PHIDIAS

Prototype of HPC/Data Infrastructure for On-demand Services

Cloud services for marine and oceanographic data access and data management

Peter Thijsse (MARIS) | June 4, 2020, 11:25 AM CEST

Co-financed by the Connecting Europe

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

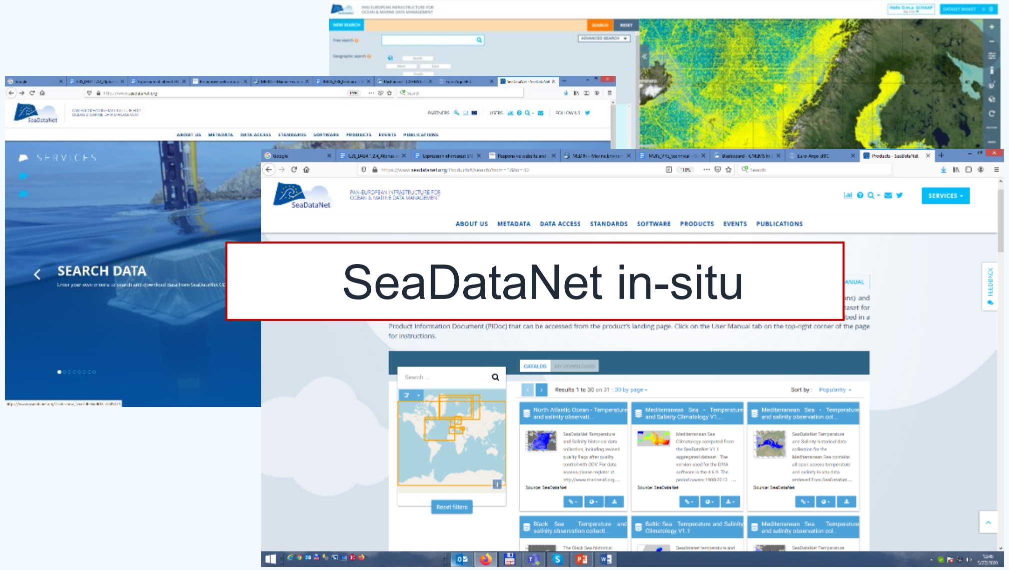
Outline

- Introduction
- Data resources in scope
- Discovery service
- Prototype Data Lake for processing

