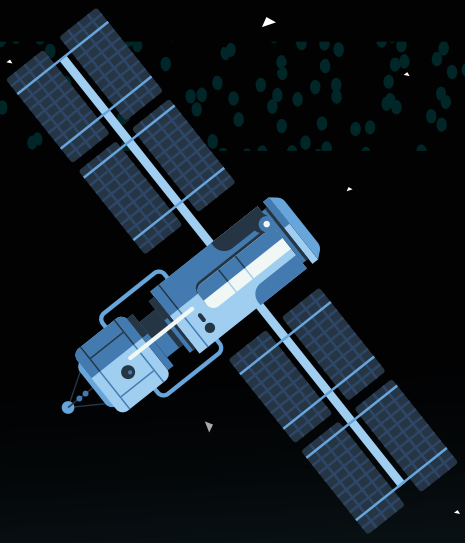
PHIDIAS

Prototype of HPC/Data Infrastructure for On-demand Services

earth observation

High Performance Computing

science

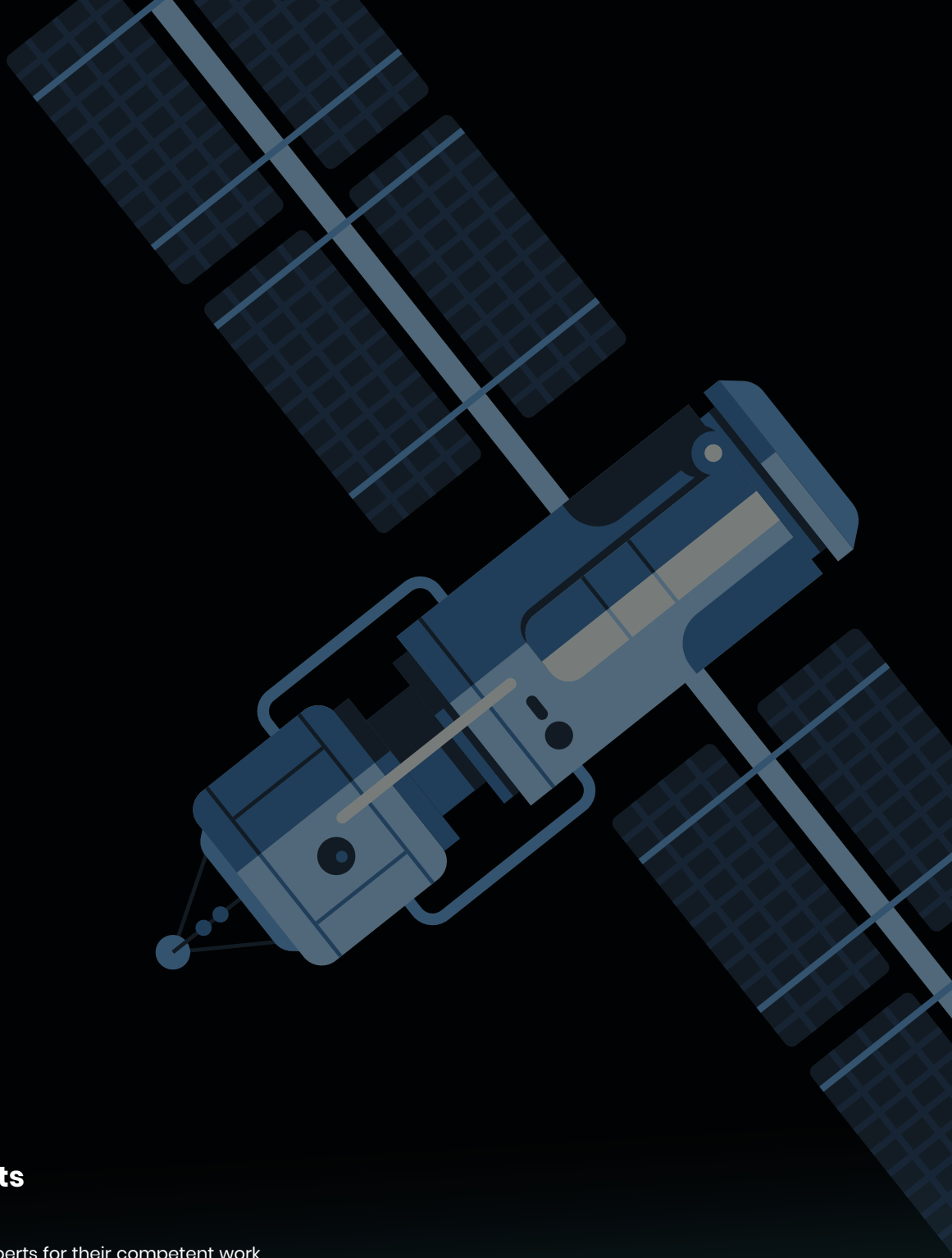


satellite data

ocean

PHIDIAS: creating access services to increase the High-Performance Computing (HPC) & Data capacities

Reader's Digest
July 2020



Acknowledgements

PHIDIAS is grateful to all the experts for their competent work and contributions to shape this document.

We would also like to show our gratitude for their efforts to the following individuals: Boris Dintrans (CINES, PHIDIAS Project Coordinator), Marion Lepaytre (CINES, PHIDIAS), Pascal Prunet (SPASCIA), Jean-Christophe Desconnets (IRD), Gilbert Maudire (Ifremer), Charles Truopin (Université de Liège), Francesco Osimanti (Trust-IT Services), Lorenzo Calamai (Trust-IT Services) for the Graphic support.

To PHIDIAS use cases participating partners such as CNRS, HYGEOS, ICARE and SRON for the Use Case 1 (Intelligent screening of large amount of satellite data for detection and identification of anomalous atmospheric composition events), GEOMATYS for the Use Case 2 (Big data earth observations: processing on-demand services for environmental monitoring), and the Université de Liège, Maris, CNRS, CSC and the Finnish Environment Institute for Use Case 3 (Ocean).

Table of contents

PHIDIAS Consortium

Disclaimer

The PHIDIAS - Creating Access services to increase the HPC and Data capacities was produced by the PHIDIAS project CEF GA no. INEA/CEF/ICT/ A2018/1810854.

PHIDIAS - Prototype of HPC/Data Infrastructure for On-demand Services is an H2020 project co-financed by the Innovation and Networks Executive Agency (INEA) under the European Union's Connecting Europe Facility (CEF).

Introduction

Intro-duction

