	rejected	rejected	rejected	rejected	rejected	rejected	rejected	rejected
Significance of relation	positive	positive	positive	positive	positive	positive	positive	positive

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22916430

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

B4: Influencing factor: Topics

Selection of platform depending on topic of search request.

Please note: since there was a topic "others" the percentages do not add up to 100%.

Selection of platform per topic in %	Products & Brands			Services			Tourism			Health & Fitness		
	All	All	Gen Z	Gen A	Gen Z	Gen A	Gen A	Gen Z	Gen A	Gen A	Gen Z	Gen A
Search Engines	50,8	48,2	59,8	64,7	64,0	48,4	52,5	54,4	44,1	51,8	51,8	n.a.
Instagram	6,0	5,2	10,4	6,5	6,5	16,1	10,8	9,6	22,7	7,2	7,2	n.a.
TikTok	6,8	6,1	9,5	22,8	22,8	29,0	10,8	9,7	20,2	8,2	8,2	n.a.
Facebook	0,0	0,0	0,0	0,0	0,0	0,0	0,3	0,4	0,0	0,0	0,0	n.a.
YouTube	1,6	1,5	1,9	0,0	0,0	0,0	2,3	2,1	4,3	16,4	16,4	n.a.
KI-Search Engines	1,0	1,1	1,1	1,1	1,1	0,0	0,2	0,2	0,0	4,0	4,0	n.a.
KI-Chatbots	2,7	3,0	1,0	1,1	1,1	6,5	3,9	3,4	8,8	11,6	11,6	n.a.
Marketplaces	17,4	18,8	10,6	0,0	0,0	0,0	0,3	0,4	0,0	0,0	0,0	n.a.
Comparison portals	7,4	9,1	0,0	0,5	0,5	0,0	17,0	17,5	0,0	0,3	0,3	n.a.

Selection of platform per topic in %	World affairs & Pop-Cult.			Tutorial & Tipps			Learning content		
	All	All	Gen Z	Gen A	Gen Z	Gen A	Gen A	Gen Z	Gen A
Search Engines	57,0	57,6	50,7	44,3	47,3	29,5	40,6	39,2	47,2
Instagram	11,8	11,5	11,3	9,0	9,7	5,9	0,5	0,6	0,0
TikTok	12,5	12,1	19,4	14,7	14,4	16,5	1,7	1,7	1,6
Facebook	0,8	0,9	0,0	0,8	1,0	0,0	0,3	0,3	0,0
YouTube	7,5	6,8	11,3	17,7	13,6	37,6	23,0	22,8	23,8
KI-Search Engines	2,7	2,5	5,9	1,5	1,4	1,4	4,4	4,3	4,7
KI-Chatbots	4,5	5,1	1,4	7,6	7,9	6,3	26,8	27,6	22,7
Marketplaces	0,4	0,4	0,0	0,2	0,3	0,0	0,7	0,9	0,0
Comparison portals	0,2	0,2	0,0	0,2	0,3	0,0	0,0	0,0	0,0

B5: Influencing factor: preferences

Please note: since there was a topic "others" the percentages do not add up to 100%.

Preferences per generation (in %)	All	Gen Z	Gen A
Simple formulation of the search query	19,2	20,5	10,7
Images and voice search	1,9	1,9	2,0
Interaction and exchange of views	3,9	3,8	4,9
Possibility to save search results	5,1	5,2	4,8
Personalised search results	3,8	3,6	5,6
Visual results	12,3	11,8	16,2
Creative presentation, e.g. with storytelling	3,8	3,6	4,8
Clear presentation of results	12,4	12,9	9,3
Info from people with real experiences (UCG)	8,4	8,9	4,4
Info from experts	7,4	6,8	11,3
Current information and trends	6,9	5,9	14,2
Info from different perspectives	4,9	4,9	4,7
Feeling that one has all the relevant information	10,0	10,4	7,7

Preferences depending on search intent (in %)	Info	Navi	Trans
Simple formulation of the search query	16,3	25,3	17,2
Images and voice search	1,7	2,4	1,7
Interaction and exchange of views	5,1	2,0	3,6
Possibility to save search results	5,5	4,0	6,4
Personalised search results	5,1	1,7	3,2
Visual results	15,0	9,6	7,0
Creative presentation, e.g. with storytelling	5,4	2,1	0,6
Clear presentation of results	9,9	16,8	12,7
Info from people with real experiences (UCG)	9,3	4,0	15,3
Info from experts	7,0	8,2	7,0
Current information and trends	6,0	8,7	6,7
Info from different perspectives	5,8	2,5	6,6
Feeling that one has all the relevant information	8,0	12,8	12,1

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22904651

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

DESIGNING A POWER BI DASHBOARD FOR DATA-DRIVEN SUPPLIER COMPARISON IN MANUFACTURING PROCUREMENT

Beulah Devathasan[1], Emolin Indran Rajanbabu[2], Nidhi[3]
[3]Department of Business Development and Technology, Aarhus University, Herning, Denmark
Emails: [1]Beulahdevathasan@outlook.dk, [2]emolin-1@hotmail.com, [3]nidhi@btech.au.dk

Abstract

Procurement departments often rely on manually maintained spreadsheets when comparing supplier quotations. Such approaches often result in inefficient workflows, inconsistent data structures, and limited transparency in decision-making. This paper investigates how standardised supplier data templates combined with Business Intelligence dashboards can support data-driven procurement decision-making. Using two manufacturing companies as case studies, a standardised supplier quotation template was designed to ensure consistent data collection across suppliers. The structured data were then integrated into a Power BI dashboard that enables interactive supplier comparison based on key procurement metrics such as landed cost, freight cost, and forecasted order volumes.

The study demonstrates how improved data standardisation enables automated analytics and visualisation of supplier information. The results show that the proposed approach enhances procurement transparency, reduces manual analysis effort, and supports more informed supplier selection decisions. The paper contributes by demonstrating how structured data templates and Business Intelligence tools can jointly enable procurement analytics in manufacturing organisations.

Keywords: Business Intelligence, Dashboard, Procurement Analytics, Power BI, Supplier Evaluation, Data Standardisation, Decision Support.

INTRODUCTION

Digitalisation has accelerated the emergence of data-driven organisations, where operational and strategic decisions increasingly rely on data analytics and information systems. As organisations collect larger volumes of operational data, the ability to transform these data into actionable insights has become critical for maintaining competitiveness and improving decision-making processes [1][2]. Consequently, a data-driven organisational culture has become essential for sustained business success [3].

Procurement functions are particularly affected by this shift as it demands to process large volumes of supplier data, cost structures, and logistical information. In order to select suppliers, these information must be evaluated [3] [4]. However, traditional procurement practices often rely on manual spreadsheet-based analysis that lacks automation and structured analytics [4]. These approaches can lead to time-consuming workflows, inconsistencies in data formats, and limited transparency in supplier comparisons. As procurement data becomes more complex, businesses need advanced analytics and visualisation tools to support decision-making to facilitate real-time insights [5].

Business Intelligence (BI) tools such as Microsoft Power BI (Power BI) provides an effective solution for organisations seeking to improve analytical practices and data-driven processes, offering modern functionality with greater adaptability, user-friendliness, accuracy and speed [6]. It enables organisations to integrate data, perform analytics and draw data-driven insights into business decisions.

Despite these capabilities, many procurement departments still struggle to effectively integrate BI tools into their workflows because supplier quotation data are often unstructured or inconsistent. Without standardised data structures, it becomes difficult to automate analytics and create reliable dashboards [7].

In this paper, Power BI is introduced as a tool for supplier comparison within the procurement departments of two manufacturing companies referred to as Company A and Company B. In both organisations, the existing supplier comparison processes are manual, time-consuming, and lacks a structured analytical framework, underscoring the need for more efficient, transparent, and data-driven decision support.

The paper therefore addresses the following research question:

How can standardised supplier data templates and Business Intelligence dashboards support data-driven supplier comparison in manufacturing procurement?

This paper contributes to the literature on procurement analytics and BI in three ways. *First*, it demonstrates how standardised supplier quotation templates can enable automated procurement analytics by providing structured and comparable data. *Second*, it develops and evaluates a Power BI dashboard artifact that integrates multiple procurement cost components to support real-time supplier comparison. *Third*, it provides empirical insights from two manufacturing case organisations, showing how BI tools can improve procurement efficiency and decision transparency without requiring major changes to existing workflows. Together, these contributions illustrate how

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22904651

data standardization and Business Intelligence dashboards can support practical implementation of procurement analytics in manufacturing environments.

The remainder of the paper is structured as follows.

Section 2 reviews related work on Business Intelligence and procurement analytics. Section 3 presents the research methodology and data collection process. Section 4 describes the current procurement processes and the development of the standardized supplier data template. Section 5 explains the design of the Power BI dashboard and its data model. Section 6 demonstrates the prototype and its analytical capabilities. Section 7 discusses the findings and implications, and Section 8 concludes the paper.

RELATED WORK

Supplier selection is a central task in procurement management [8][9]. Organisations typically assess suppliers based on criteria such as price, logistics costs, delivery performance, and quality reliability. Accurate supplier comparison requires structured datasets that enable consistent analysis across multiple suppliers.

Several studies have highlighted the potential of BI tools to enhance procurement performance through improved supplier evaluation, cost analysis and inventory management. For example, dashboards can support monitoring of supplier performance metrics such as delivery reliability, quality and cost efficiency, while spend analysis enables organisations to identify cost-saving opportunities and optimise purchasing strategies.

Illustrative scenarios developed by Microsoft demonstrate how Power BI dashboards can support procurement tasks, such as vendor comparisons, cost monitoring, and budget planning, through interactive reports and real-time data visualisation. Similarly, other empirical research demonstrates that Power BI dashboards can transform procurement spending data into actionable insights for evaluating supplier performance and managing costs. The findings highlight the importance of data preparation and integration with enterprise systems to ensure reliable analytics [10][11][12].

In procurement contexts, BI systems can support activities such as supplier evaluation, spend analysis, inventory monitoring, and cost management. Dashboards enable procurement managers to visualise supplier performance metrics such as delivery reliability, cost efficiency, and procurement volumes.

Despite these contributions, research specifically examining how Power BI can support automated supplier comparison and real-time procurement decision-making remains limited. This research addresses this gap by developing and evaluating a Power BI-based prototype designed to enhance procurement analysis and supplier comparison processes.

METHODOLOGY

This paper adopts a case study approach involving collaboration with two manufacturing companies, referred to as Company A and Company B, whose procurement departments operate in internationally distributed organisational structures. In both organisations, supplier comparison processes are currently performed manually, creating challenges related to efficiency, transparency, and structured analysis.

The research involved the following stages:

- Analysis of existing procurement processes and supplier comparison practices
- Design of a standardized supplier quotation data template
- Development of a Power BI dashboard prototype
- Evaluation of the prototype using supplier quotation data

The paper utilises procurement data obtained from the participating companies, focusing on information related to supplier quotations and purchasing activities. To enable systematic analysis and ensure comparability across suppliers, a standardised supplier template was developed. The purpose of this template was to structure supplier quotation data in a consistent format, allowing the information to be integrated and analysed within the business intelligence environment. See *Table 1*.

Table 1: Excerpt of Data Template

Supplier name	Item number	Item description	Year	Quantity	Delivery Place (City and Country)	MOQ (kg)	Raw material cost
Supplier 1	A19-1469	100 X 60 X 2MM L=170.08"4320MM	1	500	Bogense, Denmark	500	1500
Supplier 1	A19-1469	100 X 60 X 2MM L=170.08"4320MM	2	6000	Bogense, Denmark	500	1500
Supplier 1	A19-1469	100 X 60 X 2MM L=170.08"4320MM	3	14000	Bogense, Denmark	500	1500
Supplier 1	A19-1469	100 X 60 X 2MM L=170.08"4320MM	4	21000	Bogense, Denmark	500	1500
Supplier 1	A19-1469	100 X 60 X 2MM L=170.08"4320MM	5	29000	Bogense, Denmark	500	1500
Supplier 1	A19-1470	80 X 40 X 2MM L:160.24"4070MM	1	700	Bogense, Denmark	500	4000
Supplier 1	A19-1470	80 X 40 X 2MM L:160.24"4070MM	2	6000	Bogense, Denmark	500	4000
Supplier 1	A19-1470	80 X 40 X 2MM L:160.24"4070MM	3	10000	Bogense, Denmark	500	4000
Supplier 1	A19-1470	80 X 40 X 2MM L:160.24"4070MM	4	23000	Bogense, Denmark	500	4000
Supplier 1	A19-1470	80 X 40 X 2MM L:160.24"4070MM	5	30000	Bogense, Denmark	500	4000
Supplier 1	A19-1480	90 X 50 X 2MM L:157.48"4000MM	1	900	Bogense, Denmark	500	3000
Supplier 1	A19-1480	90 X 50 X 2MM L:157.48"4000MM	2	7000	Bogense, Denmark	500	3000
Supplier 1	A19-1480	90 X 50 X 2MM L:157.48"4000MM	3	11000	Bogense, Denmark	500	3000
Supplier 1	A19-1480	90 X 50 X 2MM L:157.48"4000MM	4	19000	Bogense, Denmark	500	3000
Supplier 1	A19-1480	90 X 50 X 2MM L:157.48"4000MM	5	31000	Bogense, Denmark	500	3000

