PHIDIAS

Prototype of HPC/Data Infrastructure for On-demand Services

Project Title	Prototype of HPC/Data Infrastructure for On-demand Services
Project Acronym	PHIDIAS
Grant Agreement No.	INEA/CEF/ICT/A2018/1810854
Start Date of Project	01.09.2019
Duration of Project	36 Months
Project Website	www.phidias-hpc.eu

D3.2.1 - Use cases for processing data from different Earth System compartments

Work Package	**WP3**
Lead Author (Org)	**Joël SUDRE (CNRS)**
Contributing Author(s) (Org)	**Gilbert Maudire (IFREMER)** **Jean-Christophe Desconnets (IRD)** **Pascal Prunet (SPASCIA)**
Due Date	**01.09.2020**
Date	**30.11.2020**
Version	**V1.1**

Dissemination Level

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Versioning and contribution history

Version	Date	Author	Notes
0.1	11.09.2020	Joël Sudre (CNRS) Gilbert Maudire (IFREMER) Jean-Christophe Desconnets (IRD) Pascal Prunet (SPASCIA)	TOC and compilation of information from WP3, 4, 5 &6
0.2	01.10.2020	Joël Sudre (CNRS) Gilbert Maudire (IFREMER) Jean-Christophe Desconnets (IRD) Pascal Prunet (SPASCIA)	Update after first reading phase.
0.3	01.10.2020	Jean-Christophe Desconnets (IRD)	Update and fix section 3 land surface use case
1.0	08.10.2020	Pascal Prunet (SPA)	Update and fix section 2 atmosphere use case

Disclaimer

This document contains information which is proprietary to the PHIDIAS Consortium. Neither this document nor the information contained herein shall be used, duplicated or communicated by any means to a third party, in whole or parts, except with the prior consent of the PHIDIAS Consortium.

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Table of Contents

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

List of Figures

List of Tables

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

TERMINOLOGY

Terminology/Acronym	Description
3D	Three Dimensions
4D	Four Dimensions
AERIS	Atmosphere and Service Data Pole
AI	Artificial Intelligence
API	Application Programming Interface
ATBD	Algorithm Theoretical Basis Document
BGC	Biogeochemical
CAMS	Copernicus Atmosphere Monitoring Service
CDI	Common Data Index
CEF	Connecting Europe Facility
CES	*Centre d'Expertise Scientifique* – Centre of Scientific Expertise (FR)
CF	Climate and Forecast
CIRAD	*Centre de coopération Internationale en Recherche Agronomique pour le Développement* – Agricultural Research Centre for International Development (FR)
CMEMS	Copernicus Marine Environment Monitoring Service
CNES	*Centre National d'Etudes Spatiales* – National Centre for Space Studies (FR)
CNRS	*Centre National de la Recherche Scientifique* – National Centre for Scientific Research (FR)
CO	Carbon Monoxide
CSV	Comma-Separated Values
DIAS	Data and Information Access Services
DIVAnd	Data-Interpolating Variational Analysis N dimensions
DOI	Digital Object Identifier
EC	European Commission
ECMWF	European Centre for Medium-Range Weather Forecasts
ENVRI	Environmental Research Infrastructure
EMODNET	European Marine Observation and Data Network
EO	Earth Observation
EOSC	European Open Science Cloud
ESA	European Space Agency
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
EuroHPC JU	European High Performance Computing Joint Understanding
FAIR	Findable, Accessible, Interoperable, Reusable
FTP	File Transfer Protocol
GB	Giga Byte
GOSAT	Greenhouse gases Observing SATellite

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Terminology/Acronym	Description
GPU	Graphics Processing Unit
HPC	High-Performance Computing
HPDA	High-Performance Data Analytics
HR	High Resolution
HTTP	Hypertext Transfer Protocol
IASI	Infrared Atmospheric Sounding Interferometer
IASI – NG	Infrared Atmospheric Sounding Interferometer New Generation
ICSU	International Council for Science
ID	Identity
IFCB	Imaging FlowCytobot
INEA	Innovation and Network Executive Agency
INERIS	*Institut National de l'EnviRonnement Industriel et des Risques* – National Institute for Industrial Environment and Risks (FR)
INSPIRE	Infrastructure for Spatial Information in Europe
IODD	Input Output Data Definition
IODS	Input Output Data Specification
iRODS	Integrated Rule Oriented Data System
JPEG	Joint Photographic Experts Group
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
L1	Level 1
L2	Level 2
LPIS	Land Parcel Identification System
MB	Mega Byte
MPEG	Moving Picture Experts Group
MTG – IRS	Meteosat Third Generation InfraRed Sounder
NASA	National Aeronautics and Space Administration (US)
NetCDF	Network Common Data Form
NRT	Near Real Time
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OBIA	Object-Based Image Analysis
ODATIS	Ocean Data Information and Services
OGC	Open Geospatial Consortium
OLCI	Ocean Land Colour Instrument
OTB	Orfeo Toolbox
OTBTF	Orfeo Toolbox TensorFlow
PCA	Principal Component Analysis
PEPS	*Plateforme d'Exploitation des Produits Sentinel* – Operating Platform Sentinel Products (FR)
PHIDIAS	Prototype of HPC/Data Infrastructure for On-demand Services
PNDB	*Pôle National de Données de Biodiversité* - National Biodiversity Data Pole (FR)

Terminology/Acronym	Description
PRF	Product Readme File
PUM	Product User Information
RDA	Research Data Alliance
RI	Research Infrastructure
RIS	Research Information Systems
S1	Sentinel 1
S2MP	Soil Moisture Product at Plot Scale
S5P	Sentinel 5 Precursor
SAR	Synthetic Aperture Radar
SDGs	Sustainable Development Goals
SMOS	Soil Moisture and Ocean Salinity
SNAP	Sentinel Application Platform
SPOT	*Satellite Pour l'Observation de la Terre* – Satellite for Earth Observation (FR)
SQL	Structured Query Language
SRTM	Shuttle Radar Topography Mission
SRTM HGT	Shuttle Radar Topography Mission Height
SSH	Secure Shell
SSM	Surface Soil Moisture
SWIR	ShortWave InfraRed Imagery
TETIS	*Territoire Environnement Télédétection Information Spatiale* – Land Environment Remote Sensing and Spatial Information (FR)
TIFF	Tag Image File Format
TROPOMI	TROPOspheric Monitoring Instrument
UC	Use Case
UMR	*Unité Mixte de Recherche* – Research Mixed Unit (FR)
UN	United Nations
URL	Uniform Resource Locator
UV	Ultra Violet
RGP	Rigid Gas Permeable
RI	Research Infrastructure
UVNS	UltraViolet Near-infrared Shortwave
VHR	Very High Resolution
WCS	Web Coverage Service
WDS	World Data System
WMS	Web Map Service
WMTS	Web Map Tile Service
WP	Work Package
WPS	Web Processing Service

Executive Summary

The objective of this document is to describe the different use cases that will be developed within the framework of the PHIDIAS project.
This analysis will ultimately make it possible to define the common system making it possible to carry out the processing operations and to exploit the data produced.

PHIDIAS scientific use cases are handled in the WP 4/5/6:

- WP4 (Atmosphere) Intelligent screening of large amount of satellite data for detection and identification of anomalous atmospheric composition events
- WP5 (Land Surfaces) Big data Earth Observations: processing on-demand and products dissemination for environmental monitoring
- WP6 Ocean Use Case objective is to improve the use of cloud services for marine data management, data service to user in a FAIR perspective, data processing on demand, taking into account the European Open Science Cloud (EOSC) challenge and the Copernicus Data and Information Access Services (DIAS). In those terms, this use case can be seen as one of the prototypes, for marine environmental data, of the future Blue Cloud foreseen by the European Commission.

1 Introduction

1.1 PHIDIAS project

This project aims at developing a consolidated and shared HPC and Data service by building on pre-existing and emerging infrastructure in order to create a federation of “user to infrastructure” services. Specifically, PHIDIAS Consortium will further develop and provide new services to better discover, manage and process spatial, marine and environmental data.

Nowadays, in the digital era, High-Performance Computing is the key-asset for major advances and a fundamental resource for the future of the European Union. The usage of so-called supercomputing is largely increasing as well as the number of data-intensive critical applications.

This turning point is involving multiple Stakeholders and fragment of our society:

- Industry and Small-Medium Enterprises are counting on the power of supercomputers to develop innovative, faster and cost-effective solutions and diminish time to market for their products and services to the bare minimum.
- Modern Science requires the full capacity of high computing to successfully achieve significant discoveries and progress.

The transition from petascale to exascale is in full course and it represents a window of opportunity for Europe.

The European Commission will complement the European Data Infrastructure under the European Cloud initiative with a long-term and large-scale flagship initiative on quantum technologies. The objective is to release the complete potential of quantum which holds the promise to solve computational problems beyond current supercomputers.

In this framework, European Council adopted in September 2018 the regulation that established the European High-Performance Computing Joint Undertaking (EuroHPC Joint Undertaking) to gather European and other participating countries efforts and resources with a view to building in Europe a top-notch supercomputing and data infrastructure within a competitive innovation ecosystem in relevant technologies and applications.

PHIDIAS answers INEA CEF Public Open Data Call for proposals, mostly focusing on 3rd objective: “Creation of generic access services to increase the HPC and data capacities of the European Data Infrastructure”.

Following the rationale behind the EuroHPC Joint Undertaking Initiative to underpin the ambition of making European exascale achievable in a short time whilst developing a pan-European HPC infrastructure and HPC-based services, PHIDIAS is going to propose generic

Co-financed by the Connecting Europe Facility of the European Union
The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

workflow for massive scientific data by combining computing, dissemination and archiving resources in a single framework in order to make this process this going forward.

1.2 Scientific objectives

The Earth and its external fluid envelopes are complex dynamical systems characterized by physical, chemical and biological processes interacting across a broad continuum of temporal and spatial scales. It is also the home of human societies interacting ever more closely with these environments. Observing, understanding and modelling the Earth systems' history and their integrated functioning and predicting their responses to global changes is a key and pressing research challenge and a necessity for many environmental and socio-economic applications related to the implementation of the UN Sustainable Development Goals (SDGs).

Access to all of the data diversity from the various subsystems and environments is vital to address the challenges that face us, such as natural hazards, increased anthropogenic pressures, climate change, resources and biodiversity issues and their impacts on health.

Different information systems have been developed in France to make data findable, accessible, interoperable and reusable (FAIR) in the different domains through national data centres dedicated to the atmosphere, oceans, land surfaces and solid Earth (Data Terra Research Infrastructure (RI)), the national biodiversity data centre (PNDB RI), together with results from reference climate simulations (CLIMERI-France RI).

The landscape remains however fragmented into domains of independent research that lead to a proliferation of data sources, standards and tools, together with a wide diversity of data, data products and volumes of data, some of them exceeding already the petabyte scale. Integrated, transparent and seamless access to this cornucopia of data across a continuum of interoperable and distributed infrastructures is today a critical issue for Earth system, biodiversity and environment sciences to overcome this fragmentation, accelerate science-driven extraction, composition and use of these data enabling innovative research practices and discovery processes that address inter and transdisciplinary scientific challenges and can inform decision support systems. At the same time this must enable high-quality data products synthetized from different sources to be easily transferred back to knowledge on subsystems.

The objective of the PHIDIAS project is to initiate a first technical solution (catalog, processing interfaces, ...) allowing simultaneous manipulation of data from different Earth compartments (Ocean, Atmosphere, Solid Earth, Land Surfaces, Biodiversity). The idea is to start with a specific subset of data (both spatial and in-situ) produced within the framework of 3 PHIDIAS use cases (WP4, WP5 & WP6).

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

1.3 Overall description of the WP3/4/5/6

PHIDIAS scientific use cases are handled in the WP 4/5/6:

- WP4 (Atmosphere) Intelligent screening of large amount of satellite data for detection and identification of anomalous atmospheric composition events
- WP5 (Land Surfaces) Big data Earth Observations: processing on-demand and products dissemination for environmental monitoring
- WP6 Ocean Use Case
    - The general objective of this Ocean use case is to improve the use of cloud services for marine data management, data service to user in a FAIR perspective, data processing on demand, taking into account the European Open Science Cloud (EOSC) challenge and the Copernicus Data and Information Access Services (DIAS). In those terms, this use case can be seen as one of the prototypes, for marine environmental data, of the future Blue Cloud foreseen by the European Commission.

As can be seen, these use cases deal with scientific topics which concern distinct compartments of the Earth System: Atmosphere, Ocean and Land Surfaces.
These compartments are currently scientifically and technically managed separately like independent silos.
Understanding geophysical, geodynamic and environmental processes, demands analyzing numerous and very large datasets (satellite, in situ, campaigns, long term observations but also experimentation results, model outputs, …) from different Earth Compartment.