In the digital era, High Performance Computing (HPC), also known as supercomputing, is at the core of major advances and innovation and a strategic resource for Europe's future. Indeed, in April 2016, the European Commission set the objective of providing researchers, industry, SMEs and public authorities with access to world-class supercomputers, unleashing their innovation and transformation potential. The ultimate goal is to place Europe among the current leaders in these fields, and to develop further the Digital Single Market in Europe.

The PHIDIAS project, funded by the European Union's Connecting Europe Facility (CEF), is built within this framework, aiming to become a reference point for the Earth science community enabling them to discover, manage and process spatial and environmental data, through the development of a set of High-Performance Computing (HPC) based services and tools exploiting large satellite datasets. In order to maximize its expected outcome, PHIDIAS will explore a distributed model for data transfer and resource allocation between two European computing centres: CINES in France and CSC in Finland.

The project foresees the development of three Use Cases:

- **Intelligent screening of large amount of satellite data for detection and identification of anomalous atmospheric composition events;**
- **Processing on-demand services for environmental monitoring; and**
- **Improving the use of cloud services for marine data management.**

Within the use cases, the project will develop data post-processing methods coupled with HPC capabilities, which will be deployed as services for several end-users, including scientific communities, public authorities, private players, and citizen scientists. In line with the European strategy for Open Science, the data generated and services created will be available on relevant EU portals, such as EU Open Data Portal, EUDAT, and EOSC, and will be preserved using the long-term preservation services of the EOSC.

