PHIDIAS: boosting the use of cloud services to benefit marine data management, services and processing

Charles Troupin, GHER/ULiège (Belgium), ctroupin@uliege.be
Julie Abergas-Arteza, Trust-IT Services, Pisa (Italy), j.arteza@trust-itservices.com
Alexander Barth, GHER/ULiège (Belgium), a.barth@uliege.be
Frederic Huynh, IRD (France), frederic.huynh@ird.fr
Jukka Seppälä, SYKE (Finland), jukka.seppala@ymparisto.fi
Seppo Kaitala, SYKE (Finland), Seppo.Kaitala@ymparisto.fi
Aleksi Kallio, CSC (Finland), aleksi.kallio@csc.fi
Cecile Nys, IFREMER, Plouzané (France), Cecile.Nys@ifremer.fr
Francesco Osimanti, Trust-IT Services, Pisa (Italy), f.osimanti@trust-itservices.com
Peter Thijsse, MARIS (The Netherlands), peter@maris.nl
Gilbert Maudire, IFREMER, Plouzané (France), Gilbert.Maudire@ifremer.fr
Boris Dintrans, CINES, Montpellier (France), boris.dintrans@cines.fr

The project and its goals

PHIDIAS - Prototype of HPC/Data Infrastructure for On-demand Services - addresses the development and concrete realisation of a set of High-Performance Computing (HPC) based on interdisciplinary services and tools to exploit large satellite datasets of public European interest provided by:

- Satellite observation of Earth,
- FAIR access to processed datasets,
- Value-added services through large data storage capability and high-bandwidth networks across Europe.

PHIDIAS aims at creating access services to increase the HPC and data capacities of the European Data Infrastructure in the context of the 'Connecting European Framework (CEF)' on open data. The PHIDIAS project seeks for – and establishes – links with the scientific, engineering and industrial communities that may be interested in the project's results and outcomes, in order to initiate and fulfil high-level exchanges at different levels.

The Ocean use case

Observing the ocean is challenging: missions at sea are costly, different scales of processes interact, and the conditions are constantly changing. This is why scientists say that "*a measurement not made today is lost forever*". For these reasons, it is fundamental to properly store both the data and metadata, so that access to them can be guaranteed for the widest community, in line with the FAIR principles: Findable, Accessible, Interoperable and Reusable. The PHIDIAS Ocean use case is focused on 3 aspects:

1. **Improvement of long-term stewardship of marine *in situ* data**. The SEANOE service allows users to upload, archive and publish their data, including the processed data via HPC, to which a permanent identifier (DOI) is assigned so the dataset can be cited and referenced. Efforts will be articulated around the scalability, the exchanges between data centres in charge of related data types and the protection of long-time archives. The *long-tail* data, referring to data not acquired

routinely but during scientific missions or specific events, are of particular interest. Usually they cannot make their way so easily to the data centers, as there is no automated procedure.

2. **Improvement of data storage for services to users**. The goal is to provide users with (1) fast and interoperable access to data from multiple sources, for visualization and subsetting purposes; (2) parallel processing capabilities within dedicated high-performance computing, using, for example, Jupyter notebooks or the PANGEO software ecosystem.

3. **Marine data processing workflows for on-demand processing**. The objective is that users can access data, software tools and computing resources in a seamless way to create added-value products, for example quality-controlled, merged datasets or gridded fields.

Key results

The efforts of the pilot use cases in PHIDIAS strive to improve the activities of the researchers and specialists on different aspects:

- **Data publication:** with the improved capabilities of SEANOE, they will be able to seamlessly upload large datasets, ensure their long-term archiving (also of the processed datasets) and publish them following standards, best practices and recommendations from Data Management groups. This will also enhance the ingestion of *long-tail* data, which in turn will be made available to a larger community.

- **Data access:** thanks to fast access to the most recent data collections obtained from different sources and providers (Euro-ARGO, SeaDataNet, EMODnet, CMEMS, imaging flow cytometer), users will be able to demonstrate as pilot operations such as sub-setting (based on regions, parameters), quality-control, visualisation or spatial interpolation.

- **Data processing:** the deployment of cutting-edge tools such as DIVAnd (spatial-temporal interpolation, https://github.com/gher-ulg/DIVAnd.jl) in an HPC environment will allow scientists and experts to perform spatio-temporal interpolation of large datasets. In particular, this use case will be in the North Atlantic Ocean and the Baltic Sea, which represents 10 million observations for a total of approx. 250 GBytes. The final product will consist of an inter-comparison of satellite data and in-situ data of sea surface salinity, including Inspire-compliant online services for data visualization and access.

Chlorophyll concentration (g/m³)