This is the objective of the WP3 of PHIDIAS whose goal is to build a system to remove existing data silos and to allow user to develop programs allowing to handle the full “Earth System” perimeter (e.g. programs in the field of Climate or Sustainable development). This will be done through the design and development of

- A Common Metadata Repository (based on a Common Information Model also developed in the frame of WP3) allowing to discover and access the data.
- An end-users web Common Interactive Processing Service

WP3 will be based on the process and data from WP4, WP5 & WP6.

Co-financed by the Connecting Europe Facility of the European Union
The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

2 Atmosphere use case

2.1 WP description: Intelligent screening of large amount of satellite data for detection and identification of anomalous atmospheric composition events

The goal of this WP is to exploit HPC and high-performance data management provided by WP2 for developing intelligent screening approaches for the exploitation of large amounts of satellite atmospheric data in an operational context. It will implement a prototype service on the already available Sentinel 5 Precursor (S5P) European atmospheric sounding mission:

1. Exploit an innovative technique for implementing data screening method applied to the spectral dimension of the data cubes;
2. Process 1 year of S5P data (~150 Terabytes) for the detection of a few types of dedicated events (e.g., wild fires, urban pollution episodes, industrial accidents/pollution);
3. Consolidate and prepare the extension of this service for the efficient operational exploitation of future satellite instruments Sentinel 5 (UVNS, IASI-NG) and Sentinel 4 (UNV, MTG-IRS).

This work will build on cross fertilization with WP3, WP5 and WP6 to consolidate/improve the efficiency and genericity of the intelligent screening of environmental satellite data.

Task 4.1: Implementation of PCA filtering techniques for detection of exceptional atmospheric event

- Test and validation on samples of available data measurements: IASI, S5P, GOSAT, etc.;
- Scientific consolidation of the processing on S5P for targeting atmospheric events (wild fire, anthropogenic pollution plumes, industrial accidents).

Task 4.2: Adaptation and preparation of tools, data and environment for the data screening processing of Sentinel 5 precursor data

- Access and interface with the data through common interface standards;
- Definition of Metadata information model and format (with support from pilot users);
- Preparation/implementation of the processing for common interactive processing service and portal.

Task 4.3: Processing of 1 year of global S5P data for detection of extreme atmospheric events

- S5P data processing;
- Quality analysis and verification;
- S5P products and metadata available to the pilot users.

Task 4.4.1: Use case 1 – Exploitation for monitoring and alert service development

- Qualification of the product quality and added-value (with respect to already available/planned L2 products) for real time monitoring of targeted atmospheric events;
- Specification of monitoring and alert services, and dimensioning of requirements for real-time exploitation, processing and archiving.

Task 4.4.2: Use case 2 – Scientific exploitation of the product

1) The partners will address the exploitation of the service for Carbon Monoxide (CO) plume detection and classification. 2) AERIS will develop scientific exploitation and promotion of the service with the AERIS scientific users community. For these two topics, the following WP will be carried on:

- Demonstration of added value of the products for scientific analysis of atmospheric events;
- Promotion and dissemination to the S5P scientific user communities.

Task 4.5: Update the service, and prepare for the exploitation of Copernicus Atmosphere satellite data

- Interaction with - and feedbacks from - operational users at national level (envisaged pilot already expressing interest: INERIS) and European level (envisaged pilot already expressing interest: CAMS),
- Prepare the extension of the service for the efficient operational exploitation of future satellite instruments Sentinel 5 (UVNS, IASI-NG) and Sentinel 4 (UNV, MTG-IRS): specification, dimensioning size and requirements for the storage and processing loads.

2.2 *Objectives of WP4 Use Case*

WP4 use case addresses a well identified, key requirement for the exploitation of Earth Observation: to provide the capacity of intelligent screening of large amounts of satellite data for targeting scenes of interest (e.g., extreme atmospheric events), in view of their dedicated processing or exploitation. The potential main users are: the scientific community (for the efficient specific/thematic analyses of the data); the operational services at national/European levels (for the monitoring and validation of their forecast systems); the service providers, for implementing dedicated applications for end users (e.g., alert or decision support services).

Objective of the WP4 Use Case: using HPC (high-performance computing) and HPDA (high-performance data analytics) capacities for developing intelligent screening approaches for the exploitation of large amounts of satellite atmospheric data in an operational context. It is proposed to implement these approaches as a prototype service on the already available S5P European atmospheric sounding mission.

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

The objective is to implement, demonstrate and make available to the users a service offering smart data filtering, in order to:

- detect the relevant data within the huge volume of measurements,
- identify and assess the pertinent information
- Qualify the targeted data for dedicated exploitation by users.

and provide the users with an access to the filtered data and the corresponding metadata.

Two different processing, applied to two different S5P datasets and addressing different, will be implemented:

1. PCA-based screening of L1 data (SWIR) for detection of extreme events. Based on experience and methods developed for IASI, implementation and consolidation of algorithms and tools for generic processing of atmospheric spectra recorded by S5P.

2. New AI methods for objective/automatic detection of pollutant plumes and sources from L2 products (CO, others?: CH_4 NO_2, SO_2) : Based on innovative algorithms and on SPASCIA experience with physical methods analysing specific signal enhancements.

.

2.2.1 PCA-based screening of L1 data

- Input: Applied on S5P Level 1 (L1, correspond to calibrated spectra) data.
- Objective: On the flow, real time processing of the data, for detection of extreme events: Output: selection of characterized (through metadata) and stored dataset of interest, which will be available to the users. No storage of input data is needed.
- Prototyping: during the development phase, a prototype will be tested on 1 year of S5P L1 dataset. Need for storage of this 1-year dataset at AERIS/ICARE during the development phase (i.e., duration of the PHIDIAS project).

2.2.2 Detection of gas plumes and sources from L2 products

- Input: Applied on S5P Level 2 (L2, correspond to retrieved gas concentrations) data.
- Objective: On demand re-processing of the data, for detection of plumes and sources over a region/period specified by the user: Output: characterization of this region/period dataset (through metadata) flagging the data associated to the plumes, and localizing the corresponding sources, which will be available to the users. Storage of the input data and output metadata is envisaged.
- Prototyping: during the development phase, a prototype will be tested on 1 year of S5P L2 dataset (chosen period: 1st May 2018 to 30th April 2019). AERIS/ICARE will also store "on the flow" all the L2 data from 1st May 2018.

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

2.3 Input data

2.3.1 Description

The TROPOspheric Monitoring Instrument (TROPOMI) on-board the Sentinel-5 Precursor satellite, which was successfully launched in October 2017, is a space borne nadir viewing imaging spectrometer measuring solar radiation reflected by the Earth in a push-broom configuration. It has a wide swath on the terrestrial surface and covers wavelength bands between the ultraviolet (UV) and the shortwave infrared (SWIR) (See Table 1) combining a high spatial resolution with daily global coverage. These characteristics enable the determination of gases with unprecedented level of detail on a global scale introducing new areas of application. Abundances of the atmospheric column-averaged dry air mole fractions of several gases (Level 2 products, see Table 2) are retrieved from TROPOMI's radiance measurements (Level 1 products, Table 1).

TROPOMI	UV		UVIS		NIR		SWIR	
Band	1	2	3	4	5	6	7	8
Spectral coverage [nm]	270 – 320		320 – 495		675 - 775		2305 – 2385	
Full spectral coverage [nm]	267 - 332		303 - 499		660 - 784		2299 - 2390	
Spectral resolution [nm]	0.49		0.54		0.38		0.25	
Spectral sampling ratio	6.7		2.5		2.8		2.5	
Spatial sampling [km^2]	7 x 28	7 x 3.5				7 x 3.5	7 x 7	

Table 1 – TROPOMI Level 1B products

Product	Spectrometer	Application
Ozone	UV, UVIS	Ozone layer monitoring, UV-index forecast, Climate monitoring
NO_2	UVIS	Air quality forecast and monitoring
CO	SWIR	Air quality forecast and monitoring
CH_2O	UVIS	Air quality forecast and monitoring
CH_4	SWIR	Climate monitoring
SO_2	UVIS	Air quality forecast and monitoring, Climate monitoring, Volcanic plume detection
Aerosol	UVIS, NIR	Air quality forecast and monitoring, Climate monitoring, Volcanic plume detection
Clouds	UVIS, NIR	Climate monitoring
UV-Index	UVIS	UV index forecast

Table 2 – TROPOMI Level 2 geophysical products

As for all Sentinel missions, the Sentinel-5P products are freely available to users via the Copernicus Open Access Hub. Data are available operationally from 30th April 2018. Detailed information are available on https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-5p/products-algorithms, and is reproduced below.

Level 1B – User technical documentation

The information needed to properly use the Level 1B data can be found in Table 3:

- IODS (Input Output Data Specification): description of the products that are the result from the Level 0 to Level 1b processing
- ATBD (Algorithm Theoretical Basis Document): high level description of the algorithms used in the Level-0 to 1b data processing
- PRF (Product Readme File): description of changes between different product versions and overall quality information (available a few months after launch)

Table 3 – TROPOMI Level 1B radiance/irradiance products and user documentation

File type	Spectrometer	Spectral range [nm]	Comment	User Documentation
L1B_RA_BD1	UV	270 - 300	Radiance product band 1	IODS ATBD PRF Other documents
L1B_RA_BD2		300 - 320	Radiance product band 2	
L1B_RA_BD3	UVIS	320 - 405	Radiance product band 3	
L1B_RA_BD4		405 - 500	Radiance product band 4	
L1B_RA_BD5	NIR	675 - 725	Radiance product band 5	
L1B_RA_BD6		725 - 775	Radiance product band 6	
L1B_RA_BD7	SWIR	2305-2345	Radiance product band 7	
L1B_RA_BD8		2345-2385	Radiance product band 8	
IR_UVN	UVN	270-775	Irradiance product UVN module	
IR_SIR	SWIR	2305-2385	Irradiance product SWIR module	

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Level 2 – User technical documentation

The information needed to properly use the Level 2 data can be found in Table 4:

- PUM (Product User Information): information on the technical characteristics of the S5P/TROPOMI Level 2 products
- ATBD (Algorithm Theoretical Basis Document): detailed information on the retrieval algorithms
- IODD (Input Output Data definition): description of the input and output data of the S5P/TROPOMI Level 2 processing
- PRF (Product Readme File): description of changes between different product versions and overall quality information (available few months after launch)

Table 4 – TROPOMI Level 2 geophysical products and user documentation

Product type	Parameter	User Documents
L2__O3____	Ozone (O_3) total column	PRF-O3-NRTI, PRF-03-OFFL, PUM-O3, ATBD-O3, IODD-UPAS
L2__O3_TCL	Ozone (O_3) tropospheric column	PRF-03-T, PUM-O3_T, ATBD-O3_T, IODD-UPAS
L2__O3__PR	Ozone (O_3) profile	PUM-PR , ATBD-O3_PR , IODD-NL
L2__O3_TPR	Ozone (O_3) tropospheric profile	PUM-PR , ATBD-O3_PR , IODD-NL
L2__NO2___	Nitrogen Dioxide (NO_2), total and tropospheric columns	PRF-NO2, PUM-NO2, ATBD-NO2, IODD-NL
L2__SO2___	Sulfur Dioxide (SO_2) total column	PRF-SO2, PUM-SO2, ATBD-SO2, IODD-UPAS
L2__CO____	Carbon Monoxide (CO) total column	PRF-CO, PUM-CO, ATBD-CO, IODD-NL
L2__CH4___	Methane (CH_4) total column	PRF-CH4, PUM-CH4, ATBD-CH4, IODD-NL
L2__HCHO__	Formaldehyde (HCHO) total column	PRF-HCHO, PUM-HCHO , ATBD-HCHO , IODD-UPAS

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

L2__CLOUD_	Cloud fraction, albedo, top pressure	PRF-CL, PUM-CL, ATBD-CL, IODD-UPAS
L2__AER_AI	UV Aerosol Index	PRF-AI, PUM-AI, ATBD-AI, IODD-NL
L2__AER_LH	Aerosol Layer Height (mid-level pressure)	PRF-LH, PUM-LH , ATBD-LH , IODD-NL
UV product	Surface Irradiance/erythemal dose	-
L2__NP_BDx, x=3, 6, 7	Suomi-NPP VIIRS Clouds	PRF-NPP, PUM-NPP, ATBD-NPP

Data Volume

Product type	Volume/day (GB)
L1B	475 GB
L2	35 GB
Total	510 GB

2.3.2 Access for PHIDIAS processing chain through ICARE facilities

2.3.2.1 ICARE & AERIS

A complete description of ICARE data center can be found at: https://www.icare.univ-lille.fr/about-us-2/in-a-nutshell/

The ICARE data Center was created in 2003 by CNES, CNRS, the Nord-Pas-De-Calais Regional Council, and the University of Lille, to provide various services to support the research community in fields related to atmospheric research, such as aerosols, clouds, radiation, water cycle, and their interactions. ICARE's initial emphasis is the production and distribution of remote sensing data derived from Earth observation missions from CNES, NASA, and EUMETSAT. One of ICARE's main components is the Data and Services Center, located at the University of Lille, which develops science algorithms and production codes, building on the expertise from various partner Science Computing Facilities, and distributes products to the user's community. ICARE is one of the 4 centers of AERIS, the French Data Infrastructure for Atmosphere created in 2014.