Among the impacts that PHIDIAS expects to achieve, there is the creation of sustainable HPC data-powered services for the earth, atmospheric and marine data which researchers, industry and public sectors could benefit from. At the same time the project will work towards leveraging networking infrastructures to ensure end-to-end scientific workflows; federating infrastructure to infrastructure services, including authentication and access to resources (pre- and postprocessing, management and preservation of large volumes of digital information over time); and creating a FAIR portal for the scientific community and data providers.

Being the Project Coordinator, and with the efficient support of the consortium, I am working to ensure that the PHIDIAS objectives mentioned above are being consistently pursued, in order to deliver a catalogue that will implement interoperable services for the discovery, access and processing of data, guaranteeing the largest degree of reusability of data as possible, and the improvement of the FAIRisation of satellite and environmental datasets. Essentially, paving the way and making life easier for the next generation of HPC and the Computational Scientific community.

Boris Dintrans
Director of CINES and Phidias HPC Project Coordinator

Our Goals

Our Goals

GOAL 1: BUILDING A PROTOTYPE

Develop a catalogue that will allow users to discover and access data, open-source software, public Application Programming Interfaces (APIs) and interactive processing services. This catalogue will implement interoperable services for the discovery, access and processing of the data, and be connected to other major data repositories such as the European Data Portal, the Global Earth Observation System of Systems Portal (GEOSS), the Next Global Earth Observation System of Systems Portal (NextGEOSS), and the European Open Science Cloud (EOSC).

GOAL 2: OPTIMISING & INDUSTRIALISING

Optimise and industrialise workflows to allow the largest degree of reusability of data as possible, in compliance with the INSPIRE directives and ensuring interoperability with the EUDAT (European Data), EOSC, and IS-ENES (Infrastructure for the European Network for Earth System Modelling) portals and generic services.

GOAL 3: OPEN ACCESS

Implement an end-user web common interactive processing service based on notebook and data cube technologies allowing new users to easily have access to HPC capacities and develop new algorithms.

GOAL 4: FAIRISATION

Improve the FAIRisation of satellite and environmental datasets and preserve FAIR (Findable Accessible Interoperable Reusable) datasets in a Remote Data Access (RDA) certified repository.

GOAL 5: DATA PRE-PROCESSING

Develop new data pre-processing models coupled with HPC capabilities, by building a new and innovative on-the-fly computing service for smart processing of data, addressing the problem of efficient filtering and on-the-fly first diagnostics on incoming in-situ data.

GOAL 6: DATA POST-PROCESSING

Deploy data post-processing methods as a service for several end-users, including scientific communities, public authorities, private entities and citizen scientists.

Impact on the HPC community

Each data area test case that will be developed within the PHIDIAS project is expected to produce a number of concrete outputs, that are listed below:

Data area test case 1
Intelligent screening of large amount of satellite data for detection and identification of anomalous atmospheric composition events

- Creation of a prototype processing tool, and testing of datasets for non-regression
- Elaboration of a workflow specification and product specification document
- Contribution to a Common Information Model and Bêta version of a User Guide, describing the metadata and interface for pilot users
- Creation of Sentinel 5 precursor services
- Creation of a specification document for monitoring and alert services for Use case 1, and the identification of requirements for real time exploitation, procession and archiving
- Creation of a specification document for on-demand processing and identifying requirements for scientific exploitation for Use case 2
- Elaboration of a document for the specification, dimensioning size and requirements for the storage and processing of large amounts of future atmospheric Sentinel data

Data area test case 2
Big data earth observations: processing on-demand services for environmental monitoring

- Optimization and development of THEIA data processing chains in HPC environment
- Specification and implementation of data processing chains for massive execution and on-demand execution
- Specification and implementation of a UI web environment for on-demand processing
- Specification and implementation of data workflows for the discovery and development of EO raw data and products

