The SeaDataCloud Virtual Research Environment: researching the sea from the cloud

Merret Buurman, Deutsches Klimarechenzentrum GmbH (Germany), buurman@dkrz.de
Peter Thijsse, Mariene Informatie Service 'MARIS' Bv (The Netherlands), peter@maris.nl
Sri Harsha Vathsavayi, Tieteen Tietotekniikan Keskus OY (Finland), sriharsha.vathsavayi@csc.fi
Sebastian Mieruch, Alfred Wegener Institut (Germany), sebastian.mieruch@awi.de
Gael Leblan, Institut Français de Recherche pour l'Exploitation de la Mer (France), Gael.Leblan@ifremer.fr
Giorgio Santinelli, Deltares (The Netherlands), Giorgio.Santinelli@deltares.nl
Alexander Barth, University of Liege (Belgium), a.barth@uliege.be , *et al.*

The SeaDataNet project offers a robust and state-of-the-art Pan-European infrastructure to harmonise metadata and data from marine data centres in Europe, and offers the technology to make these data accessible. The user was able to use the SeaDataNet infrastructure to download the data to local servers or local machines and work from there with the data there. As marine observation data continues to increase in size and number of datasets, transferring it over the network and processing it requires more and more efficient machines and network bandwidth. Downloading a large amount of large data sets, processing it on a laptop, and uploading the results to send them to colleagues, then receiving feedback and re-initiating the entire process, is cumbersome.

To make life easier for users and data holders, as part of the SeaDataCloud project, SeaDataNet is moving its unrestricted data to the cloud. Keeping the data centrally in high-performance data centres includes some other advantages than ease of access and download: The centralized ingestion process allows for standardized quality checks to be performed on the data before it is made available to users. Corrupt files, non-conforming formats, and duplicate observations which could lead to bias/artefacts in analyses, can be more easily detected. Loss of data is precluded by storing the data in several locations across Europe. Five major EUDAT data centres will form a cloud (DKRZ in Germany, CINECA in Italy, CSC in Finland, GRNET in Greece, and STFC in the United Kingdom) that hosts copies of the SeaDataNet datasets for highly available download.

But why only move the stored data to the cloud? Another aspect of working with large amounts of data is the resources needed for processing them. In general, the current trend is less downloading, more processing where the data is. While marine observation data continues to increase in size and number of datasets, transferring it over the network and processing it requires more and more efficient machines and network bandwidth. Downloading a large amount of large data sets, processing it on a laptop, and uploading the results to send them to colleagues, then receiving feedback and re-initiating the entire process, is cumbersome. And then, after all analyses are run (phew!), you see that a new version of the dataset had been published, correcting some important shortcoming!

To tackle these problems, the creation of the SeaDataNet Virtual Research Environment (VRE) is being developed in close collaboration of SeaDataNet developers and the 5 partners of

European Research Data Infrastructure EUDAT.

Virtual Research environments are web-based workspaces providing seamless access to all services researchers need to do their work and collaborate with their community. For SeaDataNet, this means that all the tasks that a researcher would usually do with the data, the entire workflow of data-driven science - finding data, accessing data, processing iteratively data with various tools, visualizing results, sharing results with colleagues, and publishing data - can be realized without having to download data to the desktop and using the local compute power, which might not be available to all users at the same rate.

The first use case that is being addressed is that of the SeaDataNet expert groups creating temperature and salinity climatologies products. Data from many cruises/stations/projects is aggregated to a global dataset, which is being reviewed manually to ensure its scientific quality, using the webODV tool. Then, geospatial interpolation algorithms are used to create a continuous field of temperature and salinity using the DIVA tool (Barth *et al.* 2014). These climatologies, separated by basin, are available on the Sextant catalogue for SDN products https://www.seadatanet.org/Products.

As part of the first VRE release this use case will be integrated and tested by the user groups. This use case represents a core workflow in the VRE and will be used as a template for enabling more use cases (e.g. EMODNET Bathymetry DTM processing, EMODNET Chemistry, bio-geo-chemistry QC). Of course the scientists are not restricted to predefined workflows, but encouraged to use and recombine the VRE's functionalities for custom analysis workflows.

The most common tools that the SeaDataNet community is using, notably webODV (Ocean Data View - online) and DIVA using Jupyter Notebook, are already included in the first prototype. Thanks to the extensible and extendable architecture (described in a second abstract), many more are to come. To serve expert users as well as more applied users, most services will come with an easy-to-use GUI as well as a command line interface. On top of this, a chat/forum-like communication channel for scientists will encourage more collaborative work style. A notifier mechanism, that warns users of new versions of data they are using, will prevent the use of outdated datasets. Finally, chaining several of the VRE's processing tools to an entire workflow that can easily be documented, rerun and reproduced is planned. In addition to the SDN CDI data from the SDN data cloud, other datasets relevant to the marine community will be available, for example SDN products and EMODNet data products.

As a sum-up, we can say that beyond easing the problem of resources, several other advantages are achieved by the VRE: The use of outdated data and software is prevented, slow and unnecessary downloads are reduced and jam-packed hard drives are avoided. Processing of data gets faster and multiple tasks can easily be run in parallel. Sharing input data, intermediate results or end results becomes easier. All this, facilitates collaborative science in a scientific world that relies more and more on team work, by internationally distributed teams, and open science.

Stay tuned and check out https://vre.seadatanet.org for updates on the progress of SeaDataNet's VRE!

References

Barth, A., Beckers, J.-M., Troupin, C., Alvera-Azcárate, A., and Vandenbulcke, L.: *divand-1.0: n-dimensional variational data analysis for ocean observations,* Geosci. Model Dev., 7, 225-241, doi:10.5194/gmd-7-225-2014, 2014.