For more information about AERIS, please read https://en.aeris-data.fr/

2.3.2.2 Data Access

The input data will be available through the usual AERIS/ICARE distribution channels:

- HTTP : http://www.icare.univ-lille1.fr/archive?dir=S5P/
- FTP : ftp://ftp.icare.univ-lille1.fr/SPACEBORNE/S5P

The data are publically available, but the users need to be registered to access to these services. The request form is available here: http://www.icare.univ-lille1.fr/register

To end with, the users accredited by project PIs can have a direct access to the AERIS/ICARE computational resources through SSH. The service is described here: http://www.icare.univ-lille1.fr/services/cluster.

One alternative option, to be investigated, could be to share the AERIS/ICARE S5P archive with CINES through iRODS.

2.3.3 Use in the PHIDIAS processing chain

2.3.3.1 PCA-based screening of L1 data

Operational use case (at the end of PHIDIAS project): On the flow, real time processing of the data, for detection of extreme events. This will produce a selection of characterised (through metadata) and stored dataset of interest (preliminary estimate: less than 1% of the data), available to the users.
Development phase (during the PHIDIAS project): a prototype will be tested on 1 year of S5P L1 dataset (chosen period: 1st May 2018 to 30th April 2019). There is a need for storage of this 1-year dataset at ICARE during the development phase (i.e., duration of the PHIDIAS project). Specification, volume and definition of output products will be done during this development phase.

2.3.3.2 Detection of plumes and sources from L2 products

Operational use case (during phase 2 of PHIDIAS project (pilot use cases)) : On demand re-processing of the data, for detection of plumes and sources over a region/period specified by the user : Output : characterisation of this region/period dataset (through metadata) flagging the data associated to the plumes, and localising the corresponding sources, which will be available to the users. Storage of the input data and output metadata is required.

Development phase (first 1.5 year of PHIDIAS project): Few days, dedicated regions for development and testing; **Demonstration phase** (last year of PHIDIAS): test on data selection requested by the users: SRON and SPASCIA pilot users. In a mid-term perspective (after PHIDIAS project), there are possible needs for “on the flow” processing of specific regions.

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

2.4 Processing

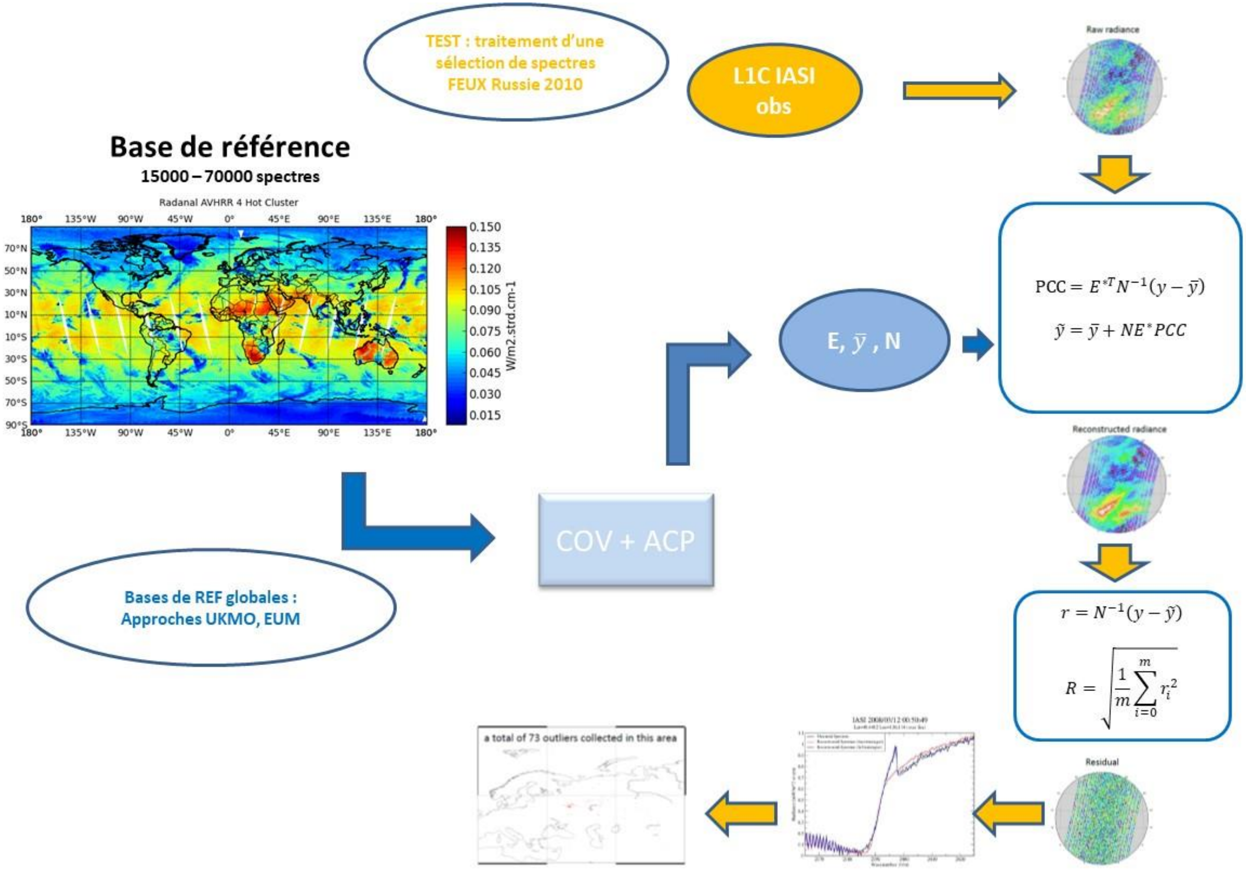
2.4.1 PCA-based screening of L1 data

2.4.1.1 High level description

The method is based on the analysis of the variability of a set of n S5P spectra y (S5P L1 product) of dimension m (the number of spectral channels): this is the learning dataset, or reference dataset. This dataset shall include observed data with different atmospheric and surface conditions, different viewing angles . It shall contain a large number of data (typically 100 000) in order to avoid non representative correlations. The behaviour of the process and its property of noise filtering and signal screening are strongly dependent on the definition of the reference dataset.

The principle is to compute from this reference dataset the covariance matrix of the measurements variability, then to apply principal component analysis (PCA) in order to extract the main Eigen structures (eigenvectors and eigenvalues). These Eigen structures will serve as a reference basis (or PC basis) for the analysis of each measurement: projection of the measurement on the reference basis; truncated reconstruction, computation of the reconstruction residuals and of the reconstruction score.

The proposed processing is illustrated on the Figure below for the analysis of IASI data. A similar processing will be implemented for S5P.

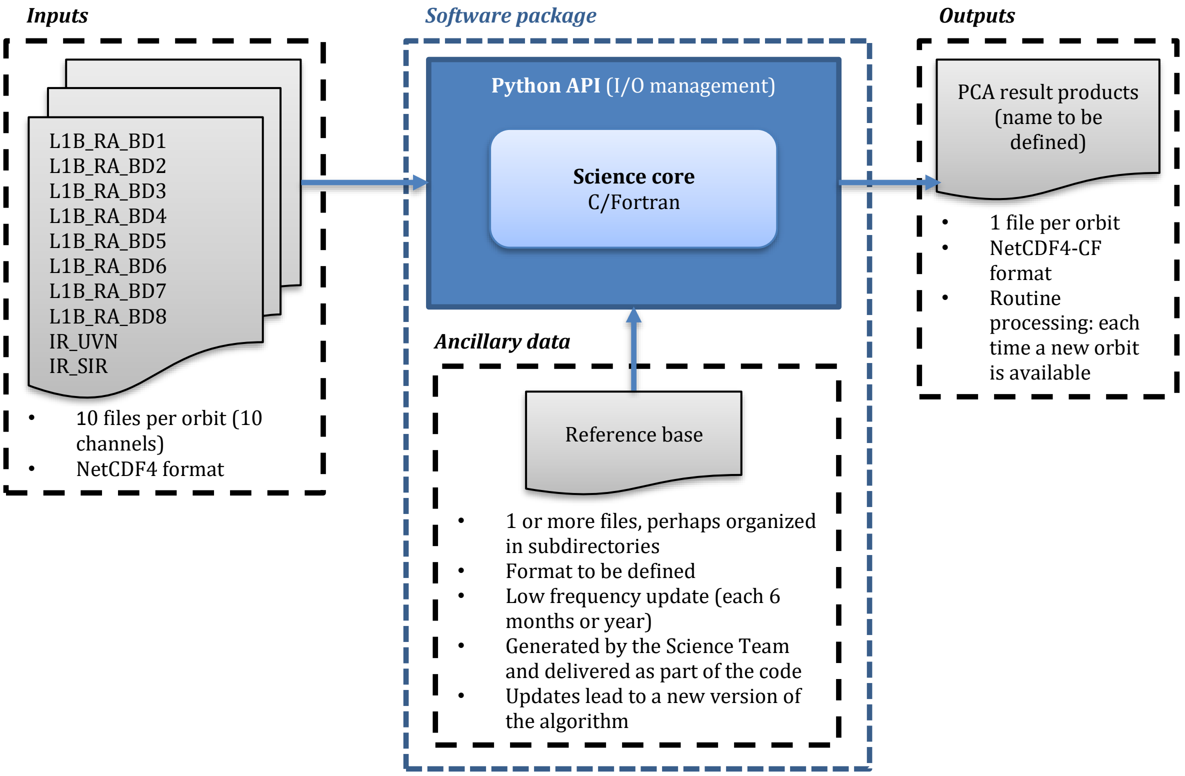

TEST : traitement d'une sélection de spectres FEUX Russie 2010
L1C IASI obs
Raw radiance
Base de référence
15000 – 70000 spectres
Radanal AVHRR 4 Hot Cluster
W/m2.strd.cm-1
0.150
0.135
0.120
0.105
0.090
0.075
0.060
0.045
0.030
0.015
$PCC = E^{*T}N^{-1}(y - \bar{y})$
$\hat{y} = \bar{y} + NE^{*}PCC$
E, $\bar{y}$, N
COV + ACP
Reconstructed radiance
$r = N^{-1}(y - \hat{y})$
$R = \sqrt{\frac{1}{m}\sum_{i=0}^{m} r_i^2}$
Residual
Bases de REF globales : Approches UKMO, EUM
a total of 73 outliers collected in this area

PCA based screening, illustration for IASI data : input are in orange, ancilliary data in blue.

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

These processing allows an efficient detection of extreme event which are not- or under-represented in the reference dataset, associated to a high reconstruction score. The identification and analysis of the specific detected event should require the exploitation of the spectral structure of the reconstruction residuals.

2.4.1.2 General Technical Scheme


Inputs
Software package
Outputs
L1B_RA_BD1
L1B_RA_BD2
L1B_RA_BD3
L1B_RA_BD4
L1B_RA_BD5
L1B_RA_BD6
L1B_RA_BD7
L1B_RA_BD8
IR_UVN
IR_SIR
• 10 files per orbit (10 channels)
• NetCDF4 format
Python API (I/O management)
Science core
C/Fortran
Ancillary data
Reference base
• 1 or more files, perhaps organized in subdirectories
• Format to be defined
• Low frequency update (each 6 months or year)
• Generated by the Science Team and delivered as part of the code
• Updates lead to a new version of the algorithm
PCA result products (name to be defined)
• 1 file per orbit
• NetCDF4-CF format
• Routine processing: each time a new orbit is available

2.4.2 Detection of plumes and sources from L2 products

2.4.2.1 High level description

The approach will use as input time series of regional maps of level 2 product associated with a pollutant (e.g.: NO_2, CO, CH4), consistently connected in space and close in time (typically daily maps over a period of a few days). The processing will explore the maps in order to:

- identify the sources points associated with point source emissions of the mapped pollutant, and flag the corresponding pixels.
- Identify all the pixels associated with the plume emitted by a given source point, and flag the corresponding pixels.
- For a pixels impacted by several sources, provide an estimation of the contribution of each sources to this pixel.

The processing will need, in addition to the input L2 product, the following ancillary information:

- 3D wind fields associated to each pixel map, interpolated from ECMWF forecast (or provided by level 2 product if available)
- A priori estimate of the position of the sources, if available.