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22904651

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

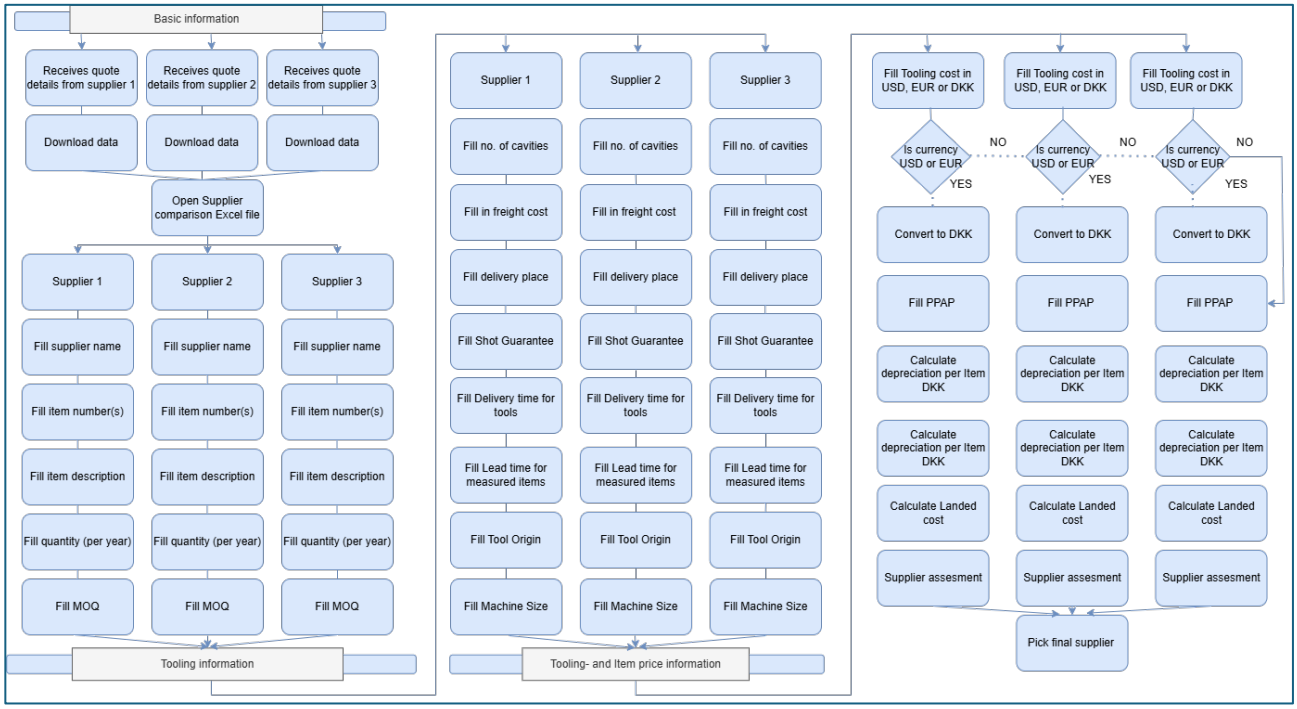


Figure 1: Current Supplier Comparison Process

The standardised template served as the foundation for organising procurement data used in the development of the Power BI prototype, thereby supporting more structured supplier comparison and analytical processing within the dashboard.

CURRENT PROCUREMENT PROCESS AND DATA TEMPLATE

Figure 1 illustrates the workflow for consolidating and comparing supplier quotations for a procurement process at the case companies. It visualises how data from multiple suppliers is collected, processed and structured into a unified comparison framework. In this example, quotation data from three suppliers is used to demonstrate the process.

However, the number of suppliers involved may vary depending on the complexity of the item and the technical or manufacturing capabilities of the suppliers. The complete process typically requires approximately 30 minutes to complete for a single supplier comparison.

The current procurement process at the case companies involves manually comparing quotes from suppliers to identify the most suitable option. The process requires extensive data entry, currency conversion, and cost calculations, followed by supplier evaluation. The reliance on manual handling makes the process time-consuming, error-prone and inefficient, highlighting potential for optimisation.

To address these limitations, a standardized supplier quotation template was developed. The template structures supplier quotation data into predefined fields that can be directly integrated into the BI environment. The template therefore serves as the foundation for organising procurement data and enabling automated supplier comparison.

DASHBOARD DESIGN

The Power BI dashboard was developed to support a more structured and transparent supplier comparison process by visualising procurement data collected through the standardised supplier template. The dashboard design followed an iterative development process in which an initial prototype was created and subsequently refined to improve data structure, usability, and visual representation

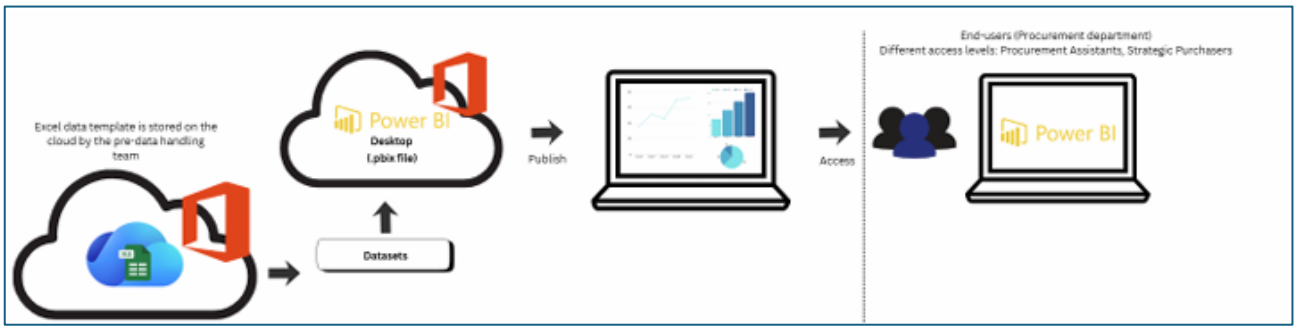


Figure 2: Architecture

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22904651

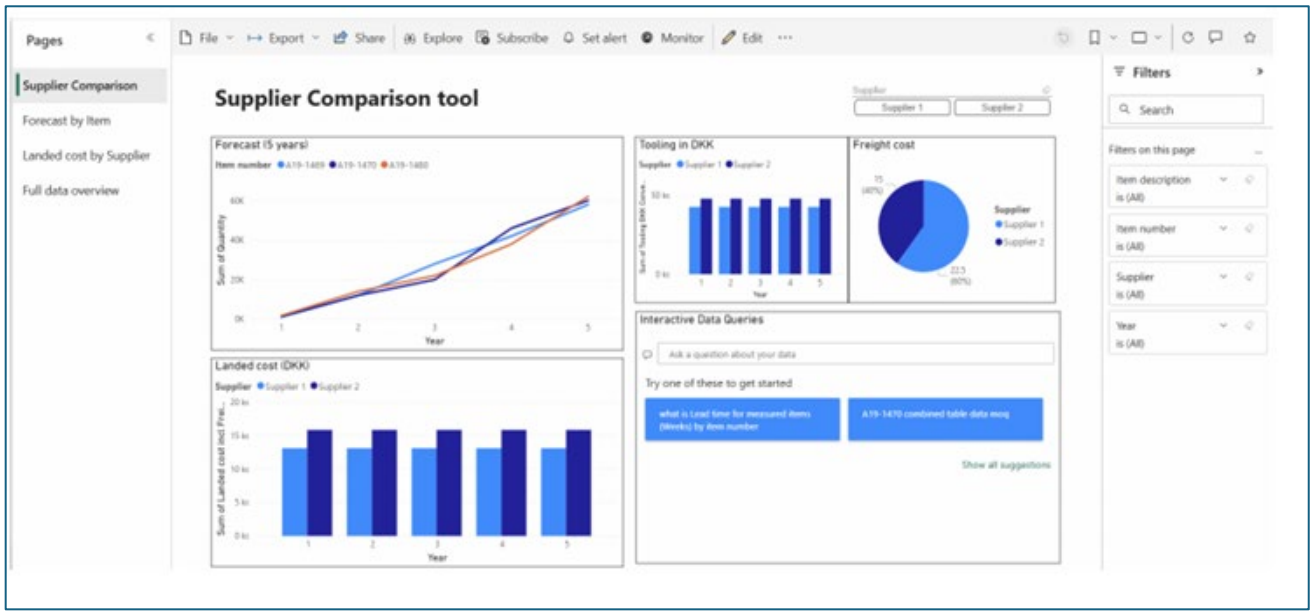


Figure 3: Dashboard Prototype

of procurement information. *Figure 2* illustrates the basic architecture.

Data Model

The dashboard relies on supplier quotation data collected through the data template provided to suppliers during the request-for-quotation (RFQ) process. Once completed by the suppliers, the templates are uploaded to a cloud-based storage environment on OneDrive, where they serve as the primary data source for the Power BI report. The data files are subsequently imported into Power BI Desktop, where they are transformed and organised within a semantic data model that structures the datasets and defines the calculated measures used in the analysis.

Visualisations

The dashboard visualises procurement data through several analytical components designed to present key information relevant to supplier evaluation. The visual interface integrates supplier and item information alongside procurement metrics such as forecasted demand volumes, tooling costs, and item prices. In addition, the dashboard presents freight costs and other cost elements that influence the total sourcing cost of a component. The dashboard also incorporates an interactive data query feature that allows users to ask questions using natural language. This functionality enables procurement professionals to interact directly with the dataset and retrieve relevant insights without the need for advanced analytical skills. Overall, the dashboard functions enable data-driven procurement by transforming raw supplier quotations into structured, comparable, and visually interpretable information. It gives decision-makers immediate access to key metrics, supports cost-efficient sourcing, and allows strategic evaluation based on standardised data rather than ad-hoc calculations. By automating this process, the supplier comparison time is reduced from 30 minutes to 6 minutes and 37 seconds, see *Table 2.*

PROTOTYPE DEMONSTRATION

To illustrate the analytical capabilities of the developed artefact, the Power BI prototype was used to analyse supplier quotation data collected through the data template. The prototype is visualised in *Figure* 3.

Forecast

The forecast visualisation presents projected annual procurement quantities for selected items over a five-year period. This view provides context for supplier comparison, as expected purchasing volumes can influence price negotiations. In the demonstrated case, forecasted quantities are identical for both suppliers, ensuring that differences in cost calculations arise from supplier-specific factors rather than variations in demand.

Table 2: Estimated Time Reduction

Process	Time (sec)
Open Outlook	5
Find Supplier 1 & Download data	60
Find Supplier 2 & Download data	60
Find Supplier 3 & Download data	60
Open overview of duty % & fill percentage into supplier documents	180
Open the OneDrive Cloud Storage	5
Create a folder	7
Upload files to the folder	20
Compare	
Total	6min 37sek

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22904651

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

Freight Cost

The freight cost visualisation shows transportation costs for sourcing items from different suppliers. Freight is a significant factor in procurement decisions, as logistical costs can significantly influence the total landed cost, particularly when suppliers are located in different countries. Therefore, freight is displayed as a separate visual element, enabling procurement employees to assess the logistical implications of each sourcing option.

Tooling Cost

The tooling cost visualisation shows the tooling costs for manufacturing the selected items for each supplier over a five-year period. Tooling costs are an important consideration, as the condition and quality of tooling influence the quality of the manufactured component. Additionally, tooling represents an investment that affects the overall cost structure and therefore plays a role in supplier evaluation.

Landed Cost

The landed cost view demonstrates how the prototype combines multiple cost elements to provide a comprehensive overview of the total sourcing cost associated with each supplier quotation. By combining item price, freight costs, tariffs and depreciation costs, the dashboard enables procurement professionals to assess suppliers based on the overall landed cost of a component. This integrated view supports a more comprehensive cost comparison and helps identify suppliers that offer the most cost-efficient sourcing option.

To enable Power BI to calculate the landed cost automatically, a DAX formula was created to combine the relevant cost components into a single measure. The DAX formula used was as follows:

```
Landed cost incl Freight Tariff & Depreciation =
(
    'Combined Table data'[Item DKK Converted]
    * 'Combined Table data'[Tariff (%)]
)
+ 'Combined Table data'[Freight cost]
+ 'Combined Table data'[Depreciation per item (DKK)]
```

The landed cost measure was calculated to include tariffs, freight costs, and depreciation in addition to the converted item price. By automating supplier comparison, the dashboard reduces supplier comparison time from approximately 30 minutes to 6 minutes and 37 seconds, representing an efficiency improvement of approximately 78%.

ANALYSIS AND DISCUSSION

The findings indicate that the solution consolidates data and standardises supplier evaluations, enabling partial automation and real-time comparison of suppliers without disrupting existing workflows.

The dashboard significantly reduces the time required to perform supplier comparisons, demonstrating the operational benefits of integrating BI tools into procurement processes. The supplier comparison time reduced from 30 minutes to 6 minutes and 37 seconds, resulting in an efficiency gain of nearly 78%. Rather than replacing current practices, it supports strategic planning and decision-making by providing structured, visual access to relevant procurement information.

At the organisational level, the dashboard enhances both operational and managerial activities: procurement staff gain improved access to comparable supplier data, while managers obtain a consolidated overview that supports oversight and coordination.