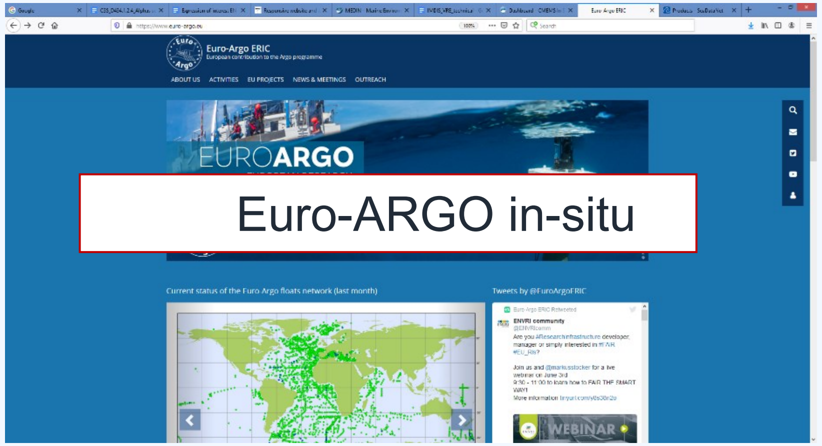
Main objective recap

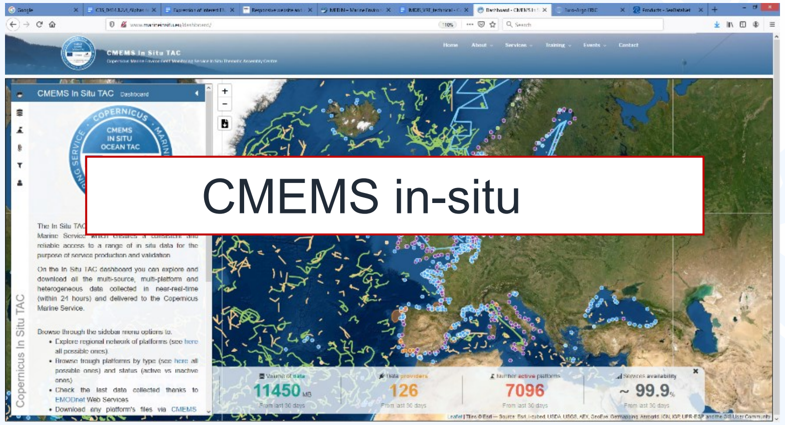
- to improve the use of cloud services for marine data management, data service to users in a FAIR perspective, data processing on demand, taking into account the European Open Science Cloud (EOSC) challenge and the Copernicus Data and Information Access Services (DIAS).

Marine data resources in scope


SeaDataNet in-situ
SeaDataNet
SERVICES
SEARCH DATA
ABOUT US METADATA DATA ACCESS STANDARDS SOFTWARE PRODUCTS EVENTS PUBLICATIONS
SERVICES
Reset filters

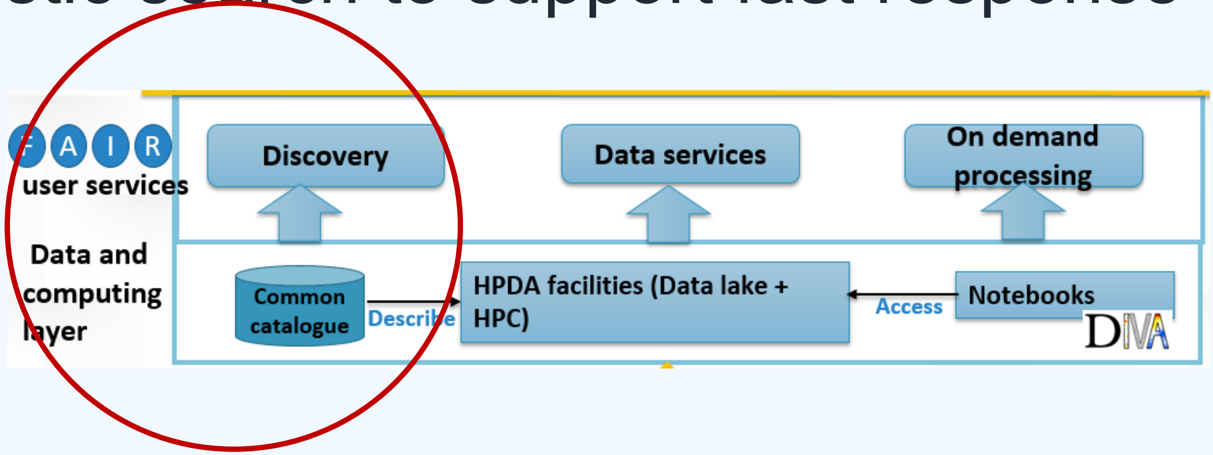
SMOS and Sentinel-3 Remote sensing


Euro-ARGO in-situ
Euro-Argo ERIC
ABOUT US ACTIVITIES EU PROJECTS NEWS & MEETINGS OUTREACH
EUROARGO
Current status of the Euro-Argo floats network (last month)
Tweets by @EuroArgoERIC
WEBINAR


CMEMS in-situ

PHIDIAS Discovery service

- Build up metadata indexes of available datasets
  - Metadata checks during import (completeness/readable/correct vocabularies)
- Include DOI's/PID's of the original datasets
  - New DOI's will be assigned for newly processed datasets (SEANOE)
- Use elastic search to support fast response on searches



