Federated HPC Workflows for Large-Scale Web Indexing

Martin Golasowski[1], Jan Martinovič[1], Michael Dinzinger[2], Sebastian Heineking[3], Gijs Hendriksen[4], Michael Granitzer[2]

[1]IT4Innovations, VSB - TU Ostrava, Czech Republic, [2]University of Passau, Germany, [3]Leipzig University Germany, [4]Radboud University, Netherlands

Motivation & Context

The global web search ecosystem is dominated by a small number of commercial providers, leading to limited transparency, algorithmic bias, and reduced user sovereignty. Europe currently lacks a publicly accessible, independent web index that reflects its values of openness, fairness, and regulatory compliance. This dependency constrains academic research and limits competition in the search and data economy. An open alternative is essential to ensure accountability and long-term sustainability of web access.

OpenWebSearch.eu Vision

OpenWebSearch.eu (OWS) aims to address this gap by laying the foundations for a sustainable ecosystem of open web search and data services. The project supports research and innovation, enables the development of alternative search technologies, and strengthens European digital sovereignty through a publicly accessible Open Web Index. OWS promotes openness, interoperability, and reuse of web data across sectors. Its vision is to empower a diverse community of users and developers with transparent and trustworthy search infrastructure.

Results & data available

9+ billion URLs indexed

185 languages covered

35.11 TiB index size

>1 TiB/day ingestion rate

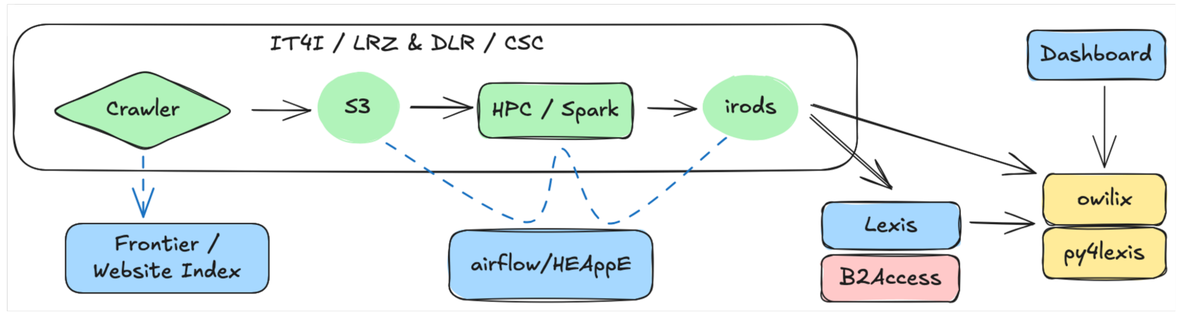
1051 datasets published

Access the OWI data at https://dashboard.ows.eu.