From a theoretical perspective, the findings highlight the critical role of data standardization as an enabler of procurement analytics. While Business Intelligence tools provide advanced analytical capabilities, their effectiveness depends heavily on the availability of structured and comparable datasets. The study therefore suggests that procurement analytics initiatives should not only focus on analytical technologies but also on the design of standardized data structures that allow automated analysis.

For practitioners, the results demonstrate how relatively simple technological solutions can significantly improve procurement processes. By implementing standardized supplier templates and BI dashboards, organizations can reduce manual analytical effort while improving transparency in supplier comparison. The solution can be implemented without replacing existing procurement systems, making it a practical approach for organizations seeking to gradually introduce data-driven decision support.

However, the effectiveness of the solution depends on data quality, standardisation, and internal analytical capabilities, highlighting that technological functionality alone is insufficient for successful implementation.

CONCLUSION

This paper presented a data-driven approach to supplier comparison in manufacturing procurement through the development of a standardized supplier data template and a Power BI dashboard. In this paper, data driven insights were derived from manual procurement processes through the development of an automated tool for cost comparison and landed cost calculation. The dashboard consolidates and standardises supplier quotations, automates key cost calculations, and provides interactive visualisations, creating an efficiency gain of approximately 78%.

A key challenge identified during the development process was the heterogeneity of supplier quotation formats and the complexity of integrating multiple cost components such as item price, freight, tariffs, and tooling costs. These challenges were addressed through the design of a standardized supplier data template and the

implementation of a semantic data model within Power BI, enabling consistent data processing and automated calculations.

The results indicate that the proposed solution improves both operational efficiency and decision transparency within procurement functions. Procurement staff gains improved access to comparable supplier data, while managers benefit from a consolidated overview that supports strategic sourcing decisions and cost analysis.

Future research could extend this work by integrating the solution with enterprise procurement systems such as ERP platforms, incorporating predictive analytics for supplier performance evaluation, and further automating data collection processes. Such developments could further strengthen the role of Business Intelligence in enabling data-driven procurement and strategic supply chain management.

REFERENCES

[1] T. Hautala-Kankaanpää, "Complementary and contingent value of SMEs' data capability and supply chain capability in the competitive environment", Ind. Manag. Data Syst., bd. 123, nr. 8, s. 2128–2149, 2023, doi: 10.1108/IMDS-01-2023-0013.

[2] M. Guida, F. Caniato, A. Moretto, og S. Ronchi, "The role of artificial intelligence in the procurement process: State of the art and research agenda", J. Purch. Supply Manag., bd. 29, nr. 2, s. 100823, mar. 2023, doi: 10.1016/j.pursup.2023.100823.

[3] G. Agag m.fl., "Understanding the relationship between marketing analytics, customer agility, and customer satisfaction: A longitudinal perspective.", J. Retail. Consum. Serv., bd. 77, 2024, doi: 10.1016/j.jretconser.2023.103663.

[4] M. M. Wittinger og K. Demeter, "Guidance on how to balance the purchasing environment and processes to save resources - a validity examination of a holistic model", Cogent Bus. Manag., bd. 11, nr. 1, 2024, doi: 10.1080/23311975.2024.2374883.

[5] V. Delke, H. Schiele, W. Buchholz, og S. Kelly, "Implementing Industry 4.0 technologies: Future roles in purchasing and supply management", *Technol. Forecast. Soc. Change*, bd. 196, s. 122847, nov. 2023, doi: 10.1016/j.techfore.2023.122847.

[6] Krishna Kishor Tirupati, Archit Joshi, Dr S P Singh, Akshun Chhapola, Shalu Jain, og Dr. Alok Gupta, "Leveraging Power BI for Enhanced Data Visualization and Business Intelligence", Univers. Res. Rep., bd. 10, nr. 2, s. 676–711, sep. 2024, doi: 10.36676/urr.v10.i2.1375.

[7] S. B. Mufor, "The Role of Analytics Solutions in Revolu-tionizing Procurement Efficiency", Am. J. Supply Chain Manag., bd. 9, s. 51–63, maj 2024, doi: 10.47672/ajscm.2024.

[8] A. Wetzstein, E. Hartmann, W. C. Benton jr., og N.-O. Hohenstein, "A systematic assessment of supplier selection literature – State-of-the-art and future scope", Int. J. Prod. Econ., bd. 182, s. 304–323, dec. 2016, doi: 10.1016/j.ijpe.2016.06.022.

[9] A. M. Khedr og S. R. S, "Enhancing supply chain management with deep learning and machine learning techniques: A review", J. Open Innov. Technol. Mark. Complex., bd. 10, nr. 4, s. 100379, dec. 2024, doi: 10.1016/j.joitmc.2024.100379.

[10] J. Fattulal og A. Shende, "Unlocking the power of business intelligence in procurement with Power BI", Int. J. Res. Manag., bd. 7, s. 684–692, apr. 2025, doi: 10.33545/26648792.2025.v7.i1h.336.

[11] K. Cantor, "Power BI For Procurement P3 Adaptive", P3 Adaptive. Set: 13. marts 2026. [Online]. Tilgængelig hos: https://p3adaptive.com/power-bi-for-procurement/

[12] Microsoft, "Power BI Purchase app - Business Central". Set: 13. marts 2026. [Online]. Tilgængelig hos: https://learn.microsoft.com/en-us/dynamics365/business-central/purchases-powerbi-app

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22924759

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

TRACEABLEAI: AN OPEN-SOURCE AGENTIC FRAMEWORK FOR PRISMA-COMPLIANT LITERATURE SYNTHESIS

Philip S. P. Ø. O. Lau[1], Nidhi[2]
Department of Business Development and Technology (BTECH), Aarhus University - Herning, Denmark

1. Introduction

The "reproducibility crisis" in Systematic Literature Reviews (SLRs) is a persistent issue, driven largely by the manual and undocumented nature of the screening and extraction phases. While Large Language Models (LLMs) and Automatic Survey Generation (ASG) systems (e.g., AutoSurvey, SurveyForge) offer great promise in accelerating discovery [1], their adoption in formal PRISMA-compliant reviews is limited by a lack of auditability. Recent studies on Deep Research Agents (DRAs) reveal that these systems often suffer from the "Mirage of Synthesis," where strong surface-level fluency and citation alignment obscure underlying factual and reasoning defects [2]. Furthermore, DRAs exhibit high variety, meaning repeated executions of the exact same query can yield vastly different research outcomes, findings, and citations [3].

To address this, the paper introduces **TraceableAI**, an open-source agentic framework designed to support PRISMA 2020-compliant [4] systematic literature reviews while maintaining transparency and reproducibility. The framework introduces a **Traceability Protocol** implemented through a machine-readable **Research Ledger** that records search queries, screening decisions, prompts, and evidence grounding for each reviewed paper. This enables AI-assisted literature synthesis to remain auditable and reproducible.

2. System Architecture & The Traceability Protocol

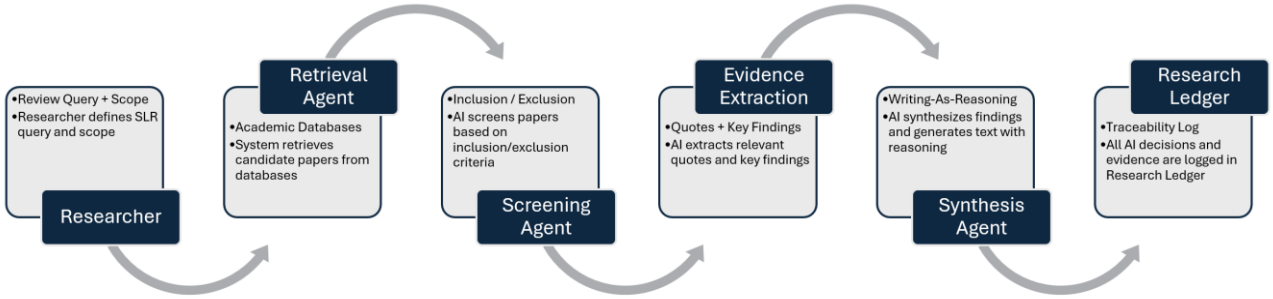


Figure 1 illustrates the architecture of TraceableAI and the Traceability Protocol's role in recording AI-assisted review decisions.

TraceableAI uses Mistral-Large as the reasoning engine, orchestrated by the ToolUniverse framework. The system processes academic sources through a standardized interface that retrieves candidate papers, evaluates inclusion criteria, and extracts relevant evidence. To ensure transparency, the system follows a Writing-As-Reasoning approach [5, 6], where evidence collection and text generation are interleaved, allowing the reasoning steps behind each decision.

The Traceability Protocol is a structured logging mechanism that records each AI-assisted decision during the SLR process, including the search query used, the prompts provided to the model, the screening decision, and the evidence supporting that decision into a Research Ledger (a JSON file). Inspired by frameworks that explicitly track an evidence-based & source verification state [7], the ledger logs:

- **Source Metadata:** The database the paper came from, the search query used, and the time it was found.

1 philip.lau@btech.au.dk
2 nidhi@btech.au.dk

https://doi.org/10.5281/zenodo.22924759

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

- **Prompt History:** The exact prompts given to the AI during the decision process.
- **Decision Logic & Grounding:** A {Decision, Criterion, Evidence} record that links each choice to a direct quote from the text. This ensures claims are backed by actual facts rather than just a basic URL link, preventing the citation-alignment-fallacy [2].
- **Parameter State:** Specific AI settings (like temperature and seed values) to help monitor and account for randomness in the model's outputs.

3. Experimental Design

To validate the framework, we propose a comparative experiment on the topic of "AI-based forecasting in the Supply Chain" with the given query: “AI-based forecasting” AND “Supply Chain”.

- **Dataset:** A candidate pool of N ≈ 200 papers retrieved from specified databases.
- **Comparative Analysis:** Two research assistants will perform a manual screening of the dataset. Their results (inclusion/exclusion and time taken) will be compared against the TraceableAI-assisted workflow.
- **Evaluation Metrics:**
    1. **Agreement Rate (Cohen’s Kappa):** To measure alignment between human and agent decisions [8].
    2. **Temporal Efficiency:** Total hours saved during the screening and extraction phases.
    3. **Citation Integrity & Evidence Grounding:** Evaluating the harmony of "Claim Attribution" (the ratio of verifiable claims cited) and "Citation Faithfulness" (whether the source genuinely supports the claim) [2, 9].
    4. **Reproducibility Completeness:** The percentage of decisions in the Research Ledger that can be independently verified by a human reviewer using the cited evidence, measuring the agent's resilience against hallucinated or orphaned citations.

4. Expected Findings

Preliminary evaluation suggests that the TraceableAI framework can significantly reduce manual screening effort while maintaining transparency and auditability required for PRISMA-compliant reviews. Current literature highlights a frequent disconnect between semantic retrieval capabilities and disciplinary narrative synthesis in ASG systems [2, 10]; we anticipate the "Research Ledger" will bridge this gap by revealing specific edge cases where human subjectivity (e.g., thematic coding) remains essential. Rather than replacing the researcher, TraceableAI serves as a verifiable co-pilot within a human-in-the-loop paradigm [11, 12], transforming the SLR process into a reproducible, data-driven, and highly transparent workflow.

References

[1] B. Guo *et al.*, ‘SurveyLens: A Research Discipline-Aware Benchmark for Automatic Survey Generation’, Feb. 11, 2026, *arXiv*: arXiv:2602.11238. doi: 10.48550/arXiv.2602.11238.
[2] E. B. Avraham *et al.*, ‘DREAM: Deep Research Evaluation with Agentic Metrics’, Feb. 21, 2026, *arXiv*: arXiv:2602.18940. doi: 10.48550/arXiv.2602.18940.
[3] H. Zhai, E. Stengel-Eskin, P. Patil, and L. Leqi, ‘Evaluating Stochasticity in Deep Research Agents’, Feb. 26, 2026, *arXiv*: arXiv:2602.23271. doi: 10.48550/arXiv.2602.23271.
[4] M. J. Page *et al.*, ‘The PRISMA 2020 statement: An updated guideline for reporting systematic reviews’, *International Journal of Surgery*, vol. 88, p. 105906, Apr. 2021, doi: 10.1016/j.ijsu.2021.105906.
[5] S. Gao *et al.*, ‘Democratizing AI scientists using ToolUniverse’, Oct. 22, 2025, *arXiv*: arXiv:2509.23426. doi: 10.48550/arXiv.2509.23426.
[6] Y. Li *et al.*, ‘AgentCPM-Report: Interleaving Drafting and Deepening for Open-Ended Deep Research’, Feb. 06, 2026, *arXiv*: arXiv:2602.06540. doi: 10.48550/arXiv.2602.06540.

https://doi.org/10.5281/zenodo.22924759

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

[7] J. Zhu *et al.*, 'RE-TRAC: REcursive TRAjectory Compression for Deep Search Agents', Feb. 02, 2026, *arXiv*: arXiv:2602.02486. doi: 10.48550/arXiv.2602.02486.
[8] R. Bedemariam *et al.*, 'Potential and Perils of Large Language Models as Judges of Unstructured Textual Data', Jan. 20, 2025, *arXiv*: arXiv:2501.08167. doi: 10.48550/arXiv.2501.08167.
[9] Y. Chen, J. Jiang, J. Liu, Y. Zhang, X. Guo, and I. King, 'TRACE: Trajectory-Aware Comprehensive Evaluation for Deep Research Agents', Feb. 05, 2026, *arXiv*: arXiv:2602.21230. doi: 10.48550/arXiv.2602.21230.
[10] D. A. Tovar, 'AI Literature Review Suite', Jul. 27, 2023, *arXiv*: arXiv:2308.02443. doi: 10.48550/arXiv.2308.02443.
[11] K. Zheng, Y. Zhong, X. Su, J. Leng, Q. Liu, and X. Chen, 'Towards Agentic Smart Design: An Industrial Large Model-driven Human-in-the-loop Agentic Workflow for Geometric Modelling', *Applied Soft Computing*, p. 113920, Sep. 2025, doi: 10.1016/j.asoc.2025.113920.
[12] S. Wu *et al.*, 'AutoSurvey2: Empowering Researchers with Next Level Automated Literature Surveys', Dec. 02, 2025, *arXiv*: arXiv:2510.26012. doi: 10.48550/arXiv.2510.26012.