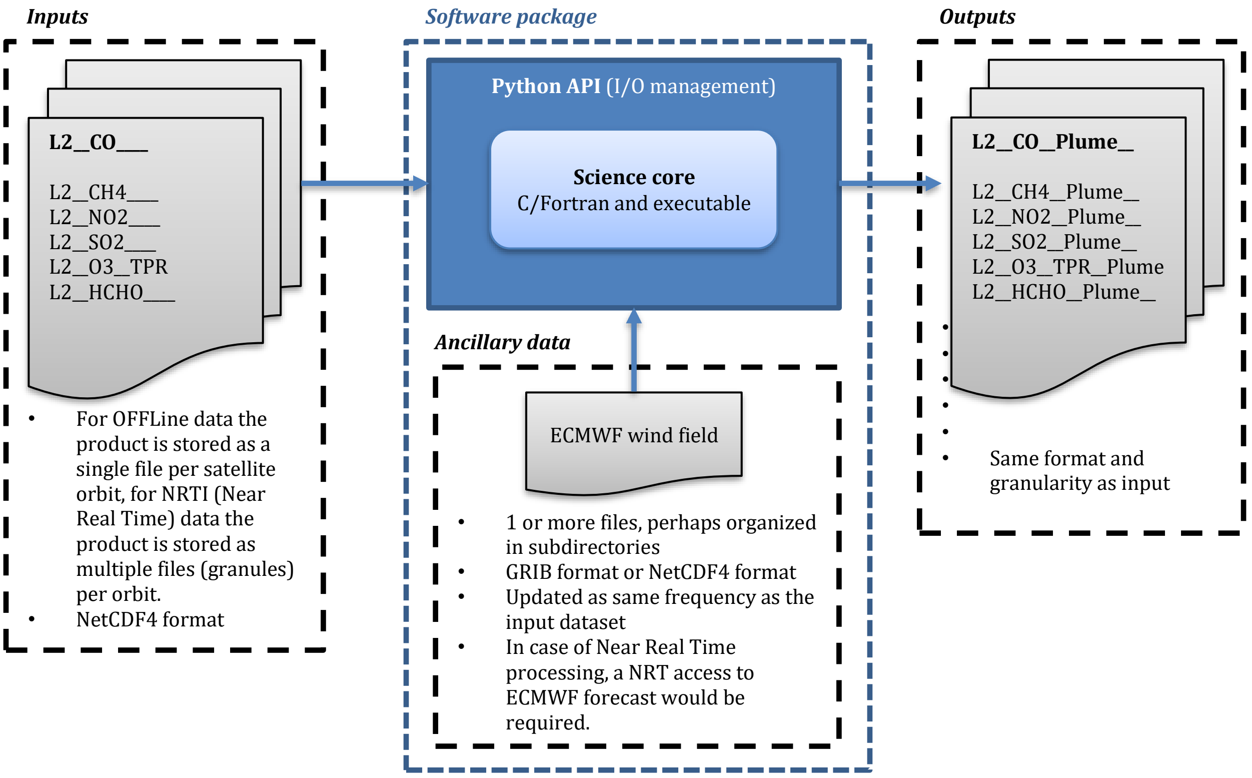
The output of the processing will be the input regional/global L2 product map with additional information, consisting in additional flags, typically:

- a flag associated to the qualification of the pixel (source/plume/none),
- a quality flag,
- additional information (number of associated sources, contribution of each source, distance to the source, ...).

Note that the above description is for the operational phase. The development phase of this processing will involve specific computation and storage resources, which will be detailed in a next version of this document.

2.4.2.2 General Technical Schema

The baseline processing will apply on off Line data. If Near Real Time (NRT) processing is required (to be discussed with the users) specific access to some data (in particular ECMWF ancillary dataset) shall be verified.



- For OFFLine data the product is stored as a single file per satellite orbit, for NRTI (Near Real Time) data the product is stored as multiple files (granules) per orbit.
- NetCDF4 format

- 1 or more files, perhaps organized in subdirectories
- GRIB format or NetCDF4 format
- Updated as same frequency as the input dataset
- In case of Near Real Time processing, a NRT access to ECMWF forecast would be required.

- Same format and granularity as input

The technical scheme above is nominally applying on one input type (L2__CO__, in bold), providing one corresponding output (L2__CO_Plume_, in bold). In a next step, it could be applied to other type of input, and/or to multiple inputs, providing the corresponding number of outputs.

2.4.3 PCA-based screening of L1 data

Use Case I: systematic processing, for use in real time and/or storage for future use.
Only selected input data and output metadata to be provided to the users.

2.4.4 Output data

2.4.4.1 PCA-based screening of L1 data

Use Case I: systematic processing, for use in real time and/or storage for future use.
Only selected input data and output metadata to be provided to the users.
Output: selection of characterised (through metadata) and stored dataset of interest, which will be available to the users. The output will be a subset of the input data. No storage of input data is needed.

2.4.4.2 Detection of plumes and sources from L2 products

Use Case I: systematic processing on selected regions of interest.

Use Case II: On demand processing: product type, region and period of interest to be specified by the users.
Input products and output meta-products to be provided to the users, over the specified region/period.
Output: characterisation of the input dataset (through metadata) by flagging the data associated to the plumes, and localising the corresponding sources, which will be available to the users. Storage of the input data and output metadata is envisaged.

The foreseen implementation is:

- Setting up an access to the software, data and processing facility through a Jupyter notebook for on-demand runs.
- Providing an interactive web interface to trigger the processing,
- Providing a RestAPI,
- Providing a WPS system.

2.4.5 Expert specific usages

Use Case III: interactive processing of both L1 PCA-based screening and L2 plume/source detection by experts. To be specified.

Implementation could be:

- Setting up an access to the software, data and processing facility through a Jupyter notebook for on-demand runs.

2.5 Technical framework

2.5.1 Processing language and access (for run by WP3)

PYTHON framework with FORTRAN or C codes embedded (for performance purpose)

The source codes will be available on the AERIS gitlab server: https://git.icare.univ-lille1.fr where an account for the CINES team will be created.
The project URLs will be provided later (projects not created at this time).

2.5.2 Input data catalogue and access (for run by WP3)

The datasets metadata will be available in the AERIS catalogue https://www.aeris-data.fr/catalogue/. It is provided in a JSON format and one example can be viewed here: jsoneditoronline.org/?url=https://sedoo.aeris-data.fr/catalogue/rest/metadatarecette/id/076096c4-c057-3c99-9ba3-d0f638dee3d0

2.5.3 Product/services Catalogue and access

TODO: define the outputs
As for the inputs, the output products catalogue will be available through the AERIS catalogue.

2.5.4 Protocols

Dataset descriptions will be given available through the PHIDIAS catalogue Rest API

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

3 Land Surface use case

3.1 WP description: Big data Earth Observations: processing on-demand and products dissemination for environmental monitoring

Optical and radar images observing Earth land 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. THEIA scientific expertise centres have design and implemented prototypes of new algorithms that meet the information needs of environmental monitoring stakeholders, such as land cover, soil moisture, natural vegetation biomass, the evolution of the artificial task combining VHR optical and radar sensors images.

Based on extended HPC environment coupled with an architecture that allows access to massive storage capacity delivered by WP2, the objectives of this WP is to enhance the scalability of EO data processing chains, disseminate in FAIR manner the mapping products for environmental monitoring coming from the end-users needs of THEIA land data centre network. The aim is to be able to:

1) Have interoperable access to catalogues of HR and VHR Earth Observation (EO) data from different sensors (Sentinel 1 and 2, SPOT) **2) Ensure massive and on-demand** EO data processing of mapping products for environmental monitoring and the dedicated web user environment, **3) Provide standardised services for the discovery and access to the information produced** in order to allow its open and interoperable dissemination according to the FAIR principles.

This activity is composed of the following tasks:

<u>Task 5.1: EO data processing chains for massive and on-demand execution</u>

It consists in optimizing (parallelizing) the THEIA data land processing chains so that they can take advantage of HPC environments and enable production in all or part of Europe (e.g. Mediterranean perimeter). Metadata production within these chains will also be specified and implemented to ensure the description of maps produced for discovery and open dissemination.

<u>Task 5.2: UI web environment dedicated for on-demand execution</u>

The current GEOSUD VHR images on-demand processing portal will be enhanced to allow the configuration and execution of on-demand processing chains on Sentinel temporal series. For this purpose, it will provide the discovery over sentinel images catalogue and available processing with the capability to filter images taking into account the processing constraints execution (format or characteristics of processing inputs).

<u>Task 5.3: Data workflows for discovery, access of EO raw data and products</u>

Starting from the existing GEOSUD data standardization pipeline (SPOT67), this task will consist in specifying and then implementing the Sentinel 1 & 2 data standardization pipeline needed to produce the maps. It will also specify the standardisation pipeline that will allow the products from the data processing workflow to be share (discovery, visualization, access) in an interoperable manner (in compliance with the INSPIRE directives).

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

3.2 Objectives of WP5 Use Case

The objective of this use case is to provide users in the academic and land management communities with an interactive environment to ensure the systematic or on-demand production of new knowledge useful for the environmental monitoring of territories. This involves, on the one hand, relying on the algorithmic developments carried out within the THEIA CES, deep learning techniques adapted to spatial data and, on the other hand, taking advantage of the complementarities (spatial and temporal resolution) of the large sets of spatial data from VHR sensors (SPOT, PLEIADES) and SENTINEL 1 and 2 time series. In order to share and reuse data and algorithms by the entire environmental community, hardware and software environments deployed shall allow:

- interoperable access to the catalogs of SENTINEL 1 and 2 images provided by the PEPS platform and SPOT6 images provided by the GEOSUD infrastructure,

- interoperable access to algorithms or processing chains encapsulating the algorithms,

- to ensure massive production (on a portion of European territory) or at the request of maps,

- exposure of the information produced via standardized discovery and access services in order to allow open and interoperable dissemination according to FAIR (Findable, Accessible, Interoperable, Reusable) principles.

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

3.3 Description of Use Cases « Environnemental monitoring »

Use case #1: Sentinel-1/Sentinel-2-derived Soil Moisture product at Plot scale (S2MP) over agricultural areas

Description

The spatio-temporal monitoring of the soil moisture in agricultural areas is of a great importance for numerous applications, particularly those related to the continental water cycle. The use of in situ sensors ensure this monitoring but this technique is very costly and it can only be carried out on a very small agricultural area, hence the importance of the spatial remote sensing which now allows large-scale operational mapping of soil moisture with high spatio-temporal resolution.

Recently, the arrival of the Sentinel-1 (S1) Synthetic Aperture Radar (SAR) satellite provided users with free open access SAR data at a high spatial resolution (10 m x 10 m) and high revisit time (six days over Europe). The S1 mission from the European Space Agency (ESA) is a constellation of two polar orbiting SAR satellites (Sentinel-1A and Sentinel-1B) operating in the C-band (~5.4 GHz). The SAR data of the S1 mission at high spatial and temporal resolutions have encouraged mapping soil moisture in an operational mode.

El Hajj et al. (2017) developed an operational method to map surface soil moisture (SSM) at the plot scale over agricultural areas based on coupling S1-SAR data and Sentinel-2 (S2) optical data using the neural network technique (S^{2}MP). The S2MP maps are produced for summer-winter crops in agricultural areas and grasslands (it is not applied to vineyards and orchards). The French Land data center Theia (https://www.theia-land.fr/en/) use the algorithm developed by El Hajj et al. (2017) for the provision of soil moisture maps at the plot scale for several sites over the world (France, Italy, Spain, Morocco, Lebanon, etc.).

M. Hajj, N. Baghdadi, M. Zribi, H. Bazzi, « Synergic Use of Sentinel-1 and Sentinel-2 Images for Operational Soil Moisture Mapping at High Spatial Resolution over Agricultural Areas », Remote Sensing, 9(12), 1292, 2017.

Target users and uses:

- **Scientific use**: especially in the context of modeling meteorological (meteorological) processes, components of the water cycle hydrologist (hydrologist) or even a plot-based crop development model (agronomist).
- **Agricultural sector use**: the data is also used to map irrigation activities for agricultural cooperatives for example

Production mode in the frame of PHIDIAS project

Co-financed by the Connecting Europe Facility of the European Union
The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

- **Systematic production mode (by default):** systematic production on a determined territory (to be defined) with a frequency to be determined ten days, twice a month, monthly, seasonal not necessarily in all seasons, privileged spring and summer..