Technical Challenges

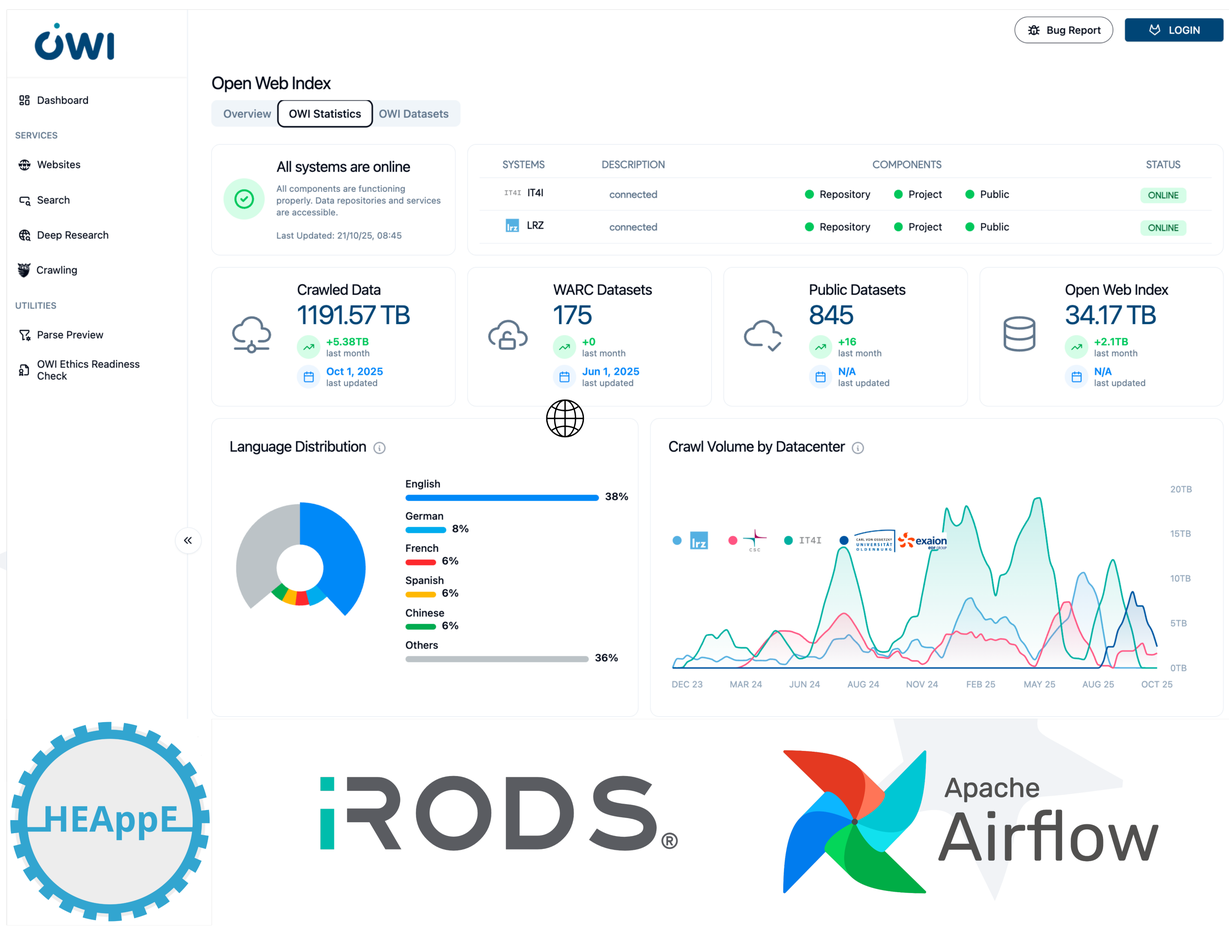
- Handle massive-scale web data with high compute and storage demands
- Support multilingual processing (Language detection, deduplication, and text extraction)
- Provide scalable, distributed infrastructure for computing and data
- Orchestrate complex data processing pipelines and data movement
- Leverage European HPC resources through the LEXIS Platform 2
- Enable independent resource providers to join the OWI infrastructure

Crawling & Processing Pipeline



The OWS processing pipeline consists of large-scale web crawling, preprocessing, index generation, and dataset publication. These workflows are executed across major European HPC centers, including IT4Innovations (CZ), Leibniz Supercomputing Centre (DE), and CSC (FI). Parallel execution enables high-throughput processing of web-scale data.

Open Web Index Dashboard



Conclusions & Outlook

- Demonstrates the feasibility of open, large-scale web indexing
- Advances transparent and reproducible search technologies
- Provides a blueprint for future data-intensive infrastructures
- Supplies web data to the LUMI AI Factory, enabling open-data-driven AI
- Enables semantic search and a human-centred open web ecosystem

References

Granitzer, M., Hayek, M., Heineking, S., Hendriksen, G., Golasowski, M., Dinzinger, M., & Zerhoudi, S. (2025). OpenWebSearch. eu-Building an Open Web Index on EuroHPC JU Infrastructures. Procedia Computer Science, 255, 43-52.

The project OpenWebSearch.EU has received funding from the European Union's Horizon research and innovation programme under grant agreement No 101070014.

Funded by the European Union

ows@openwebsearch.eu

https://openwebsearch.eu