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22925327

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

MEASURING CRAWLER LOCK-IN: TRANSLATING SCHEDULING PREFERENCES INTO LANGUAGE COVERAGE FOR Common Crawl

M. Paris*[1], S. Nagel[1]
[1]Common Crawl Foundation, Beverly Hills, USA

EXTENDED ABSTRACT

Large language models are trained predominantly on web text, and the single largest source of that text is Common Crawl—a monthly snapshot of roughly 3 billion web pages constituting 75% or more of the pre-training data for models like GPT and LLaMA [1, 2]. The language distribution of this corpus propagates directly into model capabilities. Yet this distribution follows a steep power law: English accounts for approximately 42% of pages, while dozens of languages spoken by tens of millions of people contribute less than 0.1% [3]. Normalised by speaker population, regions like Turkey, Vietnam, Cambodia, and much of Central Asia are substantially underrepresented (Fig. 1b). The question is whether this underrepresentation is a property of the web—there simply is not more content in these languages—or a property of the crawler.

Common Crawl does not decide what to crawl by language. It decides by graph structure: approximately 3 billion URLs are selected monthly from a database of 25 billion known URLs, with per-domain budget caps derived from harmonic centrality rankings over the host-level web graph [4]. This is a network metric, proven to "deter" link spam [5]. But it's not a language metric. Hosts in well-connected web regions accumulate high centrality regardless of the language they serve. The operator therefore controls the scheduling function (the input) but cares about language distribution (the output), and currently has no mechanism to translate between the two. The operator cannot distinguish "we underrepresent Turkish because there is little Turkish content on the open web" from "we underrepresent Turkish because the link graph does not route crawl attention toward Turkish-language hosts."

We address this gap by applying the temporal containment method [6] to Common Crawl's CDX index—a publicly available, SURT-sorted record of every URL fetched in every monthly crawl—stratified by country-code top-level domain (ccTLD). For two crawls of the same region separated by Δt, the containment (the fraction of one crawl's URLs present in the other) follows $\log g(\Delta t) = \log c - a \cdot \Delta t$, where the y-intercept recovers the sampling fraction c without requiring the total population size N to be known. We analyse all monthly crawls from CC-MAIN-2017-04 through CC-MAIN-2026-04 and focus on eight ccTLDs selected by per-capita normalisation [7]: `.ir` (Farsi), `.tr` (Turkish), `.kz` (Kazakh), `.uz` (Uzbek), `.pk` (Urdu), `.th` (Thai), `.kh` (Khmer), and `.vn` (Vietnamese).

Applied to Common Crawl, the containment curve reveals structure beyond the single log-linear decay observed in focused crawls. Three regimes emerge. At short Δt (1–3 months), overlap is inflated by the recrawl policy—URLs re-selected in consecutive months by the scheduling function. At medium Δt (~4–18 months), the log-linear relationship holds: the stochastic component of the crawl behaves like an urn process from which c and α are recovered per ccTLD. At large Δt (>18 months), containment flattens to a near-constant floor, implying a fixed set K of URLs appearing in every crawl—the harmonic centrality attractor. Each monthly crawl decomposes as $\text{crawl}_t = K \cup S_t$: a deterministic core K plus a stochastic component S_t that genuinely explores new content.

This decomposition provides the translation mechanism between scheduling input and language output. For each ccTLD, we obtain two operational quantities: (i) the relative sampling fraction c, which tells the operator how much of the region's web the crawl currently captures and how much room there is to grow; and (ii) the lock-in ratio $|K|/M$, which measures how much of the region's crawl budget is consumed by deterministically re-fetching the same high-centrality URLs rather than discovering new content. If a ccTLD shows low c and high lock-in, increasing the budget alone will not help—the additional capacity will be absorbed by the attractor. The intervention must instead break the attractor: inject new seeds, randomise a fraction of the URL selection, or increase per-domain budget caps for underserved hosts. Conversely, if c is low but lock-in is also low, the region is genuinely underexplored and additional budget translates directly into new coverage.

The method generalises to any partition of the SURT space—by ccTLD, by second-level domain, or by language when cross-referenced with the columnar index's CLD2 language codes [8]. It provides the feedback loop that is currently missing: measure the per-region consequences of graph-based scheduling, quantify how much of the budget translates into genuine discovery versus deterministic re-fetching, and identify where intervention would yield the largest increase in language coverage.

REFERENCES

[1] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

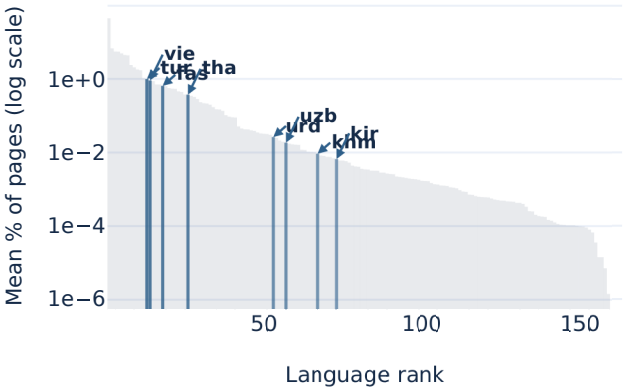
[2] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the*

* micha@commoncrawl.org

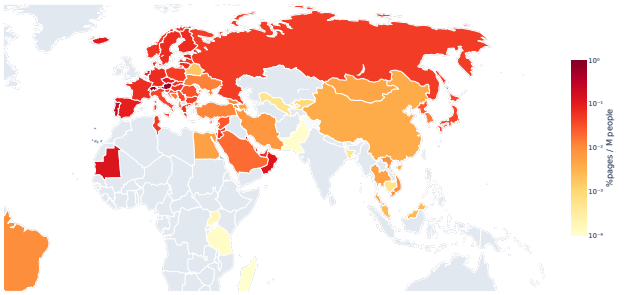
ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22925327

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI



(a) Language distribution in Common Crawl (CC-MAIN-2024-10), ranked by share of pages. English accounts for ~42%; a handful of languages (Russian, German, Japanese, French, Spanish) each contribute 4–7%; the remaining ~300 detected languages share the long tail. Languages spoken by tens of millions—Turkish, Vietnamese, Khmer, Kazakh—fall below 0.1%.



(b) Crawl share of each country's dominant language, normalised by population (millions). Countries are coloured only where a single language unambiguously dominates (89 of 316 ccTLDs); the remainder are grey (multilingual, English-dominant, or repurposed generic TLD). Language mapping from CLDR likely subtags; population from the World Bank (2023).

Figure 1: (a) Language distribution in Common Crawl: English dominates at ~42%; the long tail drops sharply. (b) Per-capita crawl share reveals that Central Asia, South Asia, and Southeast Asia are systematically underserved relative to population size.

2021 Conference on Empirical Methods in Natural Language Processing, pages 1286–1305, 2021.

[3] Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan Van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022.

[4] Paolo Boldi and Sebastiano Vigna. Axioms for centrality. *Internet Mathematics*, 10(3–4):222–262, 2014.

[5] Hsin-Tsang Lee, Derek Leonard, Xiaoming Wang, and Dmitri Loguinov. IRLbot: Scaling to 6 billion pages and beyond. *ACM Trans. Web*, 3(3), July 2009.

[6] Michael Paris, Fabian Baumann, and Grigori Paris. Estimating absolute web crawl coverage from longitudinal set intersections, 2026. `https://doi.org/10.5281/zenodo.18257246`.

[7] The World Bank. Population, total, 2025. `https://data.worldbank.org/indicator/SP.POP.TOTL`.

[8] Henry S Thompson. Improved methodology for longitudinal web analytics using common crawl. In *Proceedings of the 16th ACM Web Science Conference*, pages 59–69, 2024.

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22905553

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

Design Conflicts in Open Search Infrastructures: A Design Pattern-Based Approach to Integrating Legal and User Requirements

S. Meywirth[1], P.C. Johannes[1], C. Werther[2], M. Söllner[1], C. Geminn[1], S. Voigt[2]

[1]Research Center for Information Systems Design (ITeG), Kassel, Germany
[2]Open Search Foundation e.V., Munich, Germany

Abstract

The European Union has established an extensive regulatory framework governing digital technologies, data use, platform governance, cybersecurity, and artificial intelligence. While this framework aims to protect fundamental rights and ensure fair competition, it has also produced a high level of regulatory complexity. In addition to navigating the complex legal requirements meeting user expectations is key for building digital applications that aim for broad acceptance and widespread use and that way foster digital sovereignty in Europe.

The development of open search and Open Web Index infrastructures therefore raises not only technical questions, but fundamental design tensions. Legal requirements such as transparency, accountability, data minimization, and traceability intersect with user expectations regarding usability, personalization, performance, and simplicity. These dimensions do not merely coexist; they frequently stand in tension [1].

In practice, such tensions are often addressed reactively. Legal aspects are interpreted as constraints to be checked after architectural decisions have already been made, and user requirements are often addressed separately from normative considerations. As a consequence, conflicts emerge late in the process; sometimes in the form of compliance risks, sometimes in reduced usability, sometimes in weakened user acceptance. For infrastructures that aim at broad adoption, such reactive conflict resolution can become a structural barrier [1, 2].

For open search and Open Web Index initiatives, this issue is particularly consequential. These infrastructures do not derive their relevance from technical existence alone; their societal and economic impact unfolds only through sustained use. To become a viable alternative, open search systems must therefore be both legally robust and genuinely attractive to potential users. Legal-normative requirements and user-centred design goals cannot be treated as sequential or competing concerns. They must be integrated from the outset, as the legitimacy and the practical uptake of open search infrastructures depend on their ability to satisfy both perspectives simultaneously [3].

This talk focuses on these conflict constellations in the context of open search and index-based systems. Rather than treating legal compliance and user-centred design as separate domains, an interdisciplinary team from University Kassel and Open Search Foundation e.V. with legal and information systems research expertise employed a structured approach that integrates both perspectives directly within the design process. The goal is to support the integration of fundamental rights such as privacy and data protection — understood as "values by design" — in the Open Web Index and applications built on it, such as search engines and AI search, through the development of design patterns that combine legal and user-centred perspectives.

The approach is organized in several steps. First, both legal and user requirements are elicited separately through structured processes. Second, potential conflicts and synergies between the derived legal and user requirements are explicitly identified, documented and analysed. Third, recurring requirement constellations are translated into structured requirement patterns that capture both normative intent and user-related implications. Finally, these requirement patterns are refined into technical design patterns that provide implementable solution structures. The resulting patterns integrate normative requirements and user expectations, supporting the design of web search systems that respect fundamental rights and protect privacy.

By embedding this analysis into the design phase, conflicts are not deferred to later compliance checks but addressed proactively. At the same time, recurring solutions are documented as reusable design knowledge, enabling future projects to anticipate and manage similar tensions more systematically [1, 2].

The core argument is simple: if open search infrastructures are to gain legitimacy and sustained adoption, their central design conflicts must be identified and addressed proactively rather than reactively. In the talk we will illustrate this approach through an exemplary application. Using a concrete design scenario based on search engines and AI, we will reconstruct a typical conflict constellation and demonstrate how these tensions can be systematically identified and translated into requirement and design patterns. This example will highlight how conflicts can be addressed in the design rather than deferred to later compliance or implementation phases. One example is the tension between data minimization and privacy protection and the demands of accountability and search effectiveness. Privacy principles encourage limiting the collection and retention of user interaction data, yet high-quality ranking, personalization, and auditability often depend on exactly this data. The talk will show how such conflicts

https://doi.org/10.5281/zenodo.22905553

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

can be systematically identified, how system design alternatives can be explored, and how the resulting solutions can be translated into reusable design patterns that balance privacy preservation with traceability and performance.

.

References

[1] E. Dickhaut et. al., “Handlungsbroschüre Anforderungs- und Entwurfsmuster zur rechtsverträglichen und qualitätszentrierten Gestaltung kontextsensitiver Applikationen (AnEkA)”,ITeG, Kassel, Germany 2020, https://pubs.wi-kassel.de/wp-content/uploads/2021/02/JML_814.pdf.

[2] E. Dickhaut, L.F. Thies and A. Janson, “Die Kodifizierung von Gestaltungswissen in interdisziplinären Entwurfsmustern”, *Datenschutz und Datensicherheit*, vol 44, pp. 584–588, Sep. 2020, https://doi.org/10.1007/s11623-020-1329-6.