- **Interactive mode:** an on-demand processing mode can also be considered. This mode of production will ask the user to specify a spatial envelope (territory to be mapped), the period (start date, end date). If the request is made in an area in which the Sentinel images or the plot are not available on the planned data sources (PEPS, RGP), it will then be necessary to provide alternatives

- **Virtual product mode**: user pre-define a territory, an area on which the chain will be executed

Input data

Type	**Format**	**Data volume**	**Access**
Sentinel 1 - Level-1	Native ?	100 1 GB zip files.	https://peps.cnes.fr/rocket/#/home
Sentinel 2 - Level-3		A dozen of 2 GB zip files.	https://theia.cnes.fr/atdistrib/rocket/#/home
SRTM	SRTM HGT 1 arc second resolution	5MB per tile de 1°(110km)	https://dds.cr.usgs.gov/srtm
Parcellaire agricole	Shape file	4 GB for France coverage	**France** : https://www.data.gouv.fr/fr/datasets/registre-parcellaire-graphique-rpg-contours-des-parcelles-et-ilots-culturaux-et-leur-groupe-de-cultures-majoritaire/# [1] **Out of France** : Corine Land Cover & WMS service from Copernicus: https://land.copernicus.eu/pan-european/corine-land-cover/clc2018 RPG europe LPIS équivalent français

[1] In French

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Output data

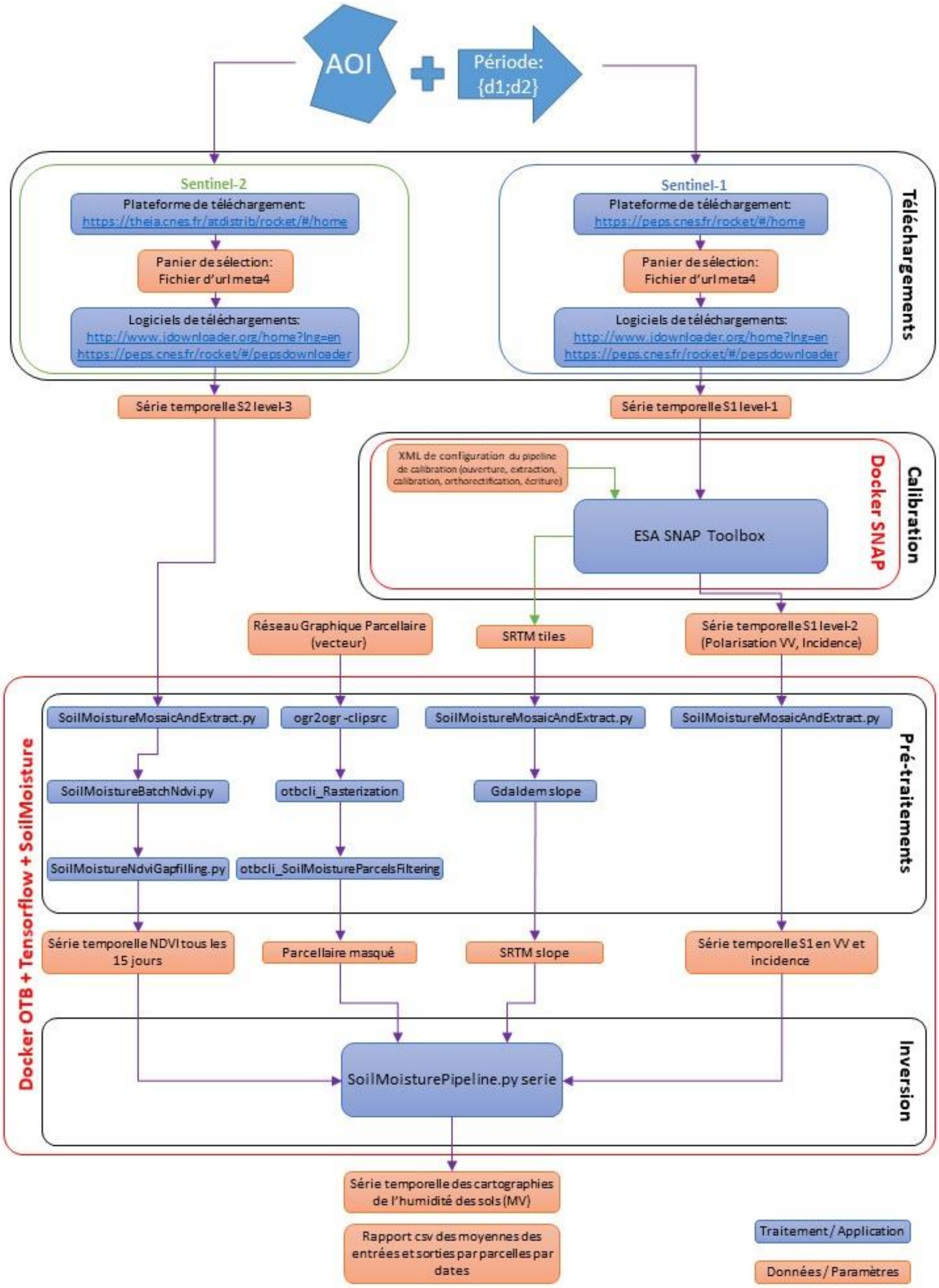
Nature	Format		Envisaged access
Soil moisture per parcel	Geotiff	Depends on the size of the study area, roughly: 10MBx 100aine/an /1000 km2.	vector. Must be exposed via ES, CSW, STAC or other metadata access API, WCS, WMS, compressed http tar.gz
Average input / output per parcel	CSV	negligible	Must be exposed and linked to the corresponding mapping in http, O&M, others?

Libraries and digital resources (storage and computing architecture)

- Python Scripts using: Orfeo Toolbox (OTB) et OTBTF (OTB TensorFlow)
- ESA SNAP Tool box (SeNtinel's Application Platform)

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.


AOI
Période:
{d1;d2}
Sentinel-2
Plateforme de téléchargement:
https://theia.cnes.fr/atdistrib/rocket/#/home
Panier de sélection:
Fichier d'url meta4
Logiciels de téléchargements:
http://www.jdownloader.org/home?lng=en
https://peps.cnes.fr/rocket/#/pepsdownloader
Sentinel-1
Plateforme de téléchargement:
https://peps.cnes.fr/rocket/#/home
Panier de sélection:
Fichier d'url meta4
Logiciels de téléchargements:
http://www.jdownloader.org/home?lng=en
https://peps.cnes.fr/rocket/#/pepsdownloader
Téléchargements
Série temporelle S2 level-3
Série temporelle S1 level-1
XML de configuration du pipeline de calibration (ouverture, extraction, calibration, orthorectification, écriture)
ESA SNAP Toolbox
Docker SNAP
Calibration
Réseau Graphique Parcellaire (vecteur)
SRTM tiles
Série temporelle S1 level-2 (Polarisation VV, Incidence)
Docker OTB + Tensorflow + SoilMoisture
SoilMoistureMosaicAndExtract.py
ogr2ogr -clipsrc
SoilMoistureMosaicAndExtract.py
SoilMoistureMosaicAndExtract.py
SoilMoistureBatchNdvi.py
otbcli_Rasterization
Gdaldem slope
SoilMoistureNdviGapfilling.py
otbcli_SoilMoistureParcelsFiltering
Pré-traitements
Série temporelle NDVI tous les 15 jours
Parcellaire masqué
SRTM slope
Série temporelle S1 en VV et incidence
SoilMoisturePipeline.py serie
Inversion
Série temporelle des cartographies de l'humidité des sols (MV)
Rapport csv des moyennes des entrées et sorties par parcelles par dates
Traitement / Application
Données / Paramètres

1 - Processing chain as a whole

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Use case #2: Very high resolution land use mapping (Moringa)

Description

CIRAD is working within THEIA's Center for Scientific Expertise in Soil Occupation to develop land use mapping methods adapted to the contexts of Southern countries. These landscapes are often dominated by family farming and characterized by specificities which limit the performance of methodological approaches adapted to European landscapes. Today, the significant changes that have occurred in the satellite imagery offer make it possible to consider solutions to meet the needs for the systematic production of land use mapping, for the management of territories. UMR TETIS is developing the Moringa processing line prototype at several study sites. The methodology consists of the **joint use of a VHR image (Spot6 / 7 or Pleiades) and one or more time series of optical images at HR (Sentinel-2 and / or Landsat-8) in an approach of classification combining OBIA and classification by the Random Forest machine learning technique trained by a training database made up of in situ surveys completed by photo-interpretation**. In order to make the method more easily reproducible for a future integration into the iota2 platform, its implementation is carried out only with free tools (OTB and Python).

Target users and uses

- **Scientific use:** input of simulation models of flows between underground / surface / atmosphere, etc.
- **Territory manager:** monitoring changes in land use

Production mode in the frame of PHIDIAS

Chain execution hypothesis: iota 2 would be executed and the resulting data would be stored close to processing (or retrieved from another data center?). These data would serve as the basis for an on-demand application restricted to a reduced area (department, etc.) on which Moringa (THRS part) would be executed, taking in particular the variables from the iota 2 chain as input.

- **Interactive mode:** this mode seems the most appropriate because it is necessary to have exogenous data (in situ) and to ensure photo interpretation

- **Expert mode:** possible, this would provide a test environment to ensure the configuration of OBIA methods (choice of segmentation methods, for example) and that of Random Forest learning

Input data

Type of data	Format	Data volume	Access

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Co-financed by the Connecting Europe Facility of the European Union

Iota 2 processing chain products	Geotiff	6 GB : France coverage & 1 year	Http download (free).
SPOT GEOSUD images	Dimap 1 & 2	10 to 20 GB : for one French department	• CSW • WCS (needs authentication) • WMTS (needs authentication) • HTTP • Elastic Search
Image LandSat-8	Geotiff + TXT + MTL	500 MB to 1 GB : one French department	• http://earthexplorer.usgs.gov/ • API Python • EarthExplorer API (non exhaustif)
In situ data	various	unknown	Contributed by the user

Output Data

Nature	**Format**	**Volume**	**Possible access**
Very high resolution land use (variable class depending on the nature of the LC / LU)	vector or raster format	Whether spatial extension	vector. Must be exposed via ES, CSW, STAC or other metadata access API, WCS, WMS, compressed http tar.gz

Libraries and digital resources (storage and computing architecture)

- OTB
- Python

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Use case #3: Remote sensing images processing with artificial intelligence: application to land cover mapping and super-resolution

Description

TETIS will provide an application based on open source libraries (Orfeo ToolBox, TensorFlow) able to apply any deep network on various remote sensing imagery. TETIS proposes to show its approach in applying it to two considered use cases:

UC #3.1: **Super-resolution of Sentinel-2 images**. A deep network is applied to the Red, Green, Blue and Near-infrared channels of the Sentinel-2 images (having 10 meters physical spacing) to derive an enhanced image at 1.5 meter physical spacing. The network will be trained using Spot images from Geosud/Dinamis and Sentinel-2 images from THEIA.

UC #3.3: **Processing of multi modal imagery for land cover mapping** using semantic segmentation. The approach consists in using two kind of images (very high resolution images from Spot, and low resolution time series from Sentinel-2) to derive a state of the art landcover mapping at very high resolution.
Both applications can be either applied on-demand though web processing services or integrated in a large scale production process. However, the super-resolution application is particularly suited for on-demand processing: typically, a user could request the super-resolution of a Sentinel-2 image over a specific region of interest at a given date.

Target users and uses

- • **Scientific use:** we are in a mainly experimental mode of use

Production mode in the frame of PHIDIAS

- • **Interactive mode and expert mode:** on the two use cases UC 3.3.1: Super-resolution of Sentinel-2 images and for which the SPOT and Sentinel images must be sought for the zone and the date or period considered. To see if it is necessary to obtain other characteristics of choice (cloud, incidence, level of production, ...)

These use cases can also be used to ensure production in:

- **Systematic mode**: UC #3.1 : Processing of multi modal imagery for land cover mapping
- **Virtual product mode**: #3.2 : Super-resolution of Sentinel-2 images

Input data:

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Type of data	**Format**	**Data volume**	**Access**
Sentinel 2 - Level-3	Given by Muscate	approximately 110Gb per tile*year.	https://theia.cnes.fr/atdistrib/rocket/#/home
SPOT GEOSUD images	Dimap 1 and 2	2Tb per year for the entire France mainland	• CSW • WCS (under authentication) • WMTS (under authentication) • http • Elastic Search

Output data:

Typz of data	**Format**	**volume**	**Possible access**
Sentinel THRS 1.5m	Geotiff	2Tb per year for the entire France mainland	vector. Must be exposed via ES, CSW, STAC or other metadata access API, WCS, WMS, compressed http tar.gz
VHR land use	Geotiff	4Tb for use case #2 in production scenario	Idem

Libraries for the processing chains

- Orfeo Toolbox (OTB) and OTBTF (OTB TensorFlow)

Processing resources

- GPUs computing architecture, equipped with hardware suited for deep learning (e.g. Tesla V100, GTX 2080Ti).

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

4 Ocean use case

4.1 WP description: Ocean use case

The general objective of this Ocean use case is to improve the use of cloud services for marine data management, data service to user in a FAIR perspective, data processing on demand, taking into account the European Open Science Cloud (EOSC) challenge and the Copernicus Data and Information Access Services (DIAS). In those terms, this use case can be seen as one of the prototypes, for marine environmental data, of the future Blue Cloud foreseen by the European Commission.

Task 6.1: Improvement of long-term stewardship of data

This task studies and proposes specifications to improve the long-term archiving of marine data. It will rely on the work conducted by WP2 for IT capacities and WP3 for metadata, and develop specific tools such as tools for controlling that actual data sets are in the described format (NetCDF using CF convention, Ocean DataView Spreadsheet) and described with the appropriate metadata (ISO 19115- INSPIRE compliant) and common vocabularies. The objective of this task is to progressively insure that marine data are preserved within procedures that can be certified by the Research Data Alliance (Core Trust Seal of Approval).

