Argo GDAC web based services

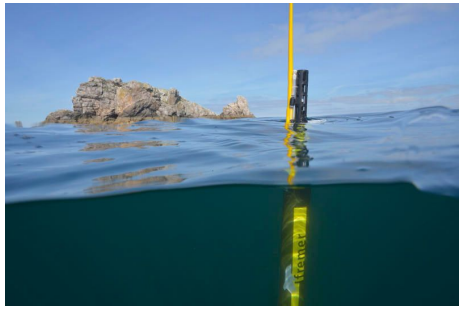
FAIR services for Argo observations

ADMT20 13th October 2019

Argo GDAC organization

A continuously managed unique and global dataset of Argo data, metadata and technical data

- **2 GDACs and 11 DACs**

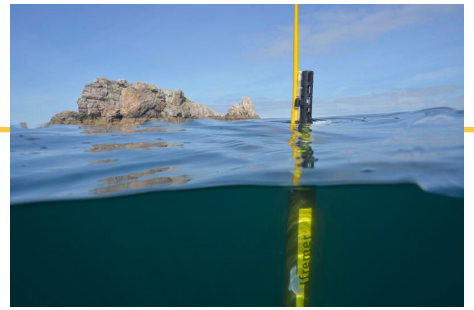


dac data assembly centres
aoml, coriolis, jma, meds, cls...

gdac global data centers
usgodae, coriolis

users
scientists, modelers, students…

Argo data flow



data sent via satellite

decode qc and format data for distribution

floats deployment and data certification

float data from satellite

4-12 hours

24 hours, real-time qc

gts : 24 hours

dm qc, few months

dac data assembly centres
aoml, coriolis, jma, meds, cls...