[3] P.C. Johannes, L. Beer and H. Koulani, “ Legal Challenges for OWI and Search Engine Use”, in *Proc. Open Search Symposium 2025 (OSSYM 25), Helsinki, Finland, 8-10 October 2025,* https://doi.org/10.5281/zenodo.17229609.

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22905662

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

PUBLIC DIGITAL INFRASTRUCTURE AS REGULATORY STABILIZER IN A RECALIBRATED AI LANDSCAPE: STRATEGIC LEGAL NAVIGATION UNDER THE EU DIGITAL OMNIBUS ON AI

Kai Erenli, FH des BFI Wien, Vienna, Austria

Abstract

Following the entry into force and phased application of the AI Act, the European Union has entered a regulatory recalibration phase often described as the "Digital Omnibus on AI". This initiative seeks to harmonize overlapping digital regulations and reduce cumulative compliance burdens resulting from the interaction of the AI Act with adjacent instruments such as the Data Act and the Data Governance Act. This paper examines the structural position of foundational datasets within this evolving regulatory framework. While such datasets do not qualify as "AI systems" under Article 3 (1) AI Act, they materially influence compliance obligations imposed on downstream actors under Article 10 (Data Governance) and Article 53 (Transparency for General-Purpose AI Models). This creates a regulatory asymmetry in which upstream data infrastructures remain formally unregulated while simultaneously bearing indirect compliance pressures through contractual and technical dependencies. The paper argues that publicly embedded data infrastructures, such as the Open Web Index, can function as regulatory stabilizers within the European AI ecosystem. By providing standardized metadata, provenance information, and traceable training data structures, such infrastructures may enable downstream developers to meet transparency and governance obligations more efficiently. Building on doctrinal analysis of the AI Act and related EU data governance instruments, the paper proposes a **"***Transparency Safe Harbor***"** mechanism for public-interest data infrastructures. Such a mechanism could reduce regulatory fragmentation, support European digital sovereignty, and mitigate market concentration risks in the emerging AI economy.

INTRODUCTION

The AI Act **[1]** establishes the first comprehensive regulatory framework for AI systems within the European internal market. Structured around a risk-based approach, the regulation introduces obligations for high-risk AI systems, governance mechanisms for general-purpose AI models, and transparency requirements designed to safeguard fundamental rights while promoting innovation. At the same time, the AI Act does not operate in isolation. It forms part of a broader European digital regulatory architecture that includes, among others, the Data Governance Act **[2]** and the Data Act **[3]**. These instruments collectively aim to create a functional European data economy while safeguarding fundamental rights and maintaining fair competition. The combined application of these regulatory instruments, however, has raised concerns regarding **cumulative compliance burdens**, particularly for smaller innovators and research-driven actors. In November 2025, the European Commission proposed the "Digital Omnibus on AI" **[4]** as part of the broader Digital Simplification Package **[5]**, with the aim of streamlining implementation and reducing cumulative compliance burdens across the EU digital regulatory framework.* Within this evolving regulatory landscape, foundational datasets, the large-scale data infrastructures used to train AI systems, occupy a structurally ambiguous position. While these datasets are essential for the development of modern AI systems, the current legal framework does not clearly define their role within the regulatory architecture. Therefore, this paper examines the role of foundational datasets, the bedrock of AI development, within this recalibrated landscape. Beyond regulatory simplification, the debate must be situated within the Union's constitutional digital framework. The AI Act, grounded in Article 114 TFEU, and to be interpreted in light of Articles 7, 8, 11, and 16 of the Charter of Fundamental Rights seeks to ensure a balance that any recalibration through the AI Omnibus must preserve internal market harmonization while safeguarding fundamental rights and innovation capacity. Foundational datasets operate precisely at this intersection.

THE DATA-AS-INFRASTRUCTURE PARADOX

Article 3(1) of the AI Act defines an AI system as a machine-based system designed to operate with varying levels of autonomy and capable of generating outputs such as predictions, recommendations, or content **[1]**. Foundational datasets do not generate such outputs and therefore fall outside the formal definition of an AI system. Despite this exclusion, datasets play a central role in enabling compliance with several key provisions of the AI Act.

First, **Article 10** establishes requirements for data governance and data quality. Providers of high-risk AI systems must ensure that training datasets are "*relevant, representative, and free of errors*" **[1]**. Article 10 further introduces governance requirements regarding data collection processes and bias mitigation. Recent Digital Omnibus discussions suggest a clearer basis for limited processing of

* Nevertheless, it must be mentioned, that a key component of the reform debate has been the introduction of a so-called "stop-the-clock" mechanism, postponing the application of certain high-risk AI obligations in Annex III in order to allow additional time for the development of harmonised technical standards

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935
https://doi.org/10.5281/zenodo.22905662

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

special categories of personal data where this is strictly necessary for bias detection and correction in AI systems. These obligations implicitly depend on the availability of high-quality data infrastructures capable of documenting provenance, structure, and potential biases. Second, Article 53 requires providers of general-purpose AI models to prepare technical documentation and provide a sufficiently detailed summary of the content used for training [1]. These transparency obligations are further operationalized by the General-Purpose AI Code of Practice published by the European AI Office on 10 July 2025, whose Transparency and Copyright chapters are intended to help providers demonstrate compliance with Article 53 AI Act **[6]**, which requires signatories to document training data sources **[7]** and respect rights-reservation mechanisms such as robots.txt signals.**[8]**

Taken together, these provisions create what may be described as a **"***data-as-infrastructure paradox***."** Although foundational datasets are not themselves regulated entities, they become essential enablers of regulatory compliance for AI developers. In practice, this results in a *contractual cascade*: developers subject to AI Act obligations increasingly require upstream data providers to deliver structured documentation and guarantees regarding the datasets used in model training **[9]**. For many actors, particularly small and medium-sized enterprises, the verification of such guarantees remains costly and technically challenging. **[10]**

FOUNDATIONAL DATASETS IN THE AI VALUE CHAIN

Within contemporary AI ecosystems, foundational datasets occupy an upstream position in the technological value chain. While AI developers remain legally responsible for compliance with Articles 10 and 53 AI Act, their ability to meet these obligations depends heavily on the availability of structured, traceable, and well-documented training data. The absence of standardized data infrastructures therefore shifts verification costs toward individual developers, creating what may be described as a compliance externality within the European AI market. Large vertically integrated technology providers can internalize these verification costs by controlling both data resources and AI development pipelines. Smaller actors, however, must rely on fragmented and often poorly documented data sources. Therefore, this dynamic risks reinforcing market concentration within the European AI ecosystem.

PROPORTIONALITY AND PUBLIC DIGITAL INFRASTRUCTURE (PDI)

The emerging "AI Omnibus" policy process offers an opportunity to address this structural imbalance through the application of the **principle of proportionality**. The AI Act already contains mechanisms intended to support innovation. For example, Article 2(6) provides exemptions for AI systems developed exclusively for research and development purposes, while Article 57 introduces regulatory sandboxes that allow controlled experimentation with AI technologies **[1]**. These provisions indicate that the EU legislator recognizes the need to balance regulatory oversight with technological innovation. Publicly embedded data infrastructures (such as the Open Web Index) offer a practical pathway to operationalize this balance. By providing standardized metadata, traceable data sources, and transparent documentation mechanisms, such infrastructures can reduce the compliance burden faced by individual developers. In effect, these infrastructures enable **"***regulatory transparency by design***."** Rather than requiring each AI developer to independently verify the quality and provenance of training datasets, a shared and auditable data infrastructure can provide the necessary information at the systemic level **[4]**.

PROPOSED POLICY RECOMMENDATION

To fully realize the benefits of such infrastructures, this paper proposes the creation of a **Transparency Safe Harbor** for public-interest data infrastructures. Under this model, non-profit or publicly governed data infrastructures, verified as Public Digital Infrastructures (PDIs), would receive limited protection from secondary liability under the AI Act, provided that they adhere to standardized transparency and governance protocols. These protocols could include:

- comprehensive dataset documentation,
- standardized metadata structures,
- traceable data provenance mechanisms, and
- publicly auditable governance frameworks.

Importantly, the proposed safe harbor would not exempt such infrastructures from transparency obligations. Rather, it would **institutionalize transparency as a shared infrastructure layer** within the European AI ecosystem **[10]**. From a regulatory perspective, such a mechanism could be implemented through interpretative guidance issued by the European AI Office or through future legislative clarification within the AI Omnibus framework.

CONCLUSION

The Digital Omnibus debate must be understood within the broader context of the European Data Union Strategy, which seeks to strengthen Europe's data infrastructure through initiatives such as AI Factories and sectoral data spaces.**[5]** Within this landscape, foundational datasets play a critical but currently under-recognized role. While these datasets fall outside the formal definition of AI systems, they function as essential compliance infrastructures for the obligations imposed on AI developers under the AI Act. Recognizing publicly governed data infrastructures as regulatory stabilizers could reduce fragmentation within the European AI ecosystem, support smaller innovators, and strengthen Europe's capacity to develop sovereign AI capabilities. Public digital infrastructures are therefore not merely technical repositories of data. They are emerging as key governance mechanisms within the European AI regulatory architecture.

https://doi.org/10.5281/zenodo.22905662

REFERENCES

[1] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

[2] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act).

[3] Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU)2020/1828 (Data Act).

[4] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI) COM/2025/836 final.

[5] COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL DATA UNION STRATEGY UNLOCKING DATA FOR AI, COM/2025/835 final

[6] European AI Office (2025). *General-Purpose AI Code of Practice*. Brussels, 10 July 2025.

[7] Spindler, G. (2024). *Data Governance and AI – The Interaction between the Data Act and the AI Act*. Journal of Intellectual Property and Information Technology Law.

[8] EDPB-EDPS Joint opinion 1/2026 on the Proposal for a Regulation as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI).

[9] Hacker, P. (2024). *The European AI Act: Regulating Core AI Technologies and High-Risk Applications*. European Law Journal.

[10] Bradford, A. (2023). *Digital Empires: The Global Battle to Regulate Technology*. Oxford University Press. (Contextualizing the 'Brussels Effect')

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22904054

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

Adaptive Unsupervised Topic Modeling for Domain-Focused Web Search

Hema Pandey*, Jens Kersten†,
Institute of Data Science, German Aerospace Center (DLR), Jena, Germany

Abstract

The exponential growth of heterogeneous web content presents major challenges for domain-specific search and retrieval, particularly in dynamic environments where annotated data, curated ontologies, and stable taxonomies are unavailable. The quality of search can be improved by metadata extraction and enrichment. Web page classification, i.e., assigning one or more thematic classes to web pages, is a central task in this regard. One example would be to find all relevant documents in a set of news pages reporting about flood events induced by heavy rainfall in a specific area. Since we would like to offer such search capabilities for a wide variety of domains and questions/requests, a flexible concept is required.

Classical supervised web page classification approaches rely on predefined category schemes and curated training corpora [1], limiting adaptability in rapidly evolving contexts such as natural disasters or emerging events. Probabilistic topic models such as Latent Dirichlet Allocation [2] offer unsupervised alternatives, yet often struggle with short, noisy, and semantically complex web documents. Recent advances in contextualized language models and neural topic modeling [3][4] provide semantically richer document representations that better capture non-linear topic structure. Density-based clustering methods, such as HDBSCAN [5] further allow automatic discovery of cluster structure without requiring a predefined number of topics.

In this abstract, we address the adaptable and content-based classification of web pages in order to support specific and domain-focused search scenarios. To achieve that, we present a modular, unsupervised pipeline for adaptive web page classification and thematic structuring designed to support focused web search scenarios. The framework combines transformer-based document embeddings with density-based clustering within a neural topic modeling architecture. Topic representations are refined using class-based TF-IDF, and interpretable labels are generated through embedding-based keyword extraction. The system supports automatic topic reduction and flexible granularity control without relying on prior labels or fixed taxonomies.

As a practical application, we demonstrate the pipeline on web data collected during the 2021 Ahrtal flood, a crisis scenario characterized by rapidly evolving and heterogeneous information sources. The results illustrate how unsupervised thematic structuring can support domain-focused web exploration and provide structured signals that can be integrated into open search infrastructures for adaptive indexing and content discovery.

* hema.pandey@dlr.de
† jens.kersten@dlr.de

The proposed methodology contributes a scalable and semantically interpretable component for open search systems operating in dynamic information environments. We contribute a modular and transferable thematic enrichment component that can be integrated into focussed crawling and open search infrastructures.

CONTRIBUTIONS

The main contributions of this work are:

- a modular neural topic modeling pipeline tailored to domain-focused web search,
- a reference-based transfer mechanism for adaptive thematic labeling of newly crawled web content, and
- an application to crisis-oriented web data demonstrating practical integration potential for open search infrastructures.

METHODOLOGY

System Overview

The proposed pipeline is designed as a modular framework to enable reusable topic modeling and scalable thematic enrichment in focused crawling scenarios. It operates in two stages: (1) construction of a reference topic model and (2) adaptive assignment of topics to newly crawled web content. The architecture integrates semantic representation learning, density-based clustering, topic refinement, and automatic labeling.

Data Preprocessing

Raw web page content is cleaned by removing URLs, non-alphabetic characters, and multilingual stopwords. Documents are filtered by language and minimum token length to reduce noise and ensure semantic richness. This preprocessing step improves robustness against heterogeneous web data.