Task 6.2: Improvement of data storage for services to users

This task studies and proposes specifications to improve the storage of marine data to provide fast and interoperable access for dissemination purposes and their processing within dedicated high-performance computing environments

Task 6.3: Marine data processing workflows for on-demand processing

The **DIVAnd software** tool has been designed and is maintained by partner **ULiège** from Belgium. DIVAnd stands for Data-Interpolating Variational Analysis N dimensions and is a software tool for spatial interpolation of in-situ data and the generation of gridded field using an efficient finite-element solver.

This task consists in optimizing DIVAnd and the IT hosting context to provide on-demand computing to user. That includes pre-processing of data using programming language and online parametrization of DIVAnd. The Jupyter notebooks seems to be suitable as they combine:

1. Code fragments that can be run successively;
2. Text cells that describe the code and at the same time constitute the user guide and;
3. Figures or animations that illustrate different steps in the process, for instance: data extraction, duplicate removal, gridded and error fields.

The use case will use the following context as a demonstrator:

- SeaDataNet will provide a base input of hundreds of thousands of data sets with nutrients ((PO_4, Total Phosphorus, NO_2+NO_3, NO_3, Total Nitrogen, NH_4 and SiO_4), chlorophyll and dissolved oxygen as collected from multiple data providers for the Atlantic Ocean and for the Baltic Sea regions;
- Thereafter this data collection will be used as input for preparing interpolated maps of specific parameters in time and depth. Depending on sufficient spatial and temporal

Co-financed by the Connecting Europe Facility of the European Union
The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

data, maps might be produced for: Nutrients (Phosphate, Nitrate), Oxygen, and Chlorophyll-a. To generate these basin maps, the **DIVAnd software** will be used;
- Inter-comparison with satellite Ocean Color data will be conducted to detect and to supersede potential issues with the link with DIAS.

Task 6.4: Data inter-comparison, collection and visualization

Starting from the existing SeaDataCloud salinity data and products and the SMOS data available at the SMOS Data Production Centre, inter-comparisons will be conducted, including visualisation using Inspire compliant services, computation and analyse of gridded products. That implies several developments because up to now these data are not located in the same technical environments:

- Improvement of data storage for both data sources;
- Adaptation of inter-comparison software to the new storage;
- Set up of Inspire compliant visualization services on top of data products and data inter-comparison grids.

4.2 Use case description

4.2.1 Objective of the WP6

The general objective of the WP6 - Ocean use case is to improve the use of cloud services for marine data management, data service to users in a FAIR perspective, data processing on demand, taking into account the European Open Science Cloud (EOSC) challenge and the Copernicus Data and Information Access Services (DIAS).

Since the marine environment is evolving continuously, and because marine observation is still expensive, observation data are unique and must be well preserved and easy to be retrieved.

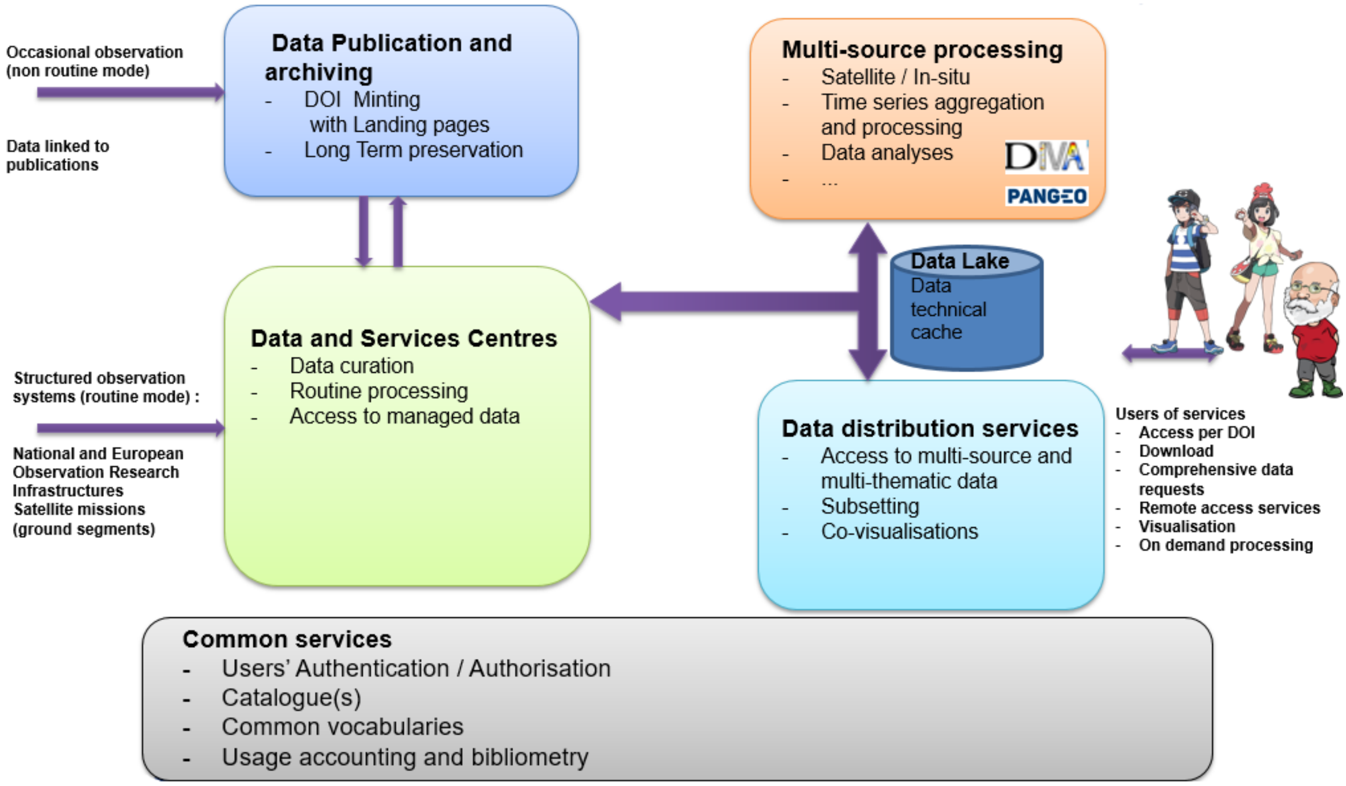
In practice, several functions (cf. Fig. below) have been implemented either in France, by the Ocean – Odatis cluster of the French Earth Observation Research Infrastructure DataTerra and, in Europe, by the SeaDataNet Research Infrastructure, the European Marine Observation and Data Network for in-situ data and the Copernicus Marine Environment Monitoring Service for Operational Oceanography and Satellite Data:

- Distributed Data and Services Centres manage data that are acquired in routine modes**.** Those data centres have established a direct interface with observation systems: satellite missions, in-situ observatories, research fleets in relationship with National Oceanographic Data Centres.

- Data that are acquired more occasionally ("long tail data") or data that require manual work at laboratories to be elaborated are often not directly interfaced with data centres. In order to facilitate the ingestion of these data in the marine data

Co-financed by the Connecting Europe Facility of the European Union
The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

management infrastructure, **publication/archiving services** have been provided to scientific teams.

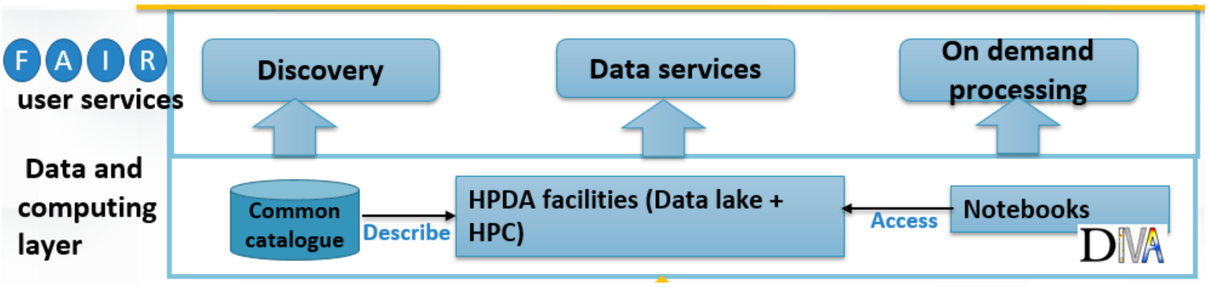

Occasional observation (non routine mode)
Data linked to publications
Data Publication and archiving
- DOI Minting with Landing pages
- Long Term preservation
Multi-source processing
- Satellite / In-situ
- Time series aggregation and processing
- Data analyses
- ...
DIVA
PANGEO
Data Lake
Data technical cache
Structured observation systems (routine mode) :
National and European Observation Research Infrastructures
Satellite missions (ground segments)
Data and Services Centres
- Data curation
- Routine processing
- Access to managed data
Data distribution services
- Access to multi-source and multi-thematic data
- Subsetting
- Co-visualisations
Users of services
- Access per DOI
- Download
- Comprehensive data requests
- Remote access services
- Visualisation
- On demand processing
Common services
- Users' Authentication / Authorisation
- Catalogue(s)
- Common vocabularies
- Usage accounting and bibliometry

General architecture of services provided by French and European Marine Data Infrastructures

Most of these services are already implemented by the French Ocean Data Management Cluster (Odatis) and by European Infrastructures such as SeaDataNet and EMODnet. However, upgrades are necessary to facilitate and to speed-up response time of data access and data browsing, to enlarge data management capacities, and to improve the long-term stewardship of marine data, especially for the long tail in-situ observations.

In this context, the technical objectives (cf. Fig. below) of the WP6 - Ocean Use Case are, by making an adapted use of HPC (high-performance computing) and HPDA (high-performance data analytics) capacities:

1. **Task 6.1: Improvement of long-term stewardship of marine in-situ data**
2. **Task 6.2: Improvement of data storage for services to users**
   - For two contexts:
     - ➢ 1- Fast and interoperable access for visualization and subsetting purposes (web portal)
     - ➢ 2- Parallel processing within dedicated high-performance computing
3. **Task 6.3: Marine data processing workflows for on-demand processing**



General sketch of the WP6 - Ocean use-case

In respect of those objectives, two main requirements have been identified about data storage:

1. Reinforce the long-term preservation capacities of observational data;
2. Manage a work copy of data with adapted structure in order to speed up and facilitate data retrieval and data processing. This working copy can be considered as a "**technical cache**" (called "**Data Lake**" in this document) that accelerates and harmonizes data requests and data processing, and may be consider as the **work copy of data**.

The need 1 – "Long-term preservation capacities" is described by the deliverable 6.1.1 - Specifications for long-term data archiving procedures in respects of RDA recommendations, since this deliverable will focus on the "Data Lake", which is the work copy of data.

The specifications of the need 2 (Data Lake) will be considered in this document.

Because, the marine data, especially in-situ observations, are managed in very distributed systems, it is difficult to provided services to users that require data from different data sources such as in-situ and satellite data or to assemble multidisciplinary datasets on a specific data areas. In this context, one of the ambitions of the present use-case is to facilitate the discovery and the assembling of data from distributed data management systems, in two cases:

- On line data distribution services such as:
  - Discovery, selection, assembling and subsetting of multi-sources data;
  - Visualization of data from different sources to allow visual comparisons.
- On-demand data processing:
  - Using Notebooks;
  - With Diva interpolation software;
  - With Pangeo Python components library for specific data browsing and analysis.

4.2.2 Managed data

Marine observations are made up of several data types, with different characteristics:

- Satellite data (large datasets, covering only the sea surface);

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

- In-situ observations (many small datasets, often with measurements below the sea surface).

And from distributed data repositories, such as:

- Operational oceanography observations that are, for example, stored in the DIAS Wekeo;
- Scientific observations that are, for example, stored using EUDAT services (SeaDataNet European infrastructure) or by the Data Centre themselves;
- Long tail data sources that are partly managed by online publication/archiving services.

Access conditions may differ from one data source to another one, for instance for some repositories, access is only granted to registered users (MarineID for SeaDataNet, Mercator Ocean International for Copernicus …).

Most of those data are recorded using one of these formats: NetCDF (following the CF-Conventions), Ocean Data View spreadsheet (CSV compatible + semantic header). Other formats must also be considered such as Geographical Shape Files (SHP) and imagery formats (TIFF, Geo-TIFF, JPEG and MPEG for the animations).

This use case will target only two specific areas: The North Atlantic (North-East for the Chlorophyll-a) and the Baltic Sea. These 2 regions represent 10 millions of observations in the North Atlantic and the Baltic Sea, accounting for a total of approx. 250 GBytes. However, the progresses achieved during the Phidias project will be extended to other data sources and data types as far as possible.

Important notice: It is considered that the long term-preservation of satellite data is out of the scope of this use case because they are under the responsibility of spatial agencies. Consequently, the requirements of long-term preservation described in this document will target only in-situ data and products derived from both in-situ data and satellite data.