pi principal investigators
howard freeland, yves desaubies …

users
scientists, modelers, students…

24 hours

perform regional analyses

gdac global data centers
usgodae, coriolis

1 day to years

information about argo

long term archive

here is the master copy of argo data

1 month

regional centres
n. atlantic, pacific, indian

aic argo information center

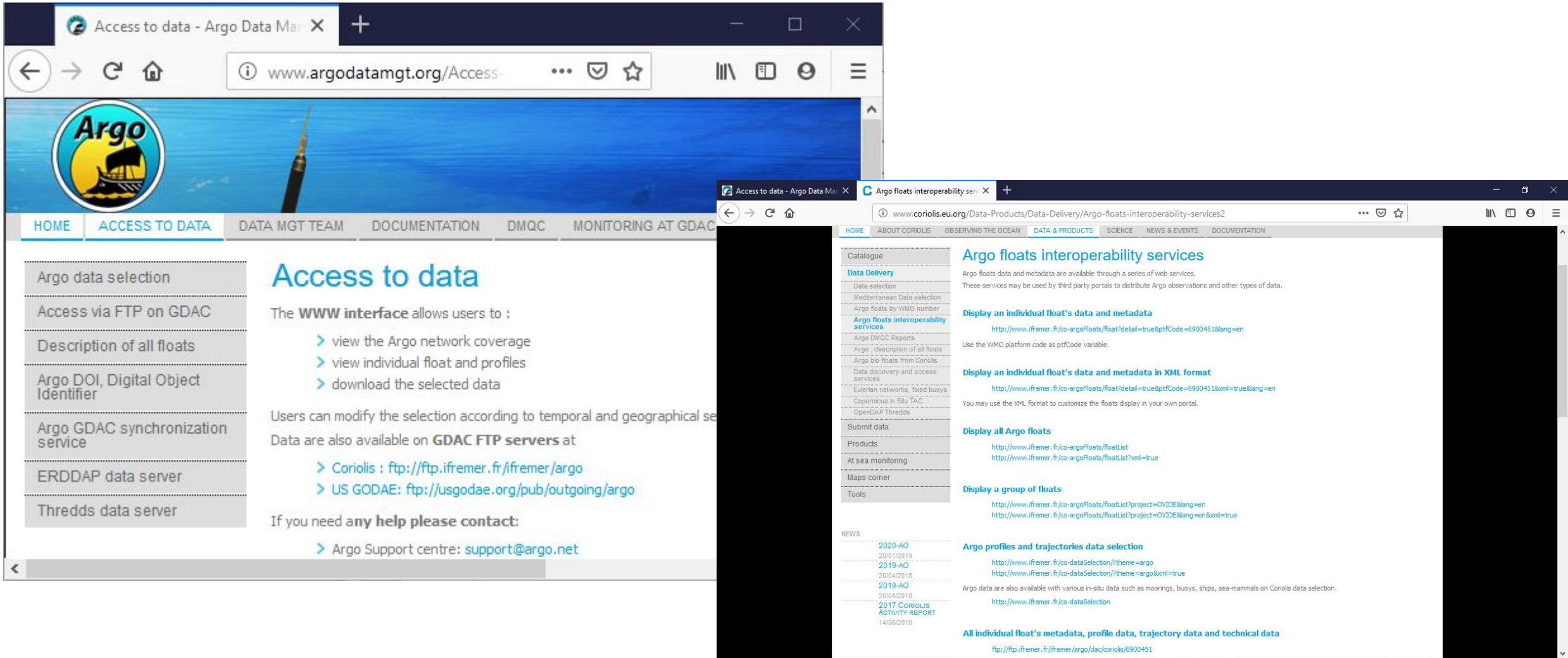
long term archive
usnodc

Argo dataset a collection of NetCDF files

- Argo dataset is a continuously managed collection of NetCDF files
  - Strict control of Argo NetCDF CF file format (https://doi.org/10.13155/29825) by the format checker (https://doi.org/10.17882/45538)
- Mirrored between US GDAC (GODAE) and EU GDAC (Coriolis)
- Distributed through different channels
  - Argo FTP server
    - ✔ Robust and popular file transfer service
  - Argo DOI
    - ✔ A one click download of Argo snapshots
    - ✔ Easy to cite
    - ✔ Cross references to bibliography, contributors, software, related datasets, documentation
    - ✔ Machine to machine catalogue services
      - How to cite
      - Content description, documentation
      - Licence (CC-BY 4.0)
      - Schema.org and JSON-LD content for SparQL queries
      - OAI-PMH harvesting protocol

Argo dataset a collection of NetCDF files

- http://www.argodatamgt.org/Access-to-data
  http://www.coriolis.eu.org/Data-Products/Data-Delivery/Argo-floats-interoperability-services2