Data area test case 3
Ocean

- Specifications for long-term data archiving procedures with respect to Remote Data Access (RDA) recommendations
- Elaboration of specifications and development of the tools to ensure that data and metadata are ready for long-term archiving
- Elaboration of specification for data storage
- Storing and archiving of large samples of SeaDataCloud in-situ data and of Satellite Sea Surface Salinity (SMOS satellite)
- Elaboration of specifications for DIVA hosting on the cloud
- Making on-demand in-situ data processing using DIVA available to users
- Report on case studies and conclusions of inter-comparison of satellite data and in-situ data for sea surface salinity including Inspire-compliant online services for data visualization

HPC community

PHIDIAS Use Cases

PHIDIAS Use Cases

Through the development of 3 Use Cases, Phidias will develop data post-processing methods coupled with HPC capabilities, which will be deployed as a service for several end-users (including scientific communities, Public authorities, private players, citizen scientists).

Use Case 1

Intelligent screening of large amount of satellite data for detection and identification of anomalous atmospheric composition events

Environmental satellite data are gradually becoming key in helping build an understanding of the changing environment and a new engine of scientific discovery. The intelligent screening of a large amount of satellite data is associated with the Earth sciences, data-derived and integrated from Earth observation, as well as terrestrial, oceanic and atmospheric data, together with data on human activity, obtained from ground sensor networks and other sources.

Goals

The data area testing case titled "Intelligent screening of a large amount of satellite data for detection and identification of anomalous atmospheric composition events" aims to use HPC and high-performance data management services for the development of intelligent screening approaches for the exploitation of large amounts of satellite atmospheric data in an operational context, implementing a prototype service on the already available Sentinel 5
Precursor (S5P) European atmospheric sounding mission.
The corresponding activities are mainly led by SPASCIA, together with its partners:

- ICARE (Sentinel 5 Precursor, Data access, data provider, formatting ad architecture),
- HYGEOS (Level 1 data processing for extreme events),
- SRON (Expertise on S5P data and products, Scientific pilot user).

Latest developments

Work has been initiated by SPASCIA, in coordination with ICARE, to select and obtain a sample of the S5P L2 product, and to start the analysis of the corresponding images. This allows the preparation of necessary inputs or test data for the development of the processing algorithm of the L2 S5P products (2nd L2 intelligent screening approach of S5P data, for the detection and characterisation of pollutant plumes). ICARE has implemented and is processing the archiving of S5P data and products on the ICARE facilities for PHIDIAS. Several exchanges between SPASCIA and ICARE have allowed the organisation of a complementary relationship between scientific and IT actions in Use Case 1 and setting up the basis for data exchanges and computing resources.

Future plans

The first phase of the project (about 1.5 years, up to next February 2021) shall implement, test and validate the processing for intelligent screening approaches for extreme event detection and pollutant plumes monitoring on 1 year of S5P data.

Then prototyping processing will be used to re-process 1 year of global data plus several specific on-demand periods/ regions (March-June 2021) and will be exploited by pilot users in demonstrating added value, improving the processing, and consolidating the products and services of WP4 (March 2021 - June 2022).

The upcoming months will be devoted to:

- HYGEOS/SPASCIA/ICARE technical development of the first L1 intelligent screening approach of S5P data,
- SPASCIA/ICARE technical development of the second L2 intelligent screening approach of the S5P product. Specification of processing based on artificial intelligence is ongoing.

Use Case 2

Big data earth observations: processing on-demand services for environmental monitoring

Optimal and radar images observing the Earth's land surface have become an essential source of information to address and analyse environmental issues. The diversity of Earth observation sensors makes it possible to consider these data as an unprecedented source of information, able to provide new insights into environmental monitoring.