The following data sets will be considered at a first stage by the use case:

4.2.2.1 In-situ Data

- SeaDataNet and EMODnet-Chemistry marine in-situ data collections, managed at EUDAT partner CSC (Finland), especially:
  - Temperature & Salinity
  - Chlorophyll-a concentration
  - Access conditions at https://www.seadatanet.org/Data-Access
- Copernicus in-situ data collections, managed by DIAS-WEKEO, Ifremer and FMI, especially:
  - Temperature & Salinity
  - Chlorophyll (BGC Argo & FerryBox)
  - Access conditions at http://www.marineinsitu.eu/access-data/

Co-financed by the Connecting Europe Facility of the European Union
The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

- Euro-Argo data managed by Ifremer, especially:
    - Temperature & Salinity
    - Chlorophyll (BGC Argo)
    - Access services:
        - ftp servers
          ftp://ftp.ifremer.fr/ifremer/argo
        - DOI (Data Object Identifiers)
          http://www.argodatamgt.org/Access-to-data/Argo-DOI-Digital-Object-Identifier
        - synchronization service (rsync)
          http://www.argodatamgt.org/Access-to-data/Argo-GDAC-synchronization-service
    - Thredds server API
      http://tds0.ifremer.fr/thredds/catalog/CORIOLIS-ARGO-GDAC-OBS/catalog.html
- Imaging FlowCytobot (IFCB): in-situ automated submersible imaging flow cytometer at Utö field station
- Long-tail observations managed by EMODnet Ingestion and SeaNoe online publication/archiving service

4.2.2.2 Remote Sensing Data

- SMOS Sea Surface Salinity products, managed at Ifremer
    - De-biased 10-day average & monthly salinity field products from SMOS satellite (mixed orbits)
    - Access services:
        - ftp server: ftp://ext-catds-cpdc:catds2010@ftp.ifremer.fr/
        - DOI: 10.12770/0f02fc28-cb86-4c44-89f3-ee7df6177e7b
    - 20 MBytes per day
- Sentinel-3 imagery, managed at ESA-DIAS (variables to be defined)
    - Sentinel 3 (OLCI) service
    - Two datasets: 1) Full resolution with 300 m spatial resolution and 2) Reduced Resolution is approximately 1.2 km on ground. Resolution 20m
    - access conditions at https://sentinel.esa.int/web/sentinel/user-guides/sentinel-3-olci

4.2.2.3 Output Data

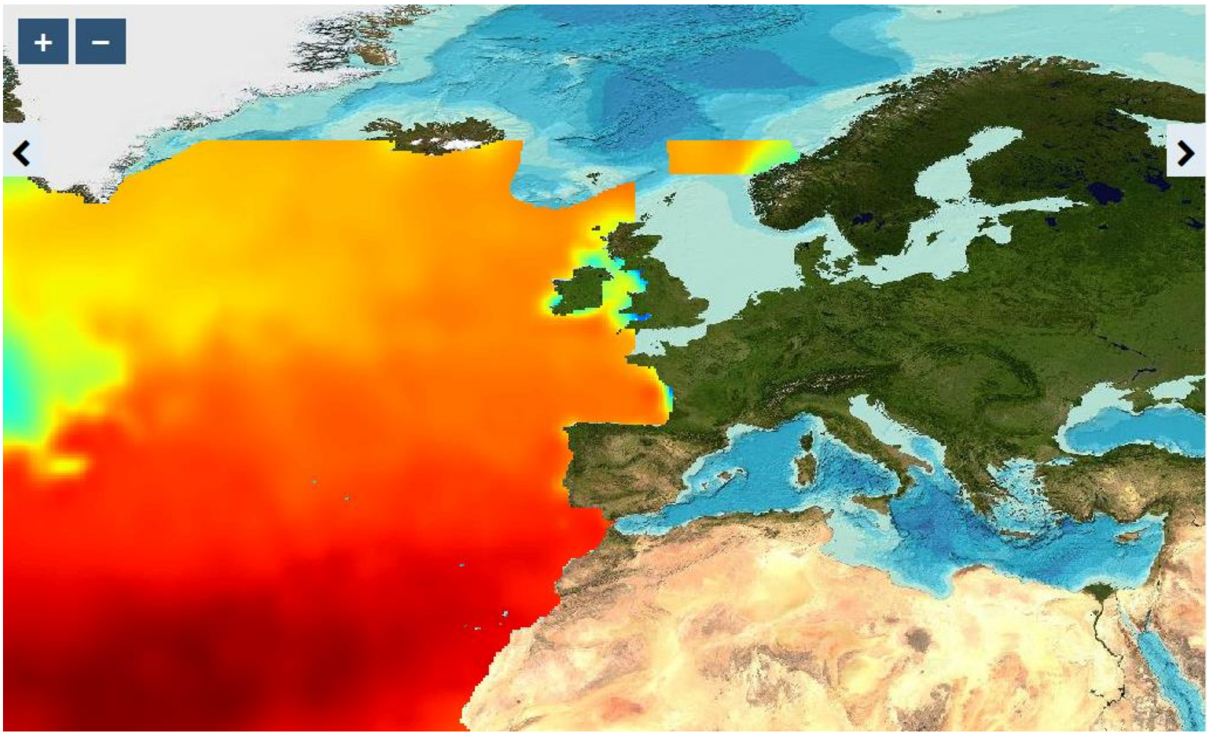
Products will be 4-D gridded fields (latitude, longitude, depth, time) or 3-D (fixed depth or fixed time) recording using NetCDF according to the CF conventions (Climate Forecast, http://cfconventions.org/).

Examples may be found at: https://www.seadatanet.org/Products#/search?from=1&to=30.

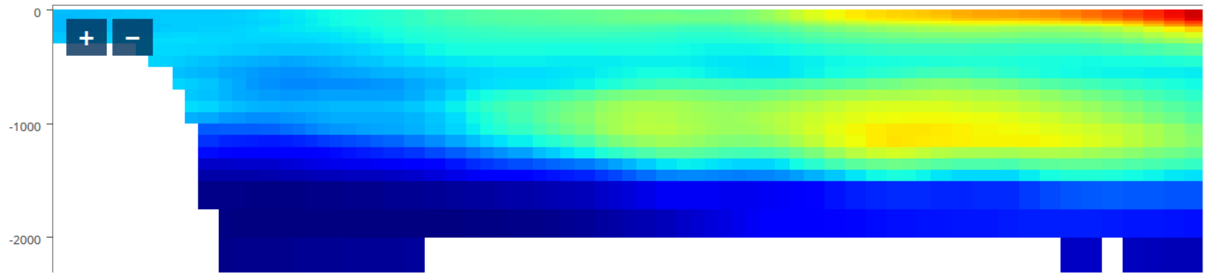
See also figure 3 and 4 below.

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.



North-East Atlantic Mean Salinity at surface in January (Ocean Browser – ULiege)



Vertical section of North-East Atlantic Mean Salinity at longitude 20°W, with Mediterranean water bodies (Ocean Browser – ULiege)

4.2.3 *Existing publication/Archiving services*

In France, the SeaNoe publication service (https://www.seanoe.org) has been set up for marine data in the framework of the Odatis cluster. In Europe, the EMODnet Ingestion service is also now available. The EMODnet ingestion service relies also on SeaNoe for marine research data.
Some other similar services also exist such as EUDAT services, Pangaea in Bremen-Germany (https://www.pangaea.de/), Zenodo (https://zenodo.org/), or the Dataverse software (***Dataverse.org***). However, these services are more generic and less dedicated to marine and earth observation sciences. For instance, Zenodo allows users not only to upload dataset, but also software codes, journal articles or presentations.

Co-financed by the Connecting Europe Facility of the European Union
The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

These publication/archiving services are in line with the usual recommendations such as:

- DataCite Metadata Working Group. (2016). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. doi:10.5438/0012.
- DataTerra, Groupe interpôle. (2016) : Bonnes pratiques autour de la citation.
- Research Data Alliance working Group: “Addressing the Gaps: Recommendations for Supporting the Long Tail of Research Data”

These services address the following requirements:

- Minting of permanent identifiers (DOI – Digital Object Identifiers),
- Management of metadata
- Generation of landing pages
- Preparation of archiving

More information about SeaNoe services may be found below and at https://www.seanoe.org/html/publish-your-data.htm.
These services are now certified as part of a certified repository by the Core Trust Seal (RDA & ICSU-WDS).

4.2.3.1 Permanent identifiers

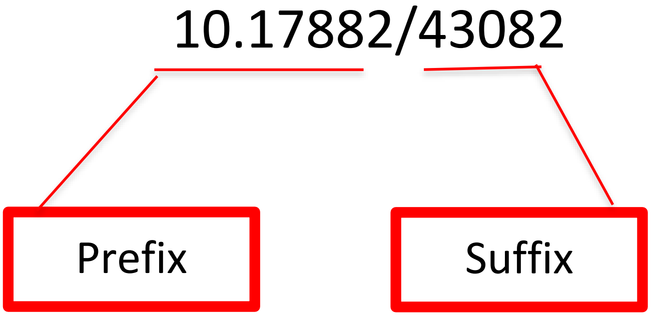
Permanent identifiers are minted in order to:

- Rationalize Data citation;
- Provide traceability on data usage.
- Simplify data access, especially for data sets that are linked to scientific publications.

The permanent identifiers in use are DOI – Digital Object Identifier, attributed by DataCite.

Minted DOI’s are not meaningful for human readers (only digits), in order to make them as persistent as possible. For example, they do not include any organization name as the organization name may change.

The used DOI syntax is:



The prefix is fixed and attributed to an organization (e.g. seanoe.org). The suffix is the identifier of the dataset within the organization.

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

In some cases, especially for managing version of constantly evolving datasets (cf. § "Preservation of evolving datasets" below) such as time series, fragments may be used (# symbol): 10.17882/42182#42350.

4.2.3.2 Metadata

Managed metadata are compliant with the Dublin Core standards. If relevant, for example for geo-referenced data, metadata are made compatible with ISO 19115 standard (e.g. by the addition of geographical extend…). Main managed information are:

- General metadata (Dublin Core)
    - Title
    - Author(s) and affiliations (link with ORC ID)
    - Publication date
    - Abstract
    - References
    - Use Conditions (Possible limitations…)
    - Reference to data user's manual (if any)
- Access conditions
    - Data Licence (Creative Commons license, ...)
    - Provided data citation according to a format suggested by DataCite: "Creator (Publication Year): Title. Publisher. Identifier"
    - Access service(s)
    - Data format and size
- Keywords (CodeList provided):
    - Variables (link with the Essential Ocean Variables Code List),
    - Method(s),
    - Instrument(s),
    - Project(s)
- Geographical extends
    - Min and Max latitudes and longitudes
    - Location map
- Temporal extends
    - Min and Max time
- Data preview(s)
    - Figures, maps, visualisation services
- List of associated datasets
- List of citing publication
- …

These metadata can be imported and/or harvested using various protocols such as OGC-Catalogue Service for the Web and OAI-PMH or downloaded in the RIS format. This format allows automatic import in bibliographic management tools (e.g. Endnote).

Co-financed by the Connecting Europe Facility of the European Union
The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Generating these metadata can be automated using tools. For example, the DIVA data interpolation tool is able to generate compatible metadata when generating interpolated fields (data products).

4.2.3.3 Published/Archived data

Using SeaNoe, data may be directly uploaded by data providers. Data centres perform checks in order to ensure that data match what has been described in metadata.

Checks include:

- Completeness of metadata in respect of data types;
- Data files are readable as described in metadata;
- Used vocabularies are the recommended vocabularies.

In case of detected non-conformance, additional information are requested from the data provider.

These checks may include format conformance tests. In the Phidias WP6-Ocean use case, only two file formats will be considered: NetCDF using CF-convention, Ocean Data View Spreadsheet (CSV compatible, including mandatory metadata headers). Specific reader software tools have been developed for those data formats.

Checksums are computed (SHA256 hash function) in order to ensure that data will not be modified.

4.2.3.4 Preservation of evolving datasets

Since data may not be modified after publication to ensure reproducibility of results, the preservation of evolving datasets such as data produced by running observation systems have been organized as follow:

- Snapshots of the datasets are periodically extracted from the evolving database (e.g. one snapshot per month);
- A DOI and the associated landing page is minted to the evolving dataset. Associated metadata describe the full datasets.
- DOI fragments are minted to the periodic snapshots and describe the temporal interval covered by the snapshot. The DOI fragment refers to the global DOI.