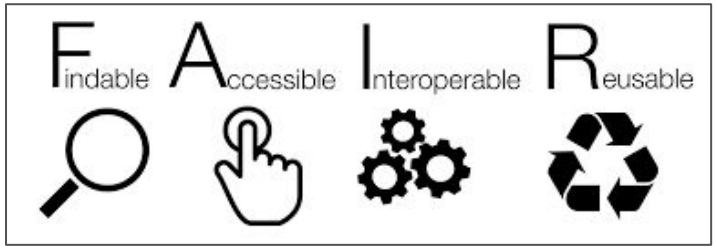


Argo GDAC data access

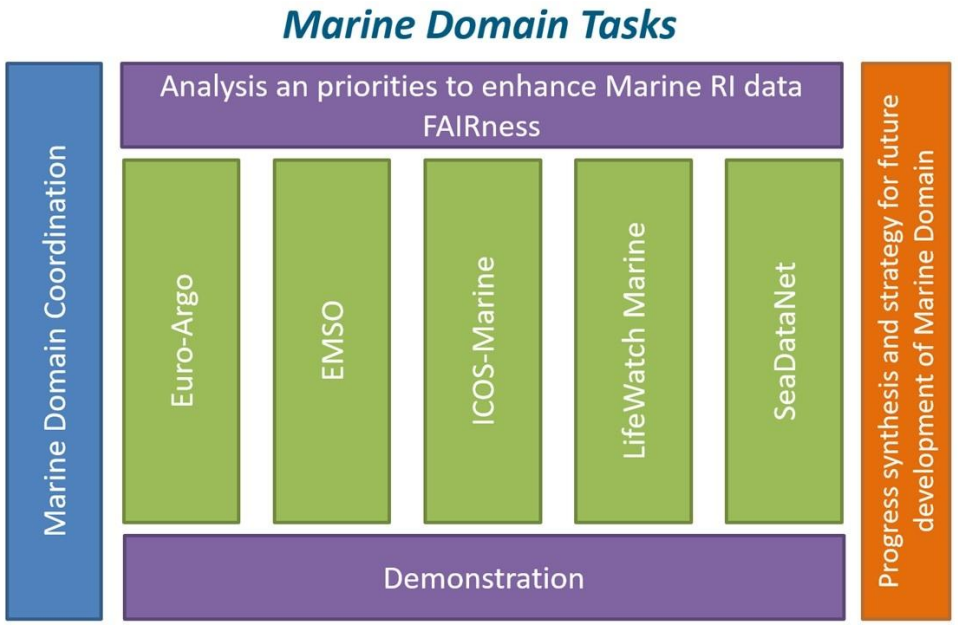
- GDAC ftp servers
  ftp://ftp.ifremer.fr/ifremer/argo
- GDAC DOI (Data Object Identifiers)
  http://www.argodatamgt.org/Access-to-data/Argo-DOI-Digital-Object-Identifier
- GDAC synchronization service (rsync)
  http://www.argodatamgt.org/Access-to-data/Argo-GDAC-synchronization-service
- GDAC Thredds server API
  http://tds0.ifremer.fr/thredds/catalog/CORIOLIS-ARGO-GDAC-OBS/catalog.html
- GDAC ERDDAP data server API
  http://www.ifremer.fr/erddap/tabledap/ArgoFloats.graph
- GDAC interactive data selection
  http://www.argodatamgt.org/Access-to-data/Argo-data-selection
- GDAC floats dashboard
  https://fleetmonitoring.euro-argo.eu/dashboard

ENVRI-FAIR [H2020 project, 2019-2022]

- ENVRI-FAIR goal is to implement the FAIR principles in ENVRI cluster (EU Research Infrastructures for Environment) and connect it to European Open Science Cloud (EOSC).



- Euro-Argo is co-coordinating the Marine Domain activities work package with EMSO *(European Multidisciplinary Seafloor and water column Observatory)*



