Contribution ID: **112** Type: **Short Talk**

Leveraging Federated Data Infrastructure for a European Open Web Index

Wednesday, 2 October 2024 12:00 (15 minutes)

In an era where web search serves as a cornerstone driving the global digital economy, the necessity for an impartial and transparent web index has reached unprecedented levels, not only in Europe but also worldwide. Presently, the landscape is dominated by a select few gatekeepers who provide their web search services with minimal scrutiny from the general populace. Moreover, web data has emerged as a pivotal element in the development of AI systems, particularly Large Language Models. The efficacy of these models is contingent upon both the quantity and calibre 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 compute resources at top-tier supercomputing centres across Europe. We then detail the use of the LEXIS platform to orchestrate and automate the execution of complex preprocessing and indexing of crawled data at each of the centres. We finally present the effort to adhere to the FAIR data principles and to make the data available to the general public.

Topic

Data innovations: Data Analytics, Sensitive Data/FAIR Data

Primary authors: Prof. GRANITZER, Michael (University of Passau); Mr HAYEK, Mohamad (Leibniz Supercomputing Centre)

Co-authors: Dr WAGNER, Andreas (CERN); Dr GOLASOWSKI, Martin (VSB - Technical University of Ostrava); Dr SHARIKADZE, Megi (Leibniz Supercomputing Centre); Mr DINZINGER, Michael (University of Passau); Ms FATHIMA, Noor Afshan (CERN); Mr ZERHOUDI, Saber (University of Passau); Mr MOIRAS, Stavros (CERN); Dr HACHINGER, Stephan (Leibniz Supercomputing Centre)

Presenter: Mr HAYEK, Mohamad (Leibniz Supercomputing Centre)

Session Classification: Maximizing Data Efficiency and Availablity: Innovations in Data Management Platforms