4.2.3.5 Preparation for long term preservation

The SeaNoe service relies on a catalogue hosted by a SQL database. However, to facilitate the long-term preservation of both metadata and data, the following file system structure (cf. Fig. below) has been adopted:

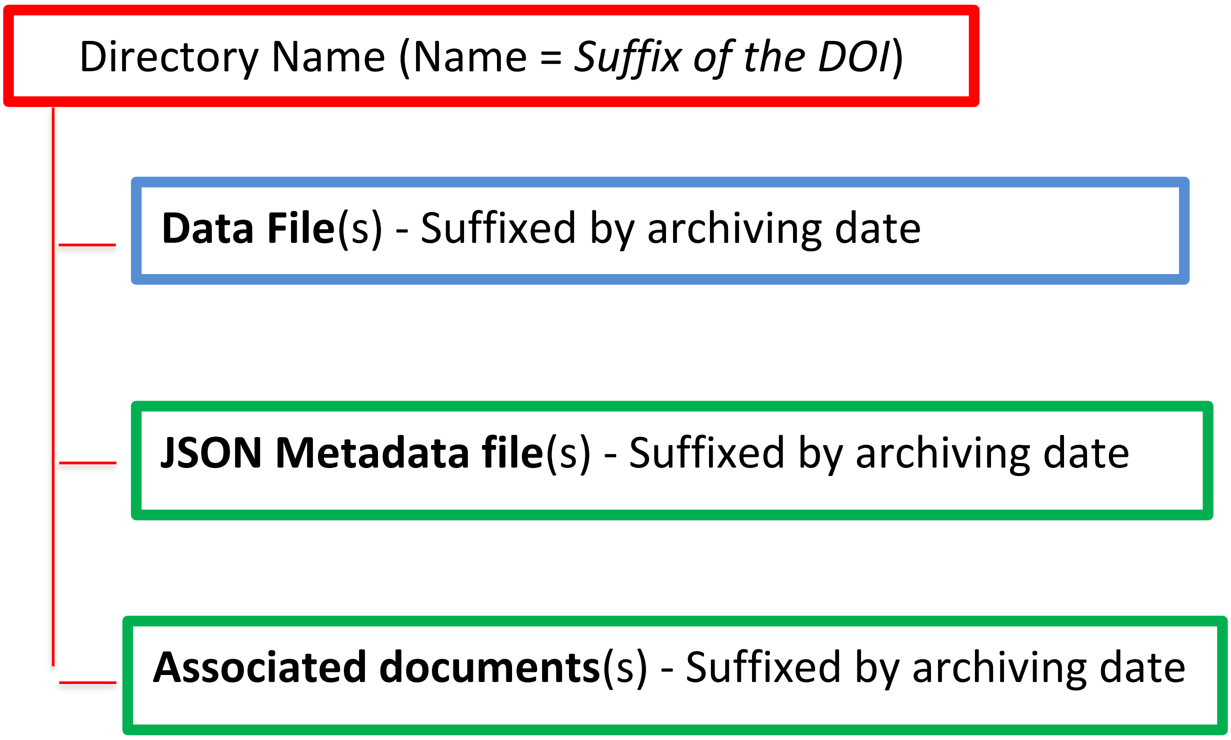
Co-financed by the Connecting Europe Facility of the European Union
The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

- One directory per published data set which includes:
- The dataset file(s) itself (themselves),
- The associated metadata formatted in JSON,
- Any document(s) referenced in the associated metadata (illustrations using image formats, texts using PDF format).

This directory is self-sufficient to describe and retrieve data and can be copied in a persistent data library.

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.


Directory Name (Name = *Suffix of the DOI*)

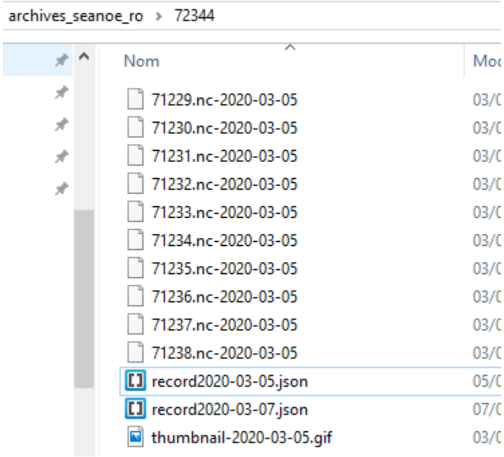
Data File(s) - Suffixed by archiving date

JSON Metadata file(s) - Suffixed by archiving date

Associated documents(s) - Suffixed by archiving date

FStructure of data files for data preservation

Example:


archives_seanoe_ro › 72344

```
Nom                              Moc
71229.nc-2020-03-05              03/(
71230.nc-2020-03-05              03/(
71231.nc-2020-03-05              03/(
71232.nc-2020-03-05              03/(
71233.nc-2020-03-05              03/(
71234.nc-2020-03-05              03/(
71235.nc-2020-03-05              03/(
71236.nc-2020-03-05              03/(
71237.nc-2020-03-05              03/(
71238.nc-2020-03-05              03/(
record2020-03-05.json            05/(
record2020-03-07.json            07/(
thumbnail-2020-03-05.gif         03/(
```

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

4.2.3.6 New requirements

The SeaNoe publication/archiving service is now well established and fits quite well with the need of marine data producers. Data management procedures have been set up in back office to check published datasets, and to notify data centres that are in charge of the related data type.

At present, these data management procedures are performed manually by EMODnet helpdesk (Odatis operated by Ifremer/Sismer in France) after each data publication:

- Check the thematic type of data and the Data Provider country, if permitted by the data licence granted by the data provider;
- Warn the thematic Data Centre in charge of this data type in France or in Europe;
- Provide access to the data files for the thematic Data Centre;
- Notify the data provider that submitted data are also transmitted to another data centre for integration in a collection, according the granted data licences

The formal identification of all data flows between data centres is now studied in the framework of EMODnet Ingestion service (“Pathways” task).

Consequently, new requirements that will be implemented as part of the use case are more oriented towards:

- Improving the scalability of the service;
- facilitating back-office exchanges between data centres in charge of related data types;
- securing long-time archive by curating several copies in several locations.

4.2.3.7 Service scalability

Due to the allocated storage resources, and due to the limitations of http protocol that is used for uploading datasets, the maximum allowed size of data sets is presently 0.2 TBytes.

To supersede this limitation, use of other protocols, that can be asynchronous, might be considered.

In addition, sharing the allocation necessary storage resources from different infrastructures may improve the total capacity of the systems.

It is foreseen to study how far the use of Virtual File Systems is relevant for those purposes.

4.2.3.8 Back office exchanges

One of the objectives of the SeaNoe service is to collect as much as possible long-tail data. As explained before, in the marine domain, long-tail data are very valuable because it is merely impossible to reproduce observations (environmental changes, cost of observation at sea).

Co-financed by the Connecting Europe Facility of the European Union
The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

As a consequence, it is of great interest to assemble these data in large “data collections”. This task has to be performed in routine mode by the Data Centres that are part of the French Ocean Cluster Odatis or by all data centres that are partners of the pan-European SeaDataNet marine data management infrastructure.

This task requires many data exchanges between the involved Data Centres. At present, this task is performed mainly manually.

Improving data exchange facilities between data centres and automatize them will be of great interest, by for example implementing iRODS data flows.

4.2.3.9 Securing long-time archive

At present, most of data safeguarding is performed by the Data Centres themselves, using they own technical infrastructures. These infrastructures are not always well adapted to this purpose, and dedicated teams for organizing and performing long term archives are not available everywhere.

Relying on professional long-term repositories (i.e. certified repositories by Core Trust Seal, *ISO* 14721 ...) seems to be extremely relevant.

Distributing, by e.g. using iRODS data flows, datasets that have to be archive in several geographically distributed repositories is also one of the measure that are recommended by long-term archive standards.

4.2.4 Specifications of the data storage (“Data Lake”)

The “Data Lake” will assemble data from several in-situ and satellite data sources, which may face different issues:

- In-situ datasets are not extremely large. However, managed data types are heterogenous: vertical profiles, times series, underway data... In addition, due to this heterogeneity, data are often managed in separate files or relational databases, that leads to very large amount of files (> hundreds of millions of individual observations). Browsing such a large number of observation is really inefficient.
- Satellite datasets may be very large (> several tens of petabytes at total), that leads to difficulties to transfer them over networks.

The “Data Lake” will be periodically synchronized (e.g. daily) with the Data Centres. Since it is impossible to transfer all data at each synchronization period, the Data Lake has to be persistent.

Important notice: The management of the original copy of the datasets assembled in the Data Lake will remain under the responsibility of the Data Centres. The Data Lake is considered as a work copy of the data.

4.2.5 Data structures within the Data Lake

Data structures within the data lake will have to be adapted to the targeted uses. Two main uses are targetted:

- Online selection and vizualization of data using a two-step discovery service via a common catalogue : 1) Selection of "Data collections" / Datasets, and then 2) selection of the subset of data of interest. This common catalogue will have to be stored. Access to data will have to be optimized to select and retrieve a small amount of data amoung a large number of data, using different selection criterions: geographical, temporal... This case will be mainly oriented towards in-situ datasets.

- On demand data processing of large data subsets using DIVA or Pangeo,

Providing storage infrastructures for these two requirements will probably necessitate two different data structures:

- One data structure adapted for selecting a few amount of data within a large number of data (e.g. NoSQL databases), especially for in-situ data;
- One data structure adapted for processing large subsets of data, in parallel mode if necessary (e.g. "Data Cubes" such as Xarray, Parquet...).

4.2.6 Storage of the common catalogue

- **Input** : Existing data indexes from SeaDataNet & EMODNET (CDI-Common Data Index, https://www.seadatanet.org/Metadata/CDI-Common-Data-Index, including planned development of an API in the framework of ENVRI-FAIR) and CMEMS in-situ data index (e.g. API https://fleetmonitoring.euro-argo.eu/swagger-ui.html)

- **Objective** : Interoperability of metadata between SeaDataNet (Common Data Index) and Copernicus Marine Services including fast detection of co-localized data, and improvement of long term preservation especially for "Long Tail" data using publication services (such as SeaNoe which is used by SeaDataNet / EMODNET).

- **Output**: 2-Step Discovery service for in-situ marine observations via a common catalogue : 1) Selection of "Data collections" / Datasets , and then 2) selection of the subset of data of interest. Storage of the common catalogue required (already partially managed with the help of EUDAT).

- **Prototyping** : Set up of the common catalogue and of a prototype user interface for selecting observations.

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

- **Storage** : The common catalogue will be stored as an **Elastic Search** NoSQL database, in order to allow faceting of the web selection portal, with optimized response time.

4.2.7 Data Lake for selection and visualization

- **Input** : Existing in-situ data sources from SeaDataNet, EMODNET, CMEMS, EuroARGO and the French Odatis Ocean data cluster. Satellite data are less used that way.

- **Objective** : Create and populate and adapted data structure within the "**Data Lake**" that facilitates and improves access to data (especially for in-situ data) for fast and interoperable access for visualization and subsetting purposes (web portal) : **"access few data among many data".**

- **Output** : "Small" extracted data subsets and web-based maps and diagrams (representation of time-series and of vertical profiles).

- **Prototyping** : set up of the Data Lake by implementing *NoSQL* Data base (e.g. Cassandra). This includes the synchronization procedures from distributed data sources to the adopted data structure within the Data Lake.

- **Storage** : The adopted data structure will be, in this case, a **Cassandra** NoSQL database, which is able to retrieve data using different selection criterions. Using the same data structure than the structure adopted for on demand-processing will be also considered and tested in order to discard the need of data duplication in several structures.

4.2.8 Data Lake for on-demand processing

- **Input** : Existing data sources from SeaDataNet, EMODNET, CMEMS, EuroARGO and the French Odatis Ocean data cluster, including both in-situ data and . Satellite data are less used that way.

- **Objective** : Create and populate and adapted data structure within the "**Data Lake**" that facilitates and improves browsing and processing of large amount of data (e.g. salinity and chlorophyll), preferably in parallel: **"access many data among many data".**

- **Output** : Gridded fields of Salinity and Chlorophyll.

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

- **Prototyping** : Data processing will be conducted using both DIVAnd software and Pangeo software component suite. These software tools will be used within notebooks for providing direct user interaction. The users will launch the processing on demand. DIVAnd will be adapted to be interfaced with the provided data structures in the Data Lakes. Tests will be conducted for exploiting HPC facilities (parallel processing using a compiled version of DIVAnd).

- **Storage** : The adopted data structure will be, in this case, a **"Data Cubes"** which are used to access data using Pangeo software components suite : e.g. zarr format, Xarray, Parquet, Arrow. In order to facilitate identification of available data sets from a Python Script, an Intake interface will be set up on top of the (https://github.com/intake/intake). Intake is a lightweight package for finding, investigating, loading and disseminating data.

4.3 Synchronisation mechanisms

Routine software will be developed and set up to synchronize source datasets with the "Data Lake". This sofware will have to manage:

- Identification of the new data or the data that have been modified or deleted from the last synchronization, using the common catalogue;
- Data transfer from data source to the Data Lake over the network;
- Reformatting of the datasets according to the data structure(s) adopted for the Data Lake.

Use of iRODS dataflows seems to be well adapted to synchronize source datasets and the Data Lake.

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Co-financed by the Connecting Europe Facility of the European Union

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.