FAIR Guiding Principles

DOI: 10.1038/sdata.2016.18

nature > scientific data > comment > article

SCIENTIFIC DATA

Comment | OPEN | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

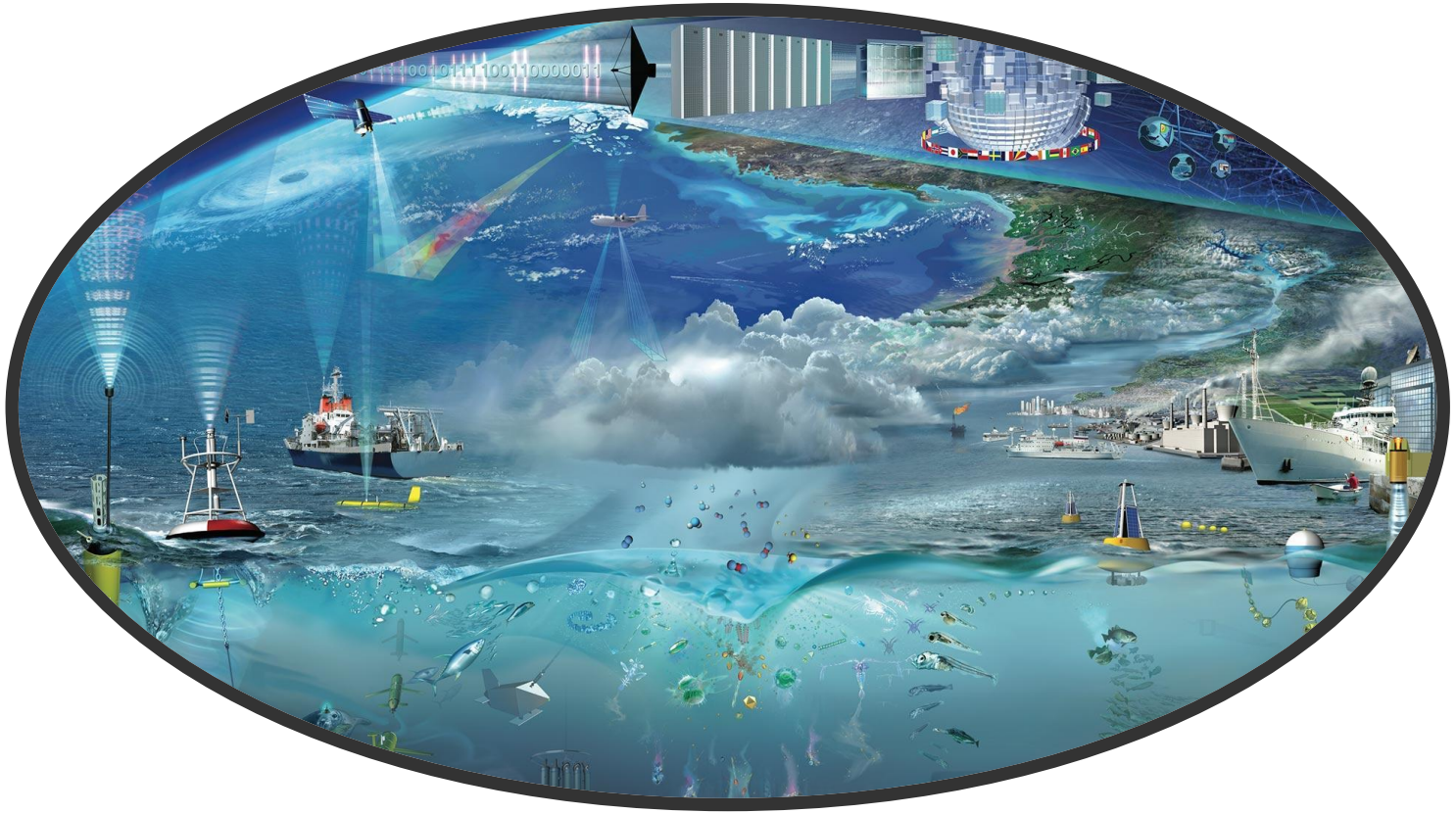
Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons - Show fewer authors

Scientific Data **3**, Article number: 160018 (2016) | Download Citation

"Data and services that are **f**indable, **a**ccessible, **i**nteroperable, **r**e-usable both for people and for machine."

Best Practices Supporting Data and Innovation

What are the benefits of FAIR?



For Ocean community:

Aggregate observations from multiple systems with machine to machine interactions to focus human efforts on analysis and knowledge

Image: https://www.atlantos-h2020.eu/ocean-observation/

What FAIR is not…

Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud DOI: 10.3233/ISU-170824

FAIR is not a standard

FAIR is not a semantic web / Linked Open Data Cloud

FAIR is not equal to 'Open' or 'Free'

Data are often Open (Access) but not FAIR

Some data can never be Open, yet be perfectly FAIR

By design, FAIR is not explicit about data quality, trustworthiness, responsibility, ethics, etc.

