Geophysical Research Abstracts
Vol. 19, EGU2017-13191, 2017
EGU General Assembly 2017

What does it take to build a medium scale scientific cloud to process significant amounts of Earth observation data?

André Hollstein, Hannes Diedrich, and Daniel Spengler
Helmholtz-Zentrum Potsdam Deutsches GeoForschungsZentrum GFZ

The installment of the operational fleet of Sentinels by Copernicus offers an unprecedented influx of freely available Earth Observation data with Sentinel-2 being a great example. It offers a broad range of land applications due to its high spatial sampling from 10 m to 20 m and its multi-spectral imaging capabilities with 13 spectral bands. The open access policy allows unrestricted use by everybody and provides data downloads for on the respective sites. For a small area of interest and shorter time series, data processing, and exploitation can easily be done manually. However, for multi-temporal analysis of larger areas, the data size can quickly increase such that it is not manageable in practice on a personal computer which leads to an increasing interest in central data exploitation platforms. Prominent examples are GoogleEarth Engine, NASA Earth Exchange (NEX) or current developments such as CODE-DE in Germany. Open standards are still evolving, and the choice of a platform may create lock-in scenarios and a situation where scientists are not anymore in full control of all aspects of their analysis. Securing intellectual properties of researchers can become a major issue in the future.

Partnering with a startup company that is dedicated to providing tools for farm management and precision farming, GFZ builds a small-scale science cloud named GTS^2 for processing and distribution of Sentinel-2 data. The service includes a sophisticated atmospheric correction algorithm, spatial co-registration of time series data, as well as a web API for data distribution. This approach is different from the drag to centralized research using infrastructures controlled by others. By keeping the full licensing rights, it allows developing new business models independent from the initially chosen processing provider.

Currently, data is held for the greater German area but is extendable to larger areas on short notice due to a scalable distributed network file system. For a given area of interest, band and time range selection, the API returns only the data that was requested in a fast manner and thereby saves storage space on the user's machine. A jupyterhub instance is a main tool for data exploitation by our users. Nearly all used software is open source, is based on open standards, and allows to transfer software to other infrastructures.

In the talk, we give an overview of the current status of the project and the service, but also want to share our experience with its development.