Semantic Representation

Documents are encoded using a transformer-based sentence embedding model derived from BERT [3]. Contextualized embeddings capture semantic similarity beyond lexical overlap, enabling grouping of topically related documents even when vocabulary differs significantly.

Neural Topic Modeling and Clustering

Topic discovery is performed using a neural topic modeling framework [4], which combines:

- Dimensionality reduction via UMAP
- Density-based clustering using HDBSCAN [5]

https://doi.org/10.5281/zenodo.22904054

- Class-based TF-IDF (c-TF-IDF) for topic representation

Unlike k-means clustering, density-based clustering does not require specifying the number of clusters in advance and can detect noise points. An automatic topic reduction mechanism merges semantically similar clusters, enabling flexible control over thematic granularity.

Topic Labeling and Transfer

For interpretability, cluster-specific keywords are extracted using embedding-based keyword extraction. The trained reference topic model is subsequently applied to newly crawled web pages via embedding-based topic assignment, enabling consistent thematic labeling across evolving web collections. This enables adaptive, taxonomy-free thematic enrichment for domain-focused search systems.

REFERENCES

[1] X. Qi and B. D. Davison, "Web page classification: Features and algorithms," *ACM computing surveys (CSUR)*, vol. 41, no. 2, pp. 1–31, 2009.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[4] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[5] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*, Springer, 2013, pp. 160–172.

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22904833

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

EXTRACTION OF GEOSPATIAL DISASTER RELATED INFORMATION FROM WEB NEWS AND TEXT DATA

Vanessa Rittlinger[a], Stefan Voigt[a], Christian Geiß[a,b], Hannes Taubenböck[a,c]

[a]German Aerospace Center (DLR),German Remote Sensing Data Center, Weßling, Germany;

[b]University of Bonn, Department of Geography, Bonn, Germany

[c]University of Würzburg, Institute of Geography and Geology, Würzburg, Germany

Abstract

Disaster management relies on timely, fine-grained information about the evolution of a disaster situation, event impact, response activities, and mitigation measures to ensure a broad assessment of the event [1]. Traditional sensor-based sources (e.g., satellite imagery, in situ measurements) provide objective physical observations but often lack contextual details, such as deployed emergency units, applied mitigation measures, or community impacts [2]. Public web news, in contrast, offer valuable insights and can often be geolocated precisely through place names and toponyms [3]. However, systematically extracting actionable, spatially referenced information from unstructured news sources remains a scientific and operational challenge [4]. This requires robust natural language processing (NLP) pipelines capable of (i) detecting disaster-related information and (ii) geolocating it with sufficient precision [5]. In this study, we investigate the geospatial level of detail and extent of relevant information extractable from news and web text data focussing on flooding events. We further compare this web-based disaster event information with other geospatial datasets, such as flood extent maps and building damage assessments derived from satellite and airborne imagery.

The focus is on flood events in Germany, with a region with extensive media coverage and scientific relevance in German language. For this purpose, we analyze a database of web news texts extracted from GDELT knowledge graph database (GKG) [6] URLs and pre-filtered by language, flood-related topics, the study area and the temporal phase of the event (including the time after the event). Filters were set to prioritize event focus and big data scalability. To enhance situational awareness in disaster management, this study synthesizes multi-source geospatial datasets, integrating Copernicus Emergency Management Service (CEMS) [7] building damage assessments with satellite-derived flood extent maps.

We present a modular workflow combining rule-based, dictionary-based, and machine learning NLP methods to semi-automatically extract disaster-relevant information from a large web news corpus. The focus lies on identifying specific disaster-related terms, which we categorize using a dictionary-based system aligned with the disaster management cycle. This enables structuring extracted terms into categories such as event impact, emergency measures, and mitigation/reconstruction. To add spatial value, we employ the FlairNER toponym identification model [8] alongside geocoding techniques. During spatial linking, we merge toponyms with disaster-relevant information via a rule-based entity linking approach, generating geospatial vector point layers enriched with event-specific data. These layers can be synergistically compared with other geospatial datasets and contextualized.

Our results demonstrate the systematic extraction of structured, spatially located disaster information. The spatial point layer integrates thematic, spatial, and temporal components, capturing data at multiple geospatial scales (e.g., municipalities, city districts, rivers, landmarks). By comparing news-derived data with existing datasets, we identify spatial patterns, such as hotspots, across thematic disaster related classes. Web news and text data reveals publicly available information on flood events, including spatial distributions of emergency relief activities (e.g., firefighters) and mitigation measures. These findings highlight the potential of web-based data to complement remote sensing-based products for disaster management support. Furthermore, it enables the potential of geospatial analysis within the text-based data output over time and across spatial and thematic classes.

Our evaluation underscores both the capabilities and limitations of rule-based and dictionary-based approaches for semi-automated text analysis. We further outline opportunities for advanced geospatial analysis, data fusion, and integration with other geospatial datasets. While our workflow highlights the value of unstructured text for disaster insights, it also emphasizes the need for domain-specific, state-of-the-art NLP methods to enhance information accuracy and geospatial referencing. Building on this, we anticipate that even richer disaster-related information can be derived for future mitigation and management support.

ACKNOWLEDGEMENTS

This work has received funding from the project OpenSearch@DLR phase II (internal DLR project).

REFERENCES

[1] J. P. De Albuquerque, B. Herfort, A. Brenning, und A. Zipf, „A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management“, *Int. J. Geogr. Inf. Sci.*, Bd. 29, Nr. 4, S. 667–689, Apr. 2015, doi: 10.1080/13658816.2014.996567.

[2] M. Wieland, S. Schmidt, B. Resch, A. Abecker, und S. Martinis, „Fusion of geospatial information from

https://doi.org/10.5281/zenodo.22904833

remote sensing and social media to prioritise rapid response actions in case of floods“, *Nat. Hazards*, Jan. 2025, doi: 10.1007/s11069-025-07120-7.

[3] I. Owuor, H. H. Hochmair, und S. Cvetojevic, „Tracking Hurricane Dorian in GDELT and Twitter“, *AGILE GIScience Ser.*, Bd. 1, S. 1–18, Juli 2020, doi: 10.5194/agile-giss-1-19-2020.

[4] M. Madruga de Brito, J. Sodoge, H. Kreibich, und C. Kuhlicke, „Comprehensive Assessment of Flood Socioeconomic Impacts Through Text-Mining“, *Water Resour. Res.*, Bd. 61, Nr. 1, S. e2024WR037813, 2025, doi: 10.1029/2024WR037813.

[5] X. Hu *u. a.*, „Location Reference Recognition from Texts: A Survey and Comparison“, *ACM Comput. Surv.*, Bd. 56, Nr. 5, S. 1–37, Mai 2024, doi: 10.1145/3625819.

[6] „GDELT-Global Knowledge Graph Codebook V2.1“. 2024. Zugegriffen: 8. Mai 2024. [Online]. Verfügbar unter: http://data.gdeltproject.org/documentation/GDELT-Global_Knowledge_Graph_Codebook-V2.1.pdf

[7] „COPERNICUS EMERGENCY MANAGEMENT SERVICE“, Copernicus EMS - Mapping. Zugegriffen: 15. Juli 2024. [Online]. Verfügbar unter: http://emergency.copernicus.eu/mapping/list-of-components/EMSR517

[8] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, und R. Vollgraf, „FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP“, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, W. Ammar, A. Louis, und N. Mostafazadeh, Hrsg., Minneapolis, Minnesota: Association for Computational Linguistics, Juni 2019, S. 54–59. doi: 10.18653/v1/N19-4010.

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22905339

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

A SPATIALLY ENABLED SEARCH SYSTEM FOR HYDROLOGICAL PUBLICATIONS: A DEMONSTRATION

Xuke Hu[†1], Lina Stein[2], Jens Kersten[1], Friederike Klan[1]
[1] Institute of Data Science, DLR, Jena, Germany
[2] iES, Institute for Environmental Sciences,
RPTU University Kaiserslautern-Landau, Landau, Germany

Abstract

In hydrology research, scholars often need to identify publications that analyze events such as floods, droughts, and landslides in a specific region. This task is still largely performed manually, which is time-consuming and difficult to reproduce. To address this need, we present a web-based geo-search system for hydrology papers that supports retrieval by geographic constraints, keywords, publication year, and journal filters. Specifically, we extract location entities (toponyms) from abstracts using Flair-based named entity recognition, and then apply an LLM-assisted geocoding method to resolve each toponym into geospatial representations such as coordinates. These structured spatial annotations enable map-based exploration and spatial filtering of the literature. The proposed system can improve the efficiency of regional evidence collection in hydrology studies.

INTRODUCTION

Hydrological studies are inherently place-based, and many hydrological analyses are formulated at the catchment or basin scale [1][2]. Analyses of floods, droughts, and landslides are typically tied to a catchment, river reach, administrative unit, or a clearly defined event footprint. When researchers compile evidence for a region (e.g., to compare past events, support model calibration, or prepare a review for a basin authority), the first step is often to identify publications that investigated similar events in the same area [3]. In practice, this step is still handled largely through manual searching and screening, which is slow and difficult to reproduce.

A central limitation is that most scholarly search interfaces are not spatially aware. They support topical retrieval and basic metadata filters (year, venue, author), but they rarely support queries such as studies within a specific region. In addition, locations are typically embedded in unstructured text, and the same study area may be described using multiple names, local spellings, or indirect references. Toponym ambiguity in scientific articles is a well-known challenge in geoparsing[4], and practical disambiguation often requires matching against a gazetteer such as GeoNames [5].

To improve location-based discovery, we develop a web-based geo-search system for hydrology publications that supports retrieval by geographic constraints in addition to conventional filters. The system indexes papers using locations mentioned in their abstracts and exposes these spatial references through an interactive map interface. Users can specify a spatial query (e.g., a bounding box) and refine results by keywords, journals, and publication year, enabling a workflow that better matches how hydrologists formulate region-focused evidence-gathering tasks.

The geo-search pipeline combines standard NLP components with an LLM-assisted geocoding step. First, we apply Flair named entity recognition (NER) [6] to extract location entities (toponyms) from article abstracts. Second, we resolve these toponyms to geographic representations, such as coordinates and canonical place references (e.g., GeoNames records), using an LLM-based toponym resolution approach [6], so that each paper can be associated with spatial indices for efficient querying. By converting textual location mentions into structured geospatial metadata, the system reduces reliance on ad hoc place-name keyword searches and supports more systematic and reproducible literature retrieval for hydrological events and study regions.

PROPOSED TOOL

Figure 1 shows the current web-based demonstration interface. The left panel provides keyword, temporal, and journal filters together with a study-location selector. The central map visualizes resolved article locations as markers and clusters; selecting a marker opens a popup with the place name, article title, year, journal, and clickable DOI link. The right panel lists the matched records, supports pagination, and offers download options (CSV/JSON). Users can draw or manually define a bounding box to restrict results to papers with at least one resolved location inside the selected area. Because a single abstract may mention multiple locations, the same paper can appear at multiple mapped points. This design supports rapid region-focused screening while preserving direct access to the underlying articles.

[†] Corresponding author: xuke.hu@dlr.de

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22905339

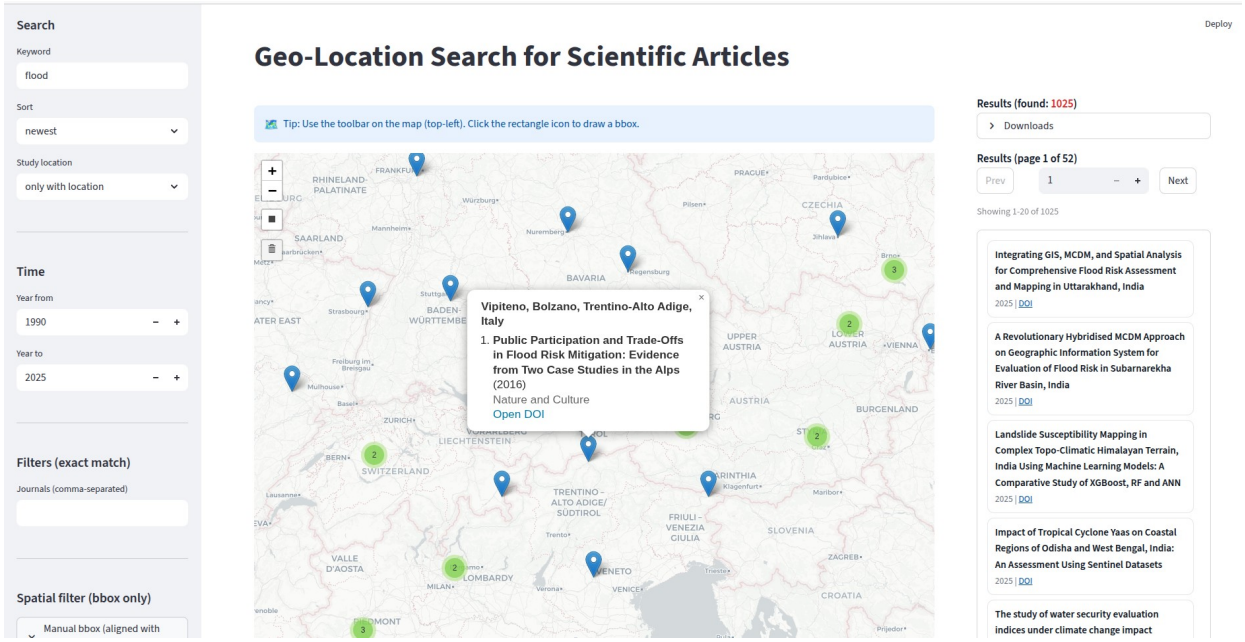


Figure 1: Screenshot of the geo-location search interface showing filters (left), mapped article locations with a popup preview (center), and the paginated results/download panel (right).