FAIR implementation will enhance Machine to Machine interactions

FAIR Guiding Principles

Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016)

Findable:

Easy to locate through a catalogue with search tools on data and metadata

Accessible:

Access through standardized protocols from an access point in the catalogue

Interoperable:

Easy to understand and compare to other dataset automatically

Reusable:

Easy to re-use and aggregate with other data by somebody who didn't participate to its acquisition or processing;

Argo Data System: pretty FAIR to People but not to Machines

Findable

A DOI for Argo dataset and monthly snapshot but only a person can identify a float cycle and a parameter

Detailed metadata but some are only described in the manual not in a Vocabulary server

Only part of the data are searchable either on Global Portals (GDACs) or at JCOMMOPS

Accessible

All data are available at GDACs identified in a DOI

All data are open and free so no authentication and authorization

FTP is not a FAIR protocol, need to know the data/metadata organization.
Thredds and ERDDAP are better but not yet robust enough to search on 20 years archive

Interoperable

NetCDF is self descriptive as long as metadata are described in Vocabulary services

Reusable

Difficult to understand the processing without the QC User manual
Need to enhance metadata on error estimation , level of QC

ENVRI-FAIR : development to enhance Machine to Machine FAIRness

Findable

Set up an Argo Vocabulary server that will be hosted at UK/BODC for all the tables of the User manual and Config/technical tables

Assign Permanent ID for an EOV acquired in a cycle

Accessible

Develop robust APIs using **big data** and **cloud** techniques

Interoperable

NetCDF + Argo **Vocab** will enhance the situation

Reusable

Enhance metadata on error estimation , level of QC interoperable with other networks

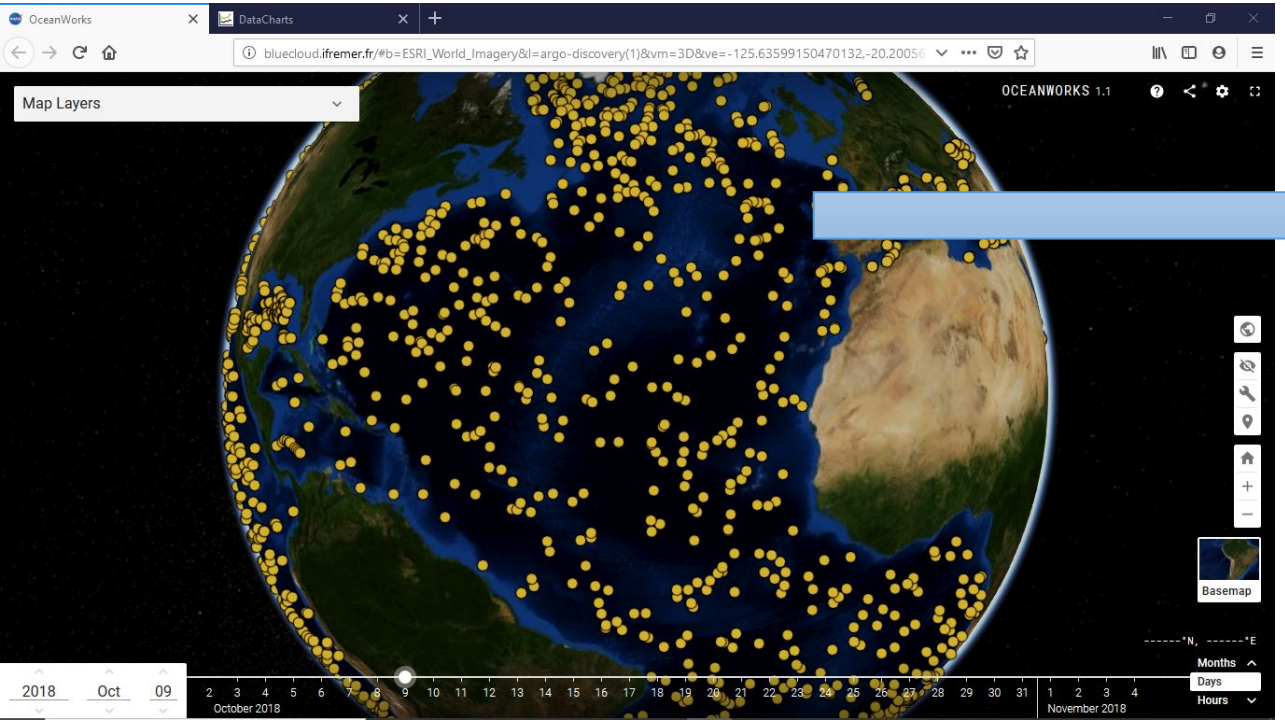
Use of Vocab is essential to facilitate aggregation with other data

