Marine Data Value Chain

Simona Simoncelli

email simona.simoncelli@ingv.it
orcid.org/0000-0003-1283-2798
https://www.ingv.it/it/organizzazione/chi siamo/personale/#987

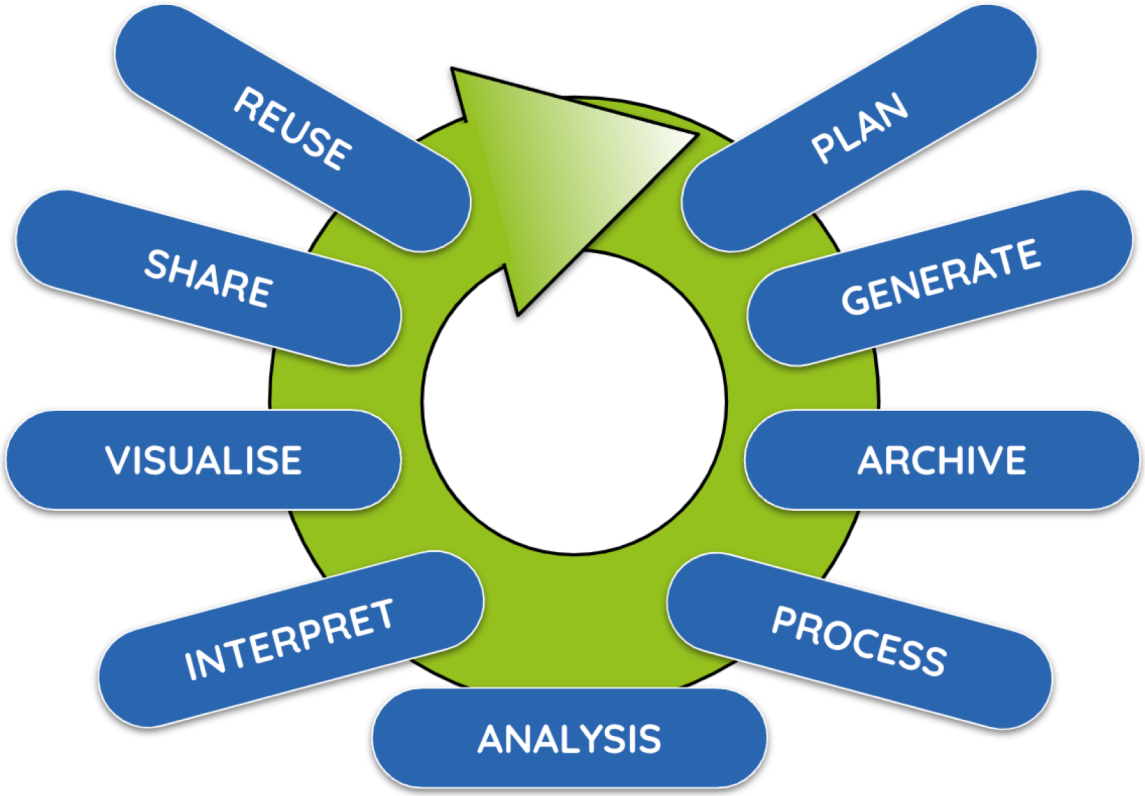
IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani

Outline

- Data life cycle and quality control
- Data products: SeaDataCloud example
- Virtual Research Environments and Operational workflows
- FAIR principles and FAIRness assessment

Research Data Life Cycle



Research Data Management:

Organization, storage, preservation, and sharing of data in a research environment throughout the **entire data life cycle and beyond**.

RDM ensures the efficiency, transparency, and reproducibility of research and include **practices and policies that aim to improve the quality and value of research data**.

Data collection: Quality Assurance and Best Practices

- Traditional view of data (collecting, processing, analyzing, and publishing of results) substituted by a **life cycle approach** that highlights the importance of finding, storing, and sharing data
- **Prior planning** has become mandatory to document data creation, content, context, but also to fulfill data quality requirements
- Data quality requires predefined Quality Assurance (QA) strategies based on the selection of internationally validated methodologies for sampling and analysis

Data Providers (PI)

- are responsible of the adequacy of the sampling strategy to the scope
- must follow specific QA procedures and protocols applied before and during the dataset creation
- have responsibilities in terms of documentation, calibration/intercalibration exercise, sampling strategy, admissible ranges of data, algorithms used, corrections, and flags

QA/QC protocols and Standard Operating Procedures

Best practices have been developed from expert groups to have agreed and broadly adopted methods across ocean research, operations and applications