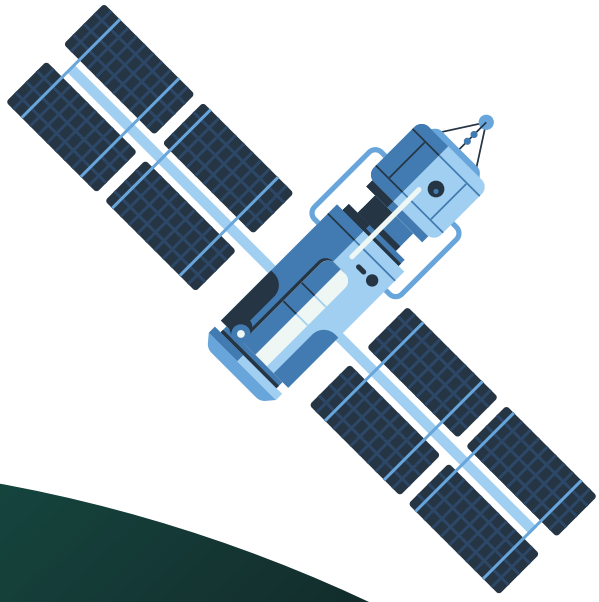
Goals

The data area testing case titled "Big data earth observations: processing on-demand services for environmental monitoring" aims to provide the academic and land management community with an interactive environment to ensure systematic or on-demand production of new knowledge useful for the environmental monitoring of territories. This will be achieved relying on the algorithmic developments carried out within the THEIA land data centre, machine and deep learning techniques adapted to spatial data, and taking advantage of the complementarities (spatial and temporal resolution) of large spatial data sets from very high-resolution sensors (SPOT, PLEIADES) and SENTINEL 1 and 2-time series.

The activities performed within this Use Case are led by IRD, together with other French academic institutes such as INRAE and CNES (involved in the THEIA land data centre), the IT private companies Geomatys and Geolabs, with the overall coordination of CINES.

Future plans

Use Case 2 output will be achieved in two different phases. The first one will be focused on the delivery of at least 24 of the optimised and deployed processing chains in the HPDA/HPC environment of the CINES. While the second phase will be centred on the deployment of demonstrators providing a UI web environment for on-demand processing, data workflows for discovery, access to earth observation (EO) raw data and products, and specifications for long-term data archiving procedures.



Use Case 3

Ocean

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.

Goals

The data area testing case titled "Ocean" aims to achieve three main goals:

1. The improvement of long-term stewardship of marine in-situ data. The SeaNoe service allows users to upload, archive and publish their data, 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 (measurements acquired more randomly, e.g. during a scientific cruise or manual work) are of particular interest.

2. The 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 visualisation and submitting 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.

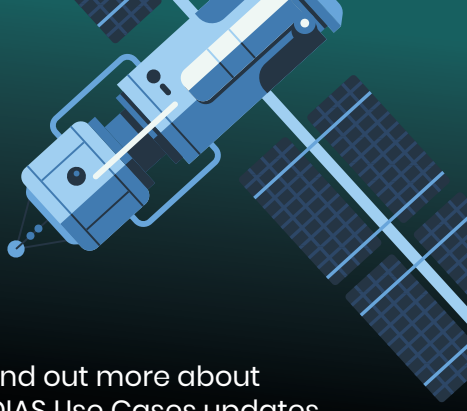
This path to achieve these objectives is led by IFREMER, together with Europe's leading research groups in ocean studies, such as the Université de Liège, MARIS, CNRS, CSC and the Finnish Environment Institute, with the coordination of CINES, the leading HPC centre in France.

Future plans

In order to fulfil the scientific goals of the use case, the work plans are mostly focused on technical developments and the implementation of tools. In particular, the tools related to the long-term archiving of both data and metadata and the storage and archiving of large salinity datasets from in situ (SeaDataCloud) and from satellite (SMOS mission) have to be developed or improved.

The team has listed different topics as part of their working plan for the forthcoming months, also providing the corresponding proposed solutions.