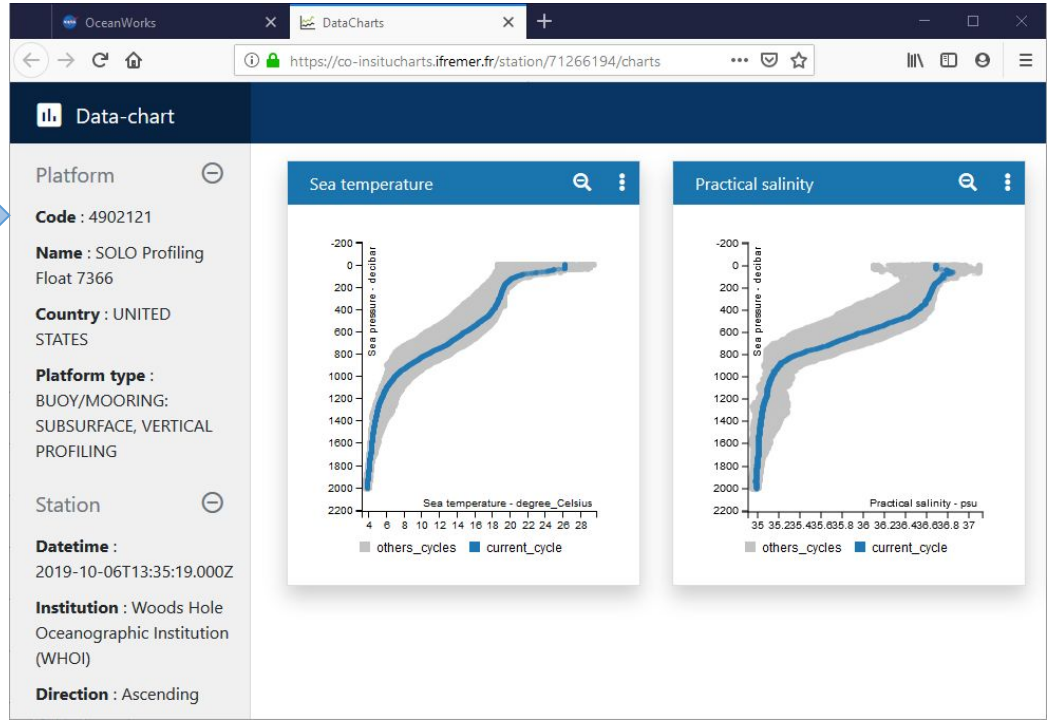
Euro-Argo-RISE and ENVRI-Fair for Data Management improvements

- Service to users
  - A survey will be conducted to identify what the European users require in term of Argo data services
  - The GDAC WWW site will be updated
    - ✔ with a new viewing and subsetting service from Argodatamgt www site
    - ✔ Robust API (OGC compliant) for machine access from GDAC and JCOMMOPS will be developed and promoted
  - Argo discovery tools at JCOMMOPS will be enhanced
  - Coherency between GDAC and JCOMMOPS metadata will be facilitated

EOSC Blue Cloud (European Open Science Cloud)

- Within EOSC, Ifremer is leading the Marine Domain activity
  - Argo data on EOSC cloud infrastructure (Openstack, Docker, Cassandra, Elasticsearch)
  - An informal collaboration with Nasa-JPL is underway to include in situ observation in the CMC (Common Mapping Client) and S-DAP (subsetting API)
    - ✔ http://bluecloud.Ifremer.fr





Findability

 The global Argo dataset is well findable. However, a rich and efficient search service at data level is missing; a local search engine should be developed.

