Geophysical Research Abstracts
Vol. 20, EGU2018-1915, 2018
EGU General Assembly 2018

PANGAEA Data Recommender: An Out of the Box Approach to Discover Earth and Environmental Datasets.

Anusuriya Devaraju (1), Uwe Schindler (2), Michael Diepenbroek (2), Jens Klump (1), and Robert Huber (2)

(1) CSIRO Mineral Resources, Australia (anusuriya.devaraju@csiro.au), (2) Center for Marine Environmental Sciences (MARUM), University of Bremen, Germany

In Earth and Environmental Sciences, various research infrastructures such as the Australian CSIRO or those of the European ENVRIplus cluster have been set up to gather and disseminate open scientific data. Examples of scientific data are facts collected through observation, measurement, simulation and modelling, experiments, or derived data products. Most of the research data portals offer keyword and faceted search. Users face several challenges when trying to discover datasets through the search tools on the portals, for example, an overload of choice among very many similar datasets, and the limitations of existing data search tools. The search tools may produce either empty search results or too many almost identical datasets. As a result, users need to spend considerable effort to find data sets that are most relevant for their applications. In addition, they need to be familiar with the structure and terminology of the portals. Text-based search may return relevant or irrelevant results, depending on how well the users articulate their data queries, and how well the authors described their data sets. Textual queries and ranking algorithms employed by the data portals may be ineffective for users looking for a diversity of data, from different sources and domains to conduct cross disciplinary studies. In most data portals, the top-ranked results may all come from the same bulk data collection, and therefore users are unlikely to discover interesting new data sets. The above challenges reflect that we need a data discovery solution that complements the existing search functionalities on the portals, which delivers relevant and novel datasets to users. A recommender system is an information filtering system that presents users with the relevant and interesting contents based on users' context and preferences. We argue that delivering data recommendations to users in a data portal can improve data discovery, and as a result, may enhance user engagement with the portal. This poster describes the design and development of a data recommender system for PANGAEA, a data publisher for earth and environmental sciences; it is managed by the Center for Marine Environmental Sciences (MARUM) at University of Bremen and the Helmholtz Center for Polar and Marine Research (AWI). PANGAEA currently hosts more than 370,000 datasets from various disciplines and used as long term data archive for several research data infrastructures. We developed a recommendation model to utilize the metadata of the datasets, and the user interactions (e.g., data search and download patterns) extracted from the PANGAEA server to produce and score data recommendations. We demonstrate how to leverage the functionalities provided by the PANGAEA's search engine (Elasticsearch) to deliver data recommendations from large collections to users.