CONCLUSION

In this abstract, we presented a demonstration of a geo-based search system for hydrology publications that combines text-based retrieval with map-based spatial filtering. By extracting toponyms from abstracts and resolving them to geospatial coordinates, the system enables faster and more reproducible region-focused literature discovery. The current prototype also has several limitations. First, it visualizes all detected locations without yet distinguishing event locations from other mentioned places (e.g., comparison regions). Second, the present version is an offline demonstration rather than a public online service. Third, the search component can be extended with more advanced ranking and semantic query capabilities. Future work will focus on event-location linking, online deployment, and improved retrieval quality and usability.

References

[1] I. Pechlivanidis, B. M. Jackson, N. R. McIntyre, and H. S. Wheater, "Catchment scale hydrological modelling: A review of model types, calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications", Global NEST J., vol. 13, no. 3, pp. 193-214, 2011. doi:10.30955/gnj.000778

[2] L. Stein, S. K. Mukkavilli, B. M. Pfitzmann, P. W. J. Staar, U. Ozturk, C. Berrospi, T. Brunschwiler, and T. Wagener, "Wealth Over Woe: Global Biases in Hydro-Hazard Research", Earth's Future, vol. 12, no. 10, e2024EF004590, 2024. doi:10.1029/2024EF004590

[3] L. Stein, S.K. Mukkavilli, and T. Wagener, 2022, "Lifelines for a drowning science-improving findability and synthesis of hydrologic publications", Hydrological Processes, 36(11), p.e14742.

[3] D. Weissenbacher, A. Akhondi, P. M. Nadkarni, and L. Gonzalez-Hernandez, "SemEval-2019 Task 12: Toponym Resolution in Scientific Papers", in Proc. 13th Int. Workshop on Semantic Evaluation (SemEval'19), Minneapolis, MN, USA, Jun. 2019, pp. 907-916. doi:10.18653/v1/S19-2155

[4] GeoNames, http://www.geonames.org/

[5] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP", in Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Minneapolis, MN, USA, Jun. 2019, pp. 54-59. doi:10.18653/v1/N19-4010

[6] X. Hu, J. Kersten, F. Klan, and S. M. Farzana, "Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge", Int. J. Geogr. Inf. Sci., pp. 1-28, 2024. doi:10.1080/13658816.2024.2405182

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22916034

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

PRIVACY-ENHANCING DIGITAL INFRASTRUCTURES - UNDERSTANDING USER PREFERENCES REGARDING INTERNET SEARCH AND THEIR INCLUSION INTO SYSTEMS DESIGN -

S. Götz, Information Systems and Systems Engineering, University of Kassel, Research Center for IS Design (ITeG), Kassel, Germany

M. Hänel, Information Systems and Systems Engineering, University of Kassel, Research Center for IS Design (ITeG), Kassel, Germany

M. Söllner, Information Systems and Systems Engineering, University of Kassel, Research Center for IS Design (ITeG), Kassel, Germany

Abstract

Search engines form a central infrastructure of modern information societies. The market is currently dominated by a few global providers who have proprietary web indexes and thus effectively act as gatekeepers of information access. This dominance is explained not only by technical superiority, but also by habituation effects, perceived service quality, usability, and other subjective evaluation dimensions. At the same time, there are considerable challenges in terms of transparency, control, and informational self-determination. Against this backdrop, the project "Privacy-enhancing digital infrastructures" (PriDi) aims to support the development of a European open search infrastructure that aims at strengthening diversity of information, individual choice, and digital sovereignty.

However, an open web index can only be a viable alternative to existing offerings if it is not only legally and technically viable but also perceived as attractive and accepted by users.

Despite their high societal relevance, search engines remain underrepresented in user acceptance research. While information systems research has produced extensive insights into the acceptance of new technologies [4, 6, 7], search engines have only rarely been examined as a specific context within this research stream. Furthermore, findings on legal compliance have only been incorporated to a limited extent into research on user acceptance of new technologies. Only the construct of privacy establishes a corresponding link in international research [1, 2]. Specific design requirements for search engines that are oriented toward the best possible realization of fundamental rights (values by design), and at the same time consider findings on user acceptance of new technologies have not yet received considerable attention in prior research. The project aims to close this gap.

The aim of the acceptance-related project is therefore to identify key user-related requirements for open internet search in a theory-driven manner, to validate them empirically, and to translate them into structured requirement patterns. Based on a literature review, we identified 26 attributes for user acceptance of open internet search. The quantitative evaluation is carried by surveying 12.000 Europeans in 12 countries. The best-worst scaling method is used to determine the importance of the attributes. According to Omre [5] 20 randomly compiled sets, four attributes are compared, with participants selecting the most important and least important attribute in each set. This approach enables the identification of those characteristics that are decisive for the acceptance of an open search infrastructure from the user's point of view.

Based on these results, acceptance-related requirement patterns are developed. These patterns structure empirically based design knowledge in a standardized form and make it systematically usable for development processes. Comparable to canvas-based innovation models, they serve as a guide for the design of new search services by making central user priorities transparent and operationalizable.

The project thus expands existing acceptance theories to include the specific context of open search infrastructures and combines empirical social research with design-oriented research in the sense of theory-driven design [3]. It contributes to the user-centered development of European search infrastructures and thus addresses a central prerequisite for their success in the European or even global market as well as in society.

At the time of the conference, we expect to be able to report on the results of the quantitative study and we expect to also provide initial insights into the corresponding pattern development.

REFERENCES

[1] Bélanger, F., & Crossler, R. E. (2011). Privacy in the Digital Age: A Review of Information Privacy Research in Information Systems. MIS Quarterly, 35(4), 1017–1041.

[2] Bélanger, F., & James, T. L. (2020). A Theory of Multilevel Information Privacy Management for the Digital Era. Information Systems Research, 31(2), 510–536.

[3] Briggs, R. O. (2006). On theory-driven design and deployment of collaboration systems. International Journal of Human-Computer Studies, 64(7), 573–582.

[4] Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. Management Science, 35(8), 982–1003.

[5] Orme, B. 2005. "Accuracy of HB Estimation in MaxDiff Experiments," Sawtooth Research Paper Series , pp. 1–7.

https://doi.org/10.5281/zenodo.22916034

[6] Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. MIS Quarterly, 27(3), 425–478.

[7] Venkatesh, V., Thong, J. Y. L., & Xu, X. (2016). Unified Theory of Acceptance and Use of Technology: A Synthesis and the Road Ahead. Journal of the Association for Information Systems, 17(5), 328–376.

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22916580

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

Renegotiating the open web: Emerging AI Preferences standards

S. Baack†, Independent

ABSTRACT

Unlike traditional web search, popular generative AI products have an extractive relationship to the open web. Products like ChatGPT or Perplexity gather web data at large scales for training and inference, yet are less likely to direct users to the original sources of this data [1]. As a result, content creators increasingly block known AI crawlers or enter exclusive licensing deals with AI companies [2]. These developments shrink the open web and reinforce concentration around a handful of resourceful AI companies, as AI builders with fewer resources increasingly lose access to high-quality data.

In response, new web standards are emerging to make the relationship between generative AI and the open web more reciprocal, giving website owners and content creators more control over how their data can be used for AI and requiring attribution. A well-designed public standard could protect public-interest access, give creators meaningful control and strengthen open search. Done poorly, however, new standards risk entrenching power imbalances and excluding researchers, journalists, non-profits, and open search projects from web data access. Due to the high stakes for building sovereign, open search indexes and products, this paper explores the development of the AI Preferences vocabulary by the Internet Engineering Task Force (IETF).

The IETF's vocabulary is one among several proposals that seek to expand the Robots Exclusion Protocol, which enables control over *who* can access data, with machine-readable signals about *how* data can be used. The IETF's work stands out as the most prominent open, multi-stakeholder effort bringing together a broad range of stakeholders negotiating primarily on a public mailing list (AI developers, publishers, user rights organizations, policymakers and more). In this paper, I study the evolution of the IETF's AI Preferences vocabulary by looking at the changes between the various draft proposals published since the group's inception in January 2025. Changes between those drafts point to major conflicts among the involved parties and reveal how the infrastructural conditions of open search are being renegotiated through technical standards processes.

A preliminary analysis of the first five draft proposals for the vocabulary [3] and meeting notes for in-person meetings of the working group [4] shows that most conflicts revolve around drawing the line between “search” and “summary.” Rights holders like publishers advocated for ways to opt out of "substitutive" uses of content, such as generated summaries, while still appearing in search engines. AI developers and user rights advocates have pushed back against what could evolve into a radical extension of copyright laws. As a result, earlier drafts clearly distinguished between "AI Use" and "Search," but this distinction has been dropped. The conflict has subsequently shifted around the precise definition of search. In the most recent draft at the time of writing (v5) "search applications" are those that a) provide a "reference to the location that the asset was obtained" from, and b) only use "excerpts that are drawn verbatim" from the source [5]. In other words, "Search" was disentangled from a specific product category ("traditional search engines") and tied to output characteristics, effectively covering both search engines and inferences by AI products responding to user queries. The only other category in draft v5, "Foundation Model Production Category," is about allowing or disallowing the use of content to train or fine-tune foundation models.

If the final IETF proposal closely mirrors the current draft, it will have implications for the ecosystem of applications built on top of data taken from open web indexes by constraining how they can present content to users, but are unlikely to constrain the indexes themselves. However, it is likely that further changes with potentially wide-ranging implications will be introduced before the standard is finalized, not least because the IETF's vocabulary competes with several prominent private initiatives with powerful backing [6]. Cloudflare's "pay-per-crawl" marketplace, facilitated by its own "content signals," is already operational and positions a single (US) infrastructure company as the arbiter of data access. Similarly, the Really Simple Licensing protocol as another prominent initiate is mostly supported by large US publishers. If such proprietary solutions dominate before public standards mature, the terms of access will be set by corporate gatekeepers with unclear implications for open search.

The open search community has a direct stake in emerging web standards around generative AI. It should take a more active role in its development and support the IETF’s and other open standard over proprietary initiatives. Otherwise, open search projects, together with researchers, non-profits or journalists relying on access to public data, risk becoming “secondary victims” [2] in the ongoing tension between data creators and commercial AI developers.

† hi@sbaack.com

REFERENCES

[1] N. Newman, “Journalism and technology trends and predictions 2026,” Reuters Institute for the Study of Journalism, 2026. doi: 10.60625/RISJ-PS1D-NP11.

[2] S. Longpre et al., “Consent in Crisis: The Rapid Decline of the AI Data Commons,” Jul. 24, 2024, arXiv: arXiv:2407.14933. doi: 10.48550/arXiv.2407.14933.

https://doi.org/10.5281/zenodo.22916580

[3] P. Keller and M. Thomson, “A vocabulary for expressing AI usage preferences,” Internet Engineering Task Force, Internet-Draft, Dec. 2025. [Online]. Available: https://datatracker.ietf.org/doc/draft-ietf-aipref-vocab/

[4] AI Preferences Working Group, “Meetings for aipref.” Accessed: Mar. 13, 2026. [Online]. Available: https://datatracker.ietf.org/wg/aipref/meetings/

[5] P. Keller and M. Thomson, “A Vocabulary For Expressing AI Usage Preferences,” Internet Engineering Task Force, Internet Draft draft-ietf-aipref-vocab-05, Dec. 2025. Accessed: Feb. 26, 2026. [Online]. Available: https://datatracker.ietf.org/doc/draft-ietf-aipref-vocab

[6] P. Keller, “Divergent Mechanisms Elusive Vocabularies,” Open Future. Accessed: Feb. 04, 2026. [Online]. Available: https://openfuture.eu/publication/divergent-mechanisms-elusive-vocabularies

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

ISBN: 978-92-9083-744-2 Open Search Symposium 2026 - #ossym2026 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.22904472

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

eMarketShield: Enhancing Consumer Protection in E-Commerce through Digital Innovation and Regulatory Oversight

C. Rinner[2], AGES - Austrian Agency for Health and Food Safety, Vienna, Austria
S. Bauernfeind, University of Applied Sciences Upper Austria, Campus Hagenberg, Austria
S. Fenz, F. Weber, Xylem Technologies, Vienna, Austria

Abstract

The rapid expansion of e-commerce introduces significant challenges for consumer protection, including limited product transparency, restricted return rights, and difficulties in regulatory oversight, particularly across borders. This study presents the eMarketShield platform—a digitally sovereign, distributed, event-driven system leveraging Apache Pulsar to enhance real-time market surveillance and consumer safety in the EU's online sector. The research outlines the platform's architecture and evaluates its efficacy in identifying vendors and unsafe products.

INTRODUCTION

E-commerce growth has complicated the roles of consumers and authorities, with issues stemming from unverifiable online information, opaque product rankings, and the anonymity of sellers. Existing legal frameworks inadequately address these challenges, allowing unsafe products to reach consumers more easily and undermining consumer protection [1]. Official controls in e-commerce should not rely on commercial search engines such as Google, Bing, Yandex, Yahoo or Baidu, as these platforms operate under proprietary algorithms and opaque indexing criteria that prioritise advertising revenue and user engagement over regulatory relevance. Dependence on such infrastructures compromises the transparency, neutrality, and completeness of market surveillance activities, potentially omitting critical data from smaller or non-optimised vendors.