Accessibility:

 Need for a rich and efficient local search engine (planned using Elastic search).

 On top of the collection of NetCDF files, API cloud services should be developed/enhanced and implemented for data discovery, visualization, download, subscription with rich sub-setting for:

o Individual scientist end users, with direct access to the original Argo NetCDF files, through a web GUI (Graphic User Interface)

o Virtual Research Environment with cloud data and metadata (API queries instead of NetCDF files transfers)

 Improvements in the field of subsetting access with OGC WMS and SOS services.

 Introduce semantic web capabilities through Linked data/SparQL endpoint development.

Interoperability:

 Schemas to be registered in common registry (e.g. on github). Apart from that implementing a data processing HUB: public and shared codes for tools such as data processing chains (Argo floats decoders), file format checkers, decimation or standardization services.

 Argo vocabulary tables are not yet implemented in the SeaDataNet vocabulary service, and not yet harmonised with other vocabularies.

Euro-Argo will implement a vocabulary server for Argo metadata to solve this. The provenance of data in the Argo Data System is underpinned by rich metadata which is standardised across the data system using vocabularies currently held in manuals and associated spreadsheets. The accuracy, controlled evolution and semantic value of this metadata will be further enhanced by migrating to a controlled vocabulary management environment and server compliant with W3C standards.

 The Argo vocabulary server will also lead to having all categories in the schema registered.

Reusability:

 Provenance info is now text only, should be coded and registered in the SeaDataNet Vocabulary service (see above).

Weakness 1: findability - no open query on Argo data
*Suggested approach: implement a **search engine** service*
Possible technologies: OpenSearch on top of an Elasticsearch metadata repository
Planning in time: POC in 2020 Q2, demo version 2021 Q1
Involved partners: Ifremer, …

