Contribution ID: **232**

Type: **Presentation**

Using Federated Data Infrastructure for a European Open Web Index

Thursday 20 March 2025 17:15 (15 minutes)

In an era where web search serves as a cornerstone driving the global digital economy, an open, impartial and transparently produced web index is a key opportunity for Europe and beyond. Currently, the landscape is dominated by a select few gatekeepers who provide their web search services with minimal scrutiny from the general public. Moreover, web data has emerged as a pivotal element in the development of AI systems, particularly Large Language Models. The efficacy of these models depends upon both the quantity and quality of the data available. Consequently, restricted access to web data and search capabilities severely curtails the innovation potential, particularly for smaller innovators and researchers who lack the resources to manage petabyte platforms.

In this talk, we present the OpenWebSearch.eu project which is currently developing the core of a European Open Web Index (OWI) as a basis for a new Internet Search in Europe. We mainly focus on the setup of a Federated Data Infrastructure leveraging geographically distributed data and computing resources at top-tier supercomputing centres across Europe. This data infrastructure leverages MINIO/S3, iRODS, EUDAT (B2SAFE, B2HANDLE) and our previous work on the LEXIS Platform for distributed computing and data management. The system developed facilitates efficient execution of complex processing and indexing workflows.

Author: HAYEK, Mohamad

Co-authors: WAGNER, Andreas (CERN); MARTINOVIČ, Jan (VSB - Technical University of Ostrava); MANKINEN, Katja (CSC –IT Center for Science); GOLASOWSKI, Martin; SHARIKADZE, Megi; GRANITZER, Michael; Ms FATHIMA, Noor Afshan (CERN); ZERHOUDI, Saber (University of Passau); HEINEKING, Sebastian (Webis--Group); HACHINGER, Stephan (Leibniz Supercomputing Centre (LRZ) of the BAdW)

Presenter: HAYEK, Mohamad

Session Classification: Data sharing infrastrcutures

Track Classification: Main sessions: CS3 federations and synergies with eResearch infrastructures.