Geophysical Research Abstracts
Vol. 21, EGU2019-4389, 2019
EGU General Assembly 2019

FAIRfication of PANGAEA datasets: Recent Developments and Lessons Learned

Anusuriya Devaraju, Uwe Schindler, Tina Dohna, Robert huber, Michael Diepenbroek, and Ketil Koop-Jakobsen
University Bremen, MARUM, Germany (adevaraju@marum.de)

PANGAEA is a data publisher for Earth and Environmental Science, operating for more than two decades. It is currently managed by the Center for Marine Environmental Sciences (MARUM) and the Alfred Wegener Institute, Helmholtz Center for Polar and Marine Research (AWI). To date, the PANGAEA system has published more than 380,000 scientific datasets from projects, cruises, institutions, and individual scientists. The datasets are identified with Digital Object Identifiers (DOIs) and are accompanied by bibliographic citations. We present the recent and ongoing developments to promote the FAIRification of PANGAEA datasets. Specifically, we discuss the practical insights and lessons learned on implementing the following applications:
(a) Representing data with rich metadata to capture the links between datasets and their related research objects: We specified the relations between datasets and related research objects (e.g., specimens, authors, instruments) as part of the dataset metadata, and make them accessible through the PANGAEA data portal. The related objects are identified with relevant persistent identifiers (e.g., IGSN, ORCID, Handles, INSDC accession numbers). The persistent identifiers ensure the accessibility of objects metadata over time.
(b) Improving data findability and discovery for users and machines: We developed a data recommender system, which enables data users to discover similar and novel scientific datasets on the PANGAEA portal. We embedded structured metadata on data pages by adopting Schema.org such that datasets are discoverable through Google and other search engines.
(c) Promoting accessibility to remote data through computational notebooks: Metadata of datasets are accessible through different means (i.e. common schemas and standardized communications protocol). To enhance data accessibility, we recently prototyped a python package that allows scientists to easily find, access and analyze PANGAEA datasets within a Jupyter notebook.
(d) Facilitating data interoperability through formal representations: To support formal and meaningful descriptions of research datasets, we developed a semi-automatic annotation approach to enrich the data header labels (e.g., observable variables, features) with common vocabularies defined ontologies and taxonomies.
(e) Ensuring meta(data) accessibility and re-usability through access rights: We released the research datasets through a common licensing term (e.g., Creative Commons Attribution License). To support data owner requirements, we ensured a restricted access to certain datasets through embargo periods.