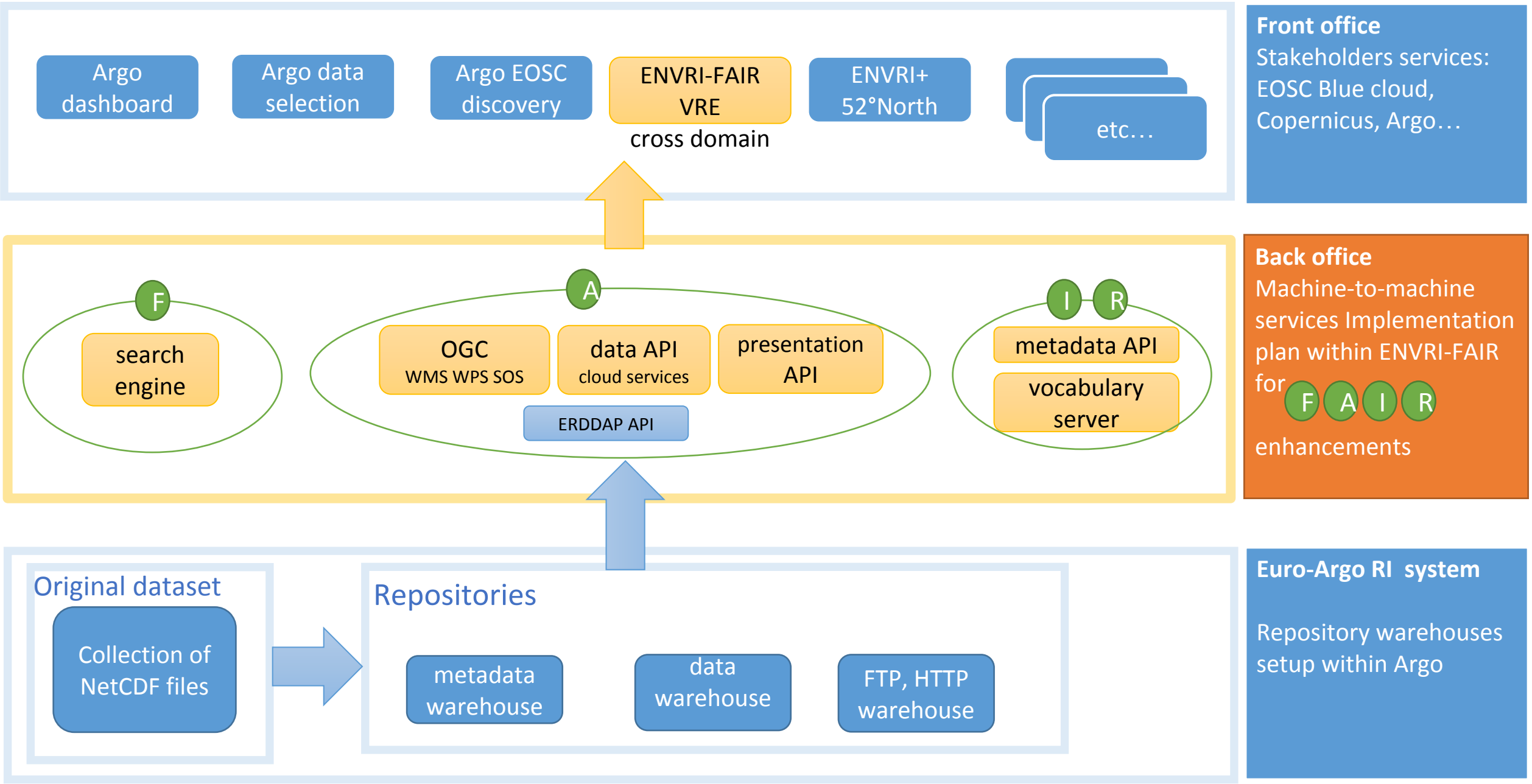
Weakness 2.1: accessibility – no machine to machine metadata access
*Suggested approach: implement a **metadata API***
Possible technologies: Elasticsearch or SOL-R on metadata repository
Planning in time: POC in 2020 Q2, demo version 2021 Q1
Involved partners: Ifremer, …

Weakness 2.2: accessibility – no machine to machine data access repository
*Suggested approach: implement a **data API***
Possible technologies: Cassandra, Parquet or HBASE on data repository
Planning in time: POC in 2020 Q2, demo version 2021 Q1
Involved partners: Ifremer, …

Weakness 2.3: accessibility – no machine to machine description of the above data and metadata API services
*Suggested approach: implement an **API description** service to facilitate the use of Argo APIs*
Possible technologies: swagger
Planning in time: POC in 2020 Q2, demo version 2021 Q1

Weakness 3: accessibility – not accessible through OGC protocols
*Suggested approach: implement the major **OGC services** on top of data and metadata APIs: **WMS** for map services, **SOS** v3 for data queries, **WPS** to activate filtering and subsetting of the **data processing HUB (ENVRI-FAIR VRE)****
Possible technologies: 52North or Geomatys (geoserver) solutions
Planning in time: POC in 2020 Q2, demo version 2021 Q1
Involved partners: Ifremer, …

Weakness 4: interoperability/re-useability – not interoperable with other RIs and scientific domains
*Suggested approach: implement a **vocabulary server** for metadata*
Possible technologies: skos queries on vocabulary server
Planning in time: POC in 2020 Q2, demo version 2021 Q1
Involved partners: BODC, Ifremer

(*)Data processing HUB (ENVRI-FAIR VRE) – demonstrated in the framework of Task 9.8 – EOV global product
Suggested approach: data processing on Cassandra repository
Possible technologies: Scala, python, Spark, SparQL query languages on top of Casandra or parquet repositories
Planning in time: POC in 2021 Q1, demo version 2022 Q1
Involved partners: Ifremer, …

ENVRI-FAIR, Argo RI implementation plan