- Service scalability: Due to technical limitations, the maximum allowed size for data upload is presently 0.2 Tbytes.
- Solution: use of other protocols, which can be asynchronous, for example, Virtual File Systems. Sharing the allocation of necessary storage resources from different infrastructures.
- Back-office exchanges: In order to make long-tail data available in data collections, many exchanges (performed manually at present) between the involved Data Centres are necessary.
- Solution: implementing iRODS (Integrated Rule-Oriented Data System) data flows to automate these exchanges and make them more efficient.
- Securing long-term archive: Data Centre infrastructures for data archiving are not always suitable for long-term archiving and dedicated staff are not always available.
- Solution: rely on professional long-term repositories instead and distribute dataset storage across different geographically distributed repositories. This could be achieved by using, for instance, iRODS data flows.
- Fast access to datasets: In situ datasets are made available among a wide range of systems, making the assembly of multidisciplinary datasets more difficult for users.
- Solution: use a working copy data, called a technical cache or Data Lake, with a suitable structure in order to speed up and facilitate data processing. Data Lake will be periodically synchronised with the Data Centres and Data publication services.
- On-demand processing: Using specialised tools requires the installation of software and the availability of computing resources. The former can be time-demanding for users.
- Solution: deployment of the DIVAnd interpolation software tool (Deliverable 6.3.1) in a virtual machine in order to provide a significant improvement on what researchers or data experts typically have access to from their office.



To find out more about PHIDIAS Use Cases updates, visit the section of our website: **www.phidias-hpc.eu/insights**

Full alignment with EU Policies to effectively address Societal Impact(s)

Synergies between earth science communities, high-performance computing, and data-driven operations have the potential of enabling novel innovation opportunities and pave the way for the realization of brand-new applications and services. PHIDIAS foresees a large social impact in terms of overall environmental monitoring capabilities enabled by the pooling of different stakeholders, namely: researchers, public authorities, private players, and citizen scientists. By federating different data sources, PHIDIAS will provide, on one hand, an interoperable and easy to use catalogue of environmental resources accessible and browsable by anybody, and, on the other hand, will provide researchers and practitioners with a platform on top of which new knowledge and new business models can be developed.

Systemic changing solution is explicitly mentioned by the "European environment — state and outlook 2020"[1] as the single unambiguous message for all policy makers in Europe. The report clearly states that besides doing numerically more in addressing sustainability challenges we need also to start thinking to tackle them in new way. In this context, PHIDIAS can support a broad range of stakeholders with actionable information that can be used to implement practicable and measurable solutions for a more sustainable future. As an example, large cloud computing providers operating in multiple continents will be able to tap into the large knowledge base provided by PHIDIAS to predict where energy could be cheaper (e.g. due to better environmental predictions, and thus moving some computing tasks to datacentres located in those areas).

Moreover, the current attempts at meeting the air quality standards defined by the EU Ambient Air Quality Directives[2] is current only possible thanks to the official reports on air quality provided by the individual countries. However, air quality monitoring is far from being a single country matter, since pollution produced in any country has undoubtedly an impact on a global scale. As result, the holistic approach pursued by PHIDIAS in consolidating environmental information on the air quality from multiple sources, is going to equip policy makers worldwide with important data for their operations and fundamental decisions.

The Mission Board on Healthy Oceans, Seas, Coastal and Inland Waters[3] aims at a measurable improvement in the conditions of our oceans and waters by 2030. Starting from the famous motto that "If You Can't Measure It, You Can't Improve It", the action of PHIDIAS can provide policy makers with actionable information to be turned into directives or new legislation. At the same time, PHIDIAS can help both researchers to explore new research directions, but it can also help citizens that want to protect their local resources by demonstrating how small everyday choices can have a tangible impact.

Finally, it is also worth stressing that the social dimension cannot be underestimated, it is however our standpoint that the transparent access to environmental data enabled by PHIDIAS can allow to provide also due care to the players and countries that could be negatively affected by low-carbon economy.

[1] *https://www.eea.europa.eu/publications/soer-2020*
[2] *https://www.eea.europa.eu/publications/air-quality-in-europe-2019*
[3] *https://op.europa.eu/en/web/eu-law-and-publications/publication-detail/-/publication/d0246783-b68a-11ea-bb7a-01aa75ed71a1*

www.phidias-hpc.eu

How to reach us

@PhidiasHpc

PHIDIAS-HPC
https://www.linkedin.com/company/phidias-hpc

PHIDIAS HPC

www.slideshare.net/PhidiasHPC1

Email: info@phidias-hpc.eu

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement No. **INEA/CEF/ICT/ A2018/1810854**.