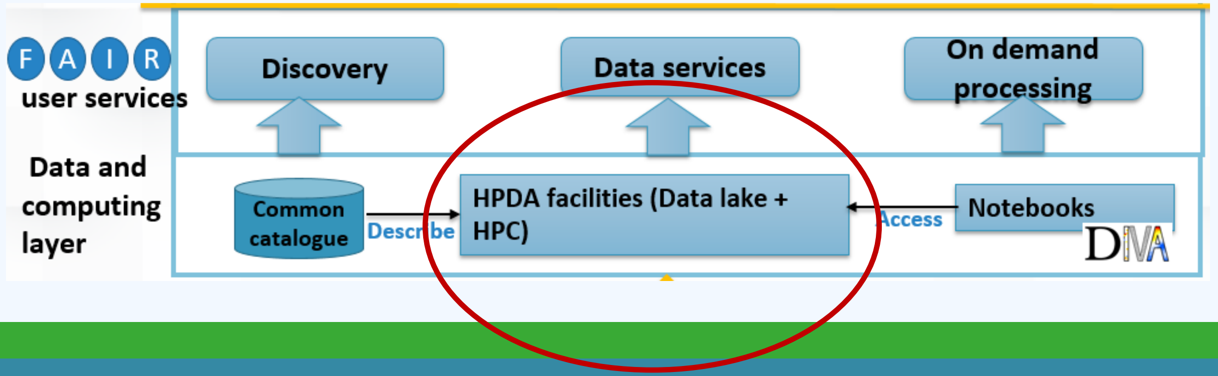
PHIDIAS Prototype of HPC/Data Infrastructure for On-demand Services

Metadata is important

- The PHIDIAS catalogue metadata model will be based on Dublin Core element (extended with ISO19115 if necessary):
  - 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 License (Creative Commons license, ...) | Provided data citation in DataCite format | Access service(s) | Data format and size
  - Keywords (CodeLists 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
  - Data preview(s)
  - List of citing publication
  - …

PHIDIAS Prototype Data Lake for processing

- Two data types:
  - In-situ datasets:
    - not extremely large, but in many small files.
    - managed data types are heterogeneous: vertical profiles, times series, underway data...
  - 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)



PHIDIAS Different use cases, different storage (1)

- For in-situ datasets - 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.
  - Example: Exploring SeaDataNet (Common Data Index) and Copernicus Marine Services data collections including fast detection of co-localized data
- Access to data will have to be optimized **to select and retrieve a small amount of data among a large number of metadata records**, using different selection criterions : geographical, temporal...
- Prototype: Elastic Search on top of (No)SQL database, in order to allow faceting of the web selection portal, with optimized response time.

Different use cases, different storage (2)

- Facilitate and improved 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).
- Prototype: 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.

PHIDIAS
Prototype of HPC/Data Infrastructure for On-demand Services

Different use cases, different storage (3)

- Support on- demand data processing of large data subsets using DIVA or Pangeo
- Requires high performance browsing and processing of large amount of data (e.g. salinity and chlorophyll), preferably in parrallel: “**access many data among many data**”.
- Output : Gridded fields of Salinity and Chlorophyll.
- Data lake prototype: “Data Cubes” which are used to access data using Pangeo software components suite : e.g. zarr format, Xarray, Parquet, Arrow.

Thank-you

Peter Thijsse (peter@maris.nl) and the PHIDIAS WP6 group

PHIDIAS

Prototype of HPC/Data Infrastructure for On-demand Services

www.phidias-hpc.eu
@PhidiasHpc
phidias-contact@cines.fr

CINES · NEOVIA INNOVATION · cnrs · MARIS · CERFACS

CSC · GEOMATYS · SYKE · IRD Institut de Recherche pour le Développement FRANCE

LIÈGE université · Ifremer · SPASCIA · Trust-IT Services Communicating ICT to markets