GoEarth - performance and challenges of a semantic search engine for CCS literature

B. D. Gutknecht (1) and M. R. Alvers (2)

(1) Christian-Albrechts-Universität zu Kiel, Institut für Geowissenschaften, Geophysik, Kiel, Germany
(benjamin@geophysik.uni-kiel.de), (2) Transinsight GmbH, Dresden, Germany

Increased interests in climate sciences have led to exponentially increasing publication numbers regarding Carbon Capture and Storage (CCS). In order to be able to handle this overload of information, an ontology-based, semantic search system was developed. The used ontology was especially designed for geo-scientific content in CCS-literature.

The semantic search engine provides users with an infrastructure that semantically categorizes textual content by facilitating natural language processing for the knowledge domain of CCS. The system works fully automated. So even search without keywords becomes possible and documents can be identified by navigating the ontology which is rendered as a tree on the left hand side of the search results. Furthermore, meta-data analysis provides useful statistical information, e.g. about author networks.

Compared to the 1045 AGU index terms and the 3587 GeoRef thesaurus terms, the new CCS ontology (1170 concepts and synonyms) links to 11.0 % and 8.6 % more documents of a geo-scientific text corpus, respectively. Of all test documents, the CCS ontology hits 96.5 % (GeoRef: 88.9 %, AGU index terms: 87.1 %). We will discuss existing challenges like disambiguation of concepts and intelligent text mining.