Best Practice:

- is a methodology that has repeatedly produced superior results relative to other methodologies with the same objective
- is method that has been adopted and employed by multiple organizations

https://www.oceanbestpractices.org/

Metadata

- Records of monitored parameters must include a minimum set of information mapped through metadata
- overview of the sensors and the methodologies (platform, instrument type, sensor's accuracy, calibration info)
- measurement position, date and time
- units
- quality information (quality flags)

Standard and Formats

- controlled vocabularies
- ISO 19115 metadata standards
- Data Transport Formats
- common QC protocols and flag scales

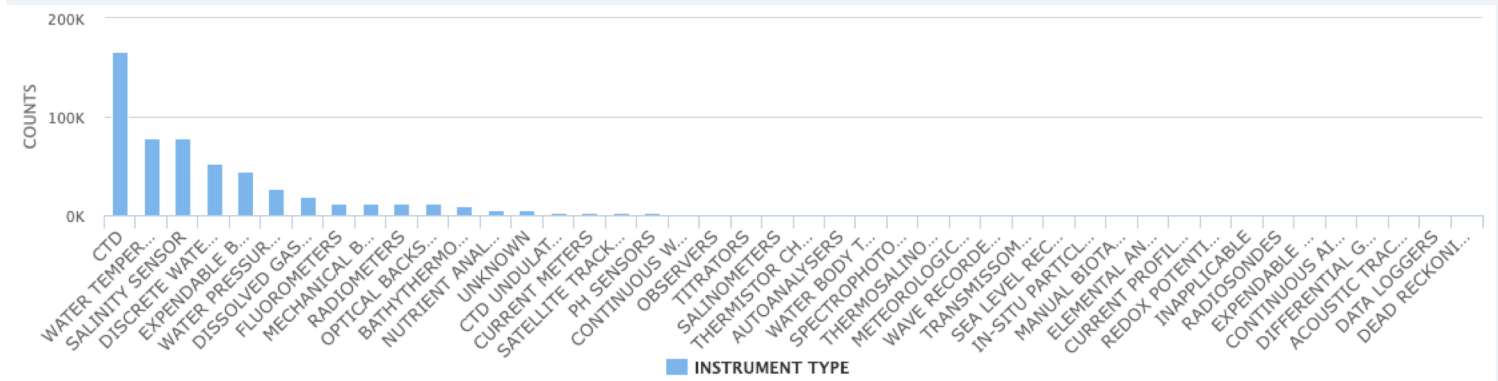
The Power of Metadata

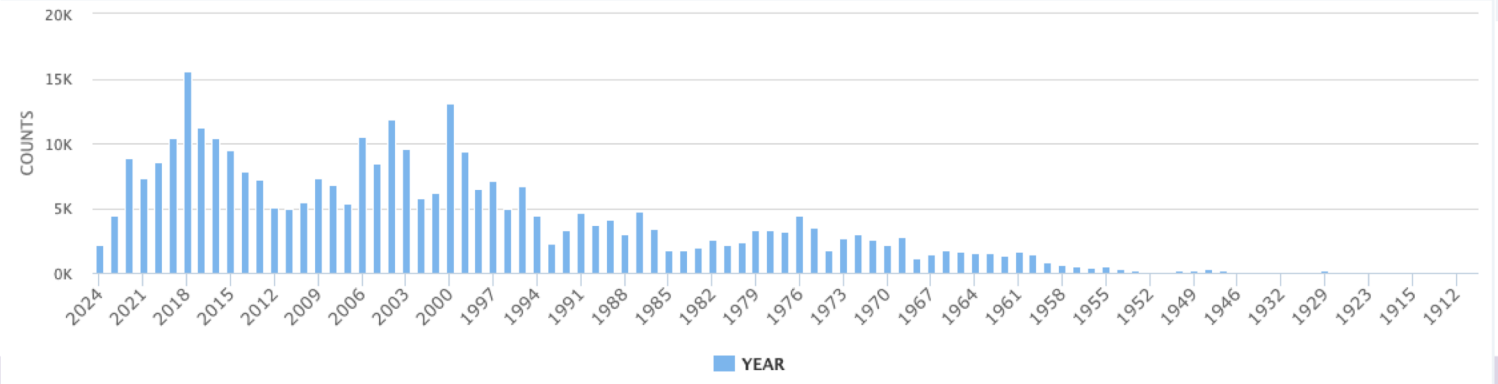
Metadata analysis can be