Open Search @ DLR - towards transparent access to web-based information in science

S. Voigt[1], German Aerospace Center, Oberpfaffenhofen, Germany
T. Hecking, German Aerospace Center, Köln, Germany
D. Jankowski, OFFIS Institute for Computer Science, Oldenburg, Germany
J. Möller, University of Oldenburg, Germany
M. Schwinger, German Aerospace Center, Oberpfaffenhofen, Germany

Abstract

Data is the raw material of the 21st century - for research, innovation, economy and society. Digital sovereignty requires free, uninfluenced & traceable access to information - in other words, open Internet search and systematic access to web data. Currently, there is a monopoly in information search: In Europe, more than 90% of all Internet searches are conducted via a single commercial and advertising-optimized search engine. This holds immense potential for intentional or unintentional manipulation in access to data, information, technology and knowledge (cognitive/economic bias). Especially for science, new concepts for a distributed and open Internet search infrastructure are needed.

The wealth of data and information on the web must be rendered more accessible through uninfluenced discovery of scientific data and information, since it is the basis for free research and innovation. Against this background, the German Aerospace Center (DLR) is contributing to the European Open Search Initiative, formed by science and computing centres. Within the Open Search @ DLR project, existing in-house capacities and know-how in data access and search are identified and pooled to set-up a cooperative crawling, indexing and search capability to web data repositories – internal and external to DLR. Furthermore, dedicated pilot applications in areas such as information retrieval, knowledge management or information evaluation and transparency, making use of the infrastructure, are developed in the project.

A primary focus of the Open Search @ DLR project is networking of in-house expertise as well as connecting with the Europe-wide Open Search Initiative.

Within this talk we present the project layout and findings during the first project phase. This includes inventorying of in-house data and heterogeneous information repositories, coordinated crawling, indexing and searching. We present architecture and set-up of a testbed for cooperative crawling, where single crawling nodes communicate URLs to crawl in a peer-to-peer fashion as basis for joint assembly of large corpora of web data.

In a second part of the talk scientific pilot applications of an open search infrastructure are discussed, including the use of georeferenced data from web- and database sources, e.g. for monitoring of news, events, geospatial analysis and early warning. Furthermore, open search approaches for exploring, linking, and indexing of information from heterogeneous scientific data sources and public web content are particularly being addressed. This includes access to (semi-)structured information in databases as well as information extraction from texts, e.g. automatic geo-tagging. In this context, especially the establishment of geographical connections between scientific, structured databases and human-readable content from the Internet play an important role.

In the last part of the talk first ideas and concepts for a long-term activity of science and computing centres to set up an open Internet search ecosystem are discussed. Such a shared activity should be based on cooperative computing, open-source software stacks and public moderation and should involve distributed scientific high-performance computing and cloud facilities forming a cooperative open search infrastructure to warrant a long-term, public and open web search environment.

As long as the digital sphere – the web – exists, free and unbiased orientation therein has to be ensured to guarantee free and unbiased access to information for science, economy and society as a whole.

1 stefan.voigt@dlr.de