EGU22-11102
https://doi.org/10.5194/egusphere-egu22-11102
EGU General Assembly 2022

Joining Geo Data Across Different Providers to Ease Machine Learning Applications

Matthes Rieke, **Benedikt Gräler**, and Simon Jirka
52°North GmbH, Münster, Germany

Data integration and harmonization has been a tedious task ever since. The increase of available data in volume and variety has further increased the need for a thorough data integration. Furthermore, the application of more and more automatic algorithms stresses the need for a sensible geo data platform to avoid the 'garbage in, garbage out' trap and to allow for a meaningful data analysis. We reviewed different projects and learned about various needs and constraints of joint spatial research data infrastructures from local to cloud based deployments. Typically, these systems are not designed from scratch and existing systems need to be integrated or interfaced. As a result or arising from the need to support the sovereignty of distributed data centers, modern infrastructures need to be capable to support federated set-ups. Often these research data infrastructures shall not only be used to store raw data for scientists, but will also provide results (maps, derived data products, tools and applications) to the public. This goes along with the need for access delegation (e.g. OAuth). A special focus is put on the provision of the joint datasets for machine learning applications. In order to facilitate efficient learning and prediction a ML processing environment needs to be aligned with the data infrastructure.

We will present commonalities among these infrastructures and outline typical design patterns. A spatial data infrastructure based on open source software components that can be deployed on the cloud will be introduced. It features open standardized interfaces and services for easy adaptation and connectivity.