METHODS

The EU Product Compliance Network outlines the key stages of official controls as targeting, sampling, compliance assessment, and follow-up. eMarketShield focuses on the critical targeting step by systematically identifying and mapping online vendors, a task made urgent by international evidence indicating that the number of unregistered or unknown sellers in the digital marketplace remains disproportionately high. We develop a cloud-based, distributed system built on Apache Pulsar, enabling real-time monitoring of e-commerce platforms. The approach adapts proven risk-based controls from traditional retail to the digital context, utilising machine learning (ML) to analyse online listings. The system integrates a dedicated crawling and scraping pipeline to reliably obtain webpage content alongside a ML pipeline to identify relevant attributes. The scraper improves throughput by employing either a lightweight HTTP request or a headless browser, based on JavaScript complexity. Scraped data is triaged in a pre-classification step that optimistically labels webpages based on content, subsequently being archived in Web ARChive (WARC) files to facilitate high-fidelity replay. After archival, the classification stage analyses each captured page to determine whether it represents a functional commercial shop, assess the completeness of the content and extract relevant attributes such as vendors details, country or commodity.

RESULTS

Our platform enables continuous monitoring of online vendors and identifies previously unknown vendors to competence authorities more efficiently than manual methods. Automated data analysis and machine learning enable a scalable approach and facilitate a responsive feedback loop between domain experts and the system's ML pipeline to continuously improve the algorithms similar to MediCrawl [2]. In a next step we want to evaluate the inclusion of open web indexing like Open Web Index [3] or other available open data sources like CommonCrawl to reduce crawling resources and improve scalability.

DISCUSSION

By leveraging sovereign digital platforms and own data indices, European competent authorities can reclaim oversight of their digital marketplaces, reducing reliance on foreign e-commerce giants and search providers. This approach aligns with the OSSYM 2026 theme of boosting digital sovereignty and sets a foundation for improved consumer protection and regulatory control in the evolving landscape of e-commerce.

REFERENCES

[1] A. Lattanzi, C. Bullon und T. Viinikainen, „Regulatory options to address food e-commerce in national legislation“ Food and Agriculture Organization of the United Nations, Rome, 2026.

[2] N. P. Gonzalez, „Medicrawl: an ML based system to automate the search of illegal products in ecommerce“ In: Proceedings of the 9th edition of the Swiss Text Analytics Conference. 2024. S. 178-178.

[3] M. Granitzer, et al. „OpenWebSearch. eu-Building an Open Web Index on EuroHPC JU Infrastructures,“ Procedia Computer Science, Bd. 255, pp. 43-52, 2025.

[2] Christoph.rinner@ages.at

SOURCE: Scalable, Open, and Comprehensive Detection of Disinformation Campaigns on the Web

Michael Granitzer, Jelena Mitrović, Laura Caspari, Hai Dang Dang, University of Passau, Germany
Martin Potthast, Sebastian Heineking, University of Kassel, hessian.AI, and ScaDS.AI, Germany
Stefan Voigt, Ursula Gmelch, Open Search Foundation, Germany Megi Sharikadze, Damir Kalimullin, Sibylle Richter, BADW-LRZ, Germany Benjamin Zeeb, Alliance4Europe, Germany

Extended abstract

The web has become an information battleground, where disinformation – deliberate falsehoods designed to deceive – undermines trust in institutions, influences elections and polarises societies. Mis- and disinformation rank as the fith most severe global risk over the next two years [1]. The threat has shifted from individual false claims to coordinated, cross-platform campaigns: in 2025, the EEAS documented 540 incidents of foreign information manipulation and interference involving about 10,500 channels, with over 90% of the most active ones covert, and the number of AI-enabled incidents more than tripled within one year [2]. Generative AI makes convincing content almost free to produce, with more than 3,700 AI-generated "news" sites tracked to date [3]. It also turns the web into an attack surface for AI itself: the "Pravda" network published about 3.6 million articles in 2024, aimed at crawlers rather than readers, and leading chatbots repeated its false narratives in 33% of test cases (*LLM grooming*) [4].

Observatories such as the European Digital Media Observatory (EDMO) and the Social Observatory for Disinformation and Social Media Analysis (SOMA) have made disinformation visible, but rely largely on manual monitoring and on commercial search engines that offer limited filtering, little provenance information and no bulk access. Research access to platform data under the Digital Services Act is only now being implemented [5]. We see four open challenges: (i) open, targeted data collection across the web and social media; (ii) robust detection of generated and manipulative content, as AI-text detectors perform well on benchmarks [6] but are evaded by paraphrasing [7]; (iii) tracing provenance and coordination, since campaigns reveal themselves through content re-use and coordinated actors rather than single items [8]; and (iv) human validation embedded in analysts' workflows.

This talk presents the SOURCE project [9], an transdisciplinary initiative designed to detect and analyze disinformation campaigns across the web and social media. The project is built on the foundation of the Open Web Index (OWI) [10], a large-scale database developed through the EU-funded OpenWebSearch.EU project [11]. SOURCE extends the OWI by developing an open source platform that combines advanced AI techniques with a collaborative, community-driven approach to disinformation detection and mitigation. The project's core innovation lies in its dual focus on both the *provenance* (origin and propagation) and the *content* (text, image, and video) of online information, providing a holistic view of disinformation campaigns.

By integrating provenance-based and content-based detection methods within a collaborative, community-driven framework, SOURCE establishes a unique, open infrastructure for real-time monitoring and analysis. The project's builds a robust, open source detection platform, a community of practice for disinformation analysts, and a sustainable infrastructure for future research. SOURCE directly addresses the critical need for transparent, scalable, and collaborative tools to combat the growing threat of disinformation.

The SOURCE project combines cutting-edge AI technology with a community-driven model of open science, harnessing the power of modern and frontier HPC capabilities. It creates a powerful, sustainable, and scalable infrastructure for the detection and analysis of disinformation. The project directly addresses the limitations of existing solutions and provides a much-needed, open, and transparent alternative. SOURCE is not just a research project; it is a foundational step towards building a more resilient and trustworthy digital information ecosystem for Europe and the world.

REFERENCES

[1] World Economic Forum. The Global Risks Report 2026, January 2026.
[2] EEAS. 4th EEAS Report on Foreign Information Manipulation and Interference Threats, March 2026.
[3] NewsGuard. AI Tracking Center: AI content farm sites, updated June 2026. https://www.newsguardtech.com/
[4] NewsGuard. A well-funded Moscow-based global "news" network has infected Western AI tools with Russian propaganda, March 2025. www.newsguardtech.com
[5] European Commission. Delegated act on data access under the Digital Services Act, July 2025.
[6] J. Bevendorff et al. Overview of the "Voight-Kampff" generative AI authorship verification task at PAN and ELOQUENT 2024. CLEF 2024 Working Notes.
[7] K. Krishna et al. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. NeurIPS 2023.
[9] https://source-project.de
[10] G. Hendriksen et al. The Open Web Index: Crawling and indexing the web for public use. ECIR 2024.
[11] https://openwebsearch.EU

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI

Appendix

List of Authors

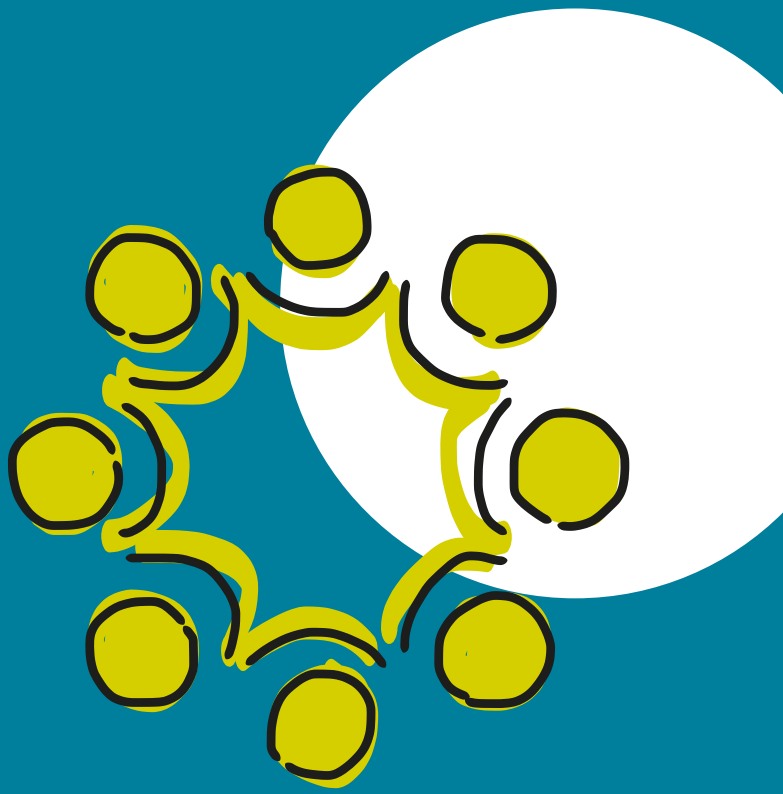
Al-Maamari, M.	RET-P01
Altmeyer, K.	ETS-P02
Baack, S.	UBP-A02
Bauernfeind, S.	WDA-A01
Baumann, F.	CRA-P01
Caspari, L.	RET-P01, WDA-A02
Christine Plote	ETS-P01
Ciz, D.	RET-P01
Dang, D. H.	RET-P01, WDA-A02
Decker, A.	UBP-P01
Devathasan, B.	WDA-P01
Dinzinger, M.	ETS-P01, RET-P01
Einsiedler, M.	UBP-P01
Erenli, K.	LEA-A02
Fathima, N. A.	RET-P02
Fenz, S.	WDA-A01
Fontana, L.	AIS-P01, AIS-P02
Frühwirth, R.	CRA-P02
Gabelmann, J.	ETS-P02
Geiß, C.	SPM-A01
Geminn, C.	LEA-A01
Ghosh Dastidar, K.	RET-P01
Gmelch, U.	ETS-P02, WDA-A02
Gottsmann, T.	ETS-P02
Götz, S.	UBP-A01
Granitzer, M.	ETS-P01, RET-P01, WDA-A02
Guijarro, J. M.	RET-P02
Gütl, C.	AIS-P01, AIS-P02, CRA-P02
Hänel, M.	UBP-A01
Hecking, T.	CRA-P02
Heel, S.	ETS-P02
Heineking, S.	ETS-P01, WDA-A02
Henriques, A.	AIS-P01, AIS-P02
Hu, X.	SPM-A02
Indran Rajanbabu, E.	WDA-P01
Jakovljevic, I.	AIS-P01, AIS-P02
Johannes, P. C.	LEA-A01, LEA-P02
Kalimullin, D.	WDA-A02
Kersten, J.	RET-A01, SPM-A02
Klan, F.	SPM-A02

Kühn, W.	SPM-P01
Lau, P.	AIS-A01
Malone, S.	ETS-P02
Meywirth, S.	LEA-A01
Mitra, B.	ETS-P03
Mitrović, J.	RET-P01, WDA-A02
Nagel, S.	CRA-A01
Neophytou, N.	ETS-P03
Nidhi	AIS-A01, WDA-P01
Niehaus, E.	SPM-P01
Nussbaumer, A.	CRA-P02, ETS-P01
Pack, S.	AIS-P01
Pandey, H.	RET-A01
Paris, G.	CRA-P01
Paris, M.	CRA-A01, CRA-P01
Platz, M.	ETS-P02
Plote, C.	ETS-P02
Pöchhacker, N.	ETS-P01
Potthast, M.	ETS-P01, WDA-A02
Rapp, J.	SPM-P01
Reese, K.	ETS-P02
Rehm, F.	RET-P02
Richter, S.	WDA-A02
Ridgway, R.	ETS-P01
Rinner, C.	WDA-A01
Rittlinger, V.	SPM-A01
Rodrigues Eismann, A.	ETS-P02
Sandner, E.	AIS-P01, AIS-P02
Scharf, D.	AIS-P02
Schick, L.	ETS-P02
Schuster, J.	SPM-P01
Sharikadze, M.	WDA-A02
Simniceanu, A.	AIS-P01, AIS-P02
Söllner, M.	LEA-A01, UBP-A01
Stein, L.	SPM-A02
Stoffel, M.	SPM-P01
Taubenböck, H.	SPM-A01
Voigt, S.	LEA-A01, SPM-A01, WDA-A02
Wagner, A.	AIS-P01, AIS-P02, RET-P02
Weber, F.	WDA-A01
Werther, C.	LEA-A01, LEA-P01
Wiesner, M.	ETS-P02
Wolf, V.	ETS-P02
Zeeb, B.	WDA-A02
Zerhoudi, S.	ETS-P01

Content from this work may be used under the terms of the CC-BY-ND 4.0 licence (© 2023). Any distribution of this work must maintain attribution to the author(s), title of the work, publisher, and DOI







open search foundation

CODE UNIVERSITY OF APPLIED SCIENCES

ISSN: 2957-4935
ISBN: 978-92-9083-744-2