Pilot for accessing distributed marine data using the OneData concept

Peter Thijsse, MARIS, Nootdorp (The Netherlands), peter@maris.nl
Dick Schaap, MARIS, Nootdorp, (The Netherlands), dick@maris.nl
Gergely Sipos, EGI Foundation, Amsterdam (The Netherlands), gergely.sipos@egi.eu

There are many challenges, also in the SeaDataNet community, in the domain of access to distributed data. A large portion of the challenges is related to the various aspects of data storage management. For example, having access to files locally available, synchronous file transfers, proprietary log formats, and log locations. OneData (adopted by EGI) offers many advanced features related to data storage management. Therefore, MARIS has spent effort as part of the EOSC-HUB project on working out several use cases with the OneData concept. As users two different groups can be distinguished: Developers facing the challenge with more and more, and larger datafiles. End-users benefitting from the faster response times and more direct access to large data collections.

About Onedata

With Onedata, users can access, store, process and publish data using a global data storage backend provided by computing centers and storage providers worldwide. Onedata focuses on instant, transparent access to distributed data sets, without unnecessary staging and migration, allowing access to the data directly from your local computer or worker node. The most important concepts of the Onedata platform are:

- Spaces - distributed virtual volumes, where users can organize their data
- Providers - entities who support user spaces with actual storage resources exposed via Oneprovider services
- Zones - federations of providers, which enable creation of closed or interconnected communities, managed by Onezone services.

More background information of Onedata can be found at www.onedata.org. For the pilot these concepts have been used, configured and investigated.

SeaDataNet practice

In the SeaDataNet practice (just like in other infrastructures) developers are confronted many times with situations that data sets are stored at different locations while we want to undertake central processing. For instance, there is great interest in so-called BioGeoChemical (BGC) data sets as these provide input for determining indicators about the quality of marine waters and as such are very relevant for the Marine Strategy Framework Directive of the EU which aims at establishing Good Environmental Status (GES). Through its engagement with EMODnet Chemistry, SeaDataNet is actively supporting Regional Sea Conventions, EU DG Environment, and European Environment Agency (EEA) in compiling and providing harmonised and validated data collections for eutrophication and contaminants which are derived from the BGC data as gathered by the SeaDataNet data centres. Moreover, SeaDataNet has established cooperations with Copernicus CMEMS as well as with Euro-Argo to work together on mutual data exchanges and on improving and innovating quality control and processing of large BGC data collections for various purposes, including MSFD. Access to the data files, as well as controlling quality and processing the distributed datasets, currently have performance issues and existing solutions might have to be replaced at some point with new concepts.

Pilot content
In order to investigate the potential of Onedata solutions for SeaDataNet purposes, during 2020 a test configuration has been set up using OneData in combination with Cassandra and Elasticsearch. OneData has been configured to give access to a Onedata “Space” to data of a number of data providers on the cloud, each provided with BGC data collections in the SeaDataNet ODV format. Data is collected via a OneData “Zone”, cached and stored in a Cassandra open source NoSQL database with wide column store, which allows high searching performance on large data sets with many numbers. Elasticsearch has been configured on top in order to optimize free text search on the metadata of the data sets to facilitate fast and precise subsetting of data collections from the master collection.

During the session we will present the conclusions with key insights, examples how this could be used and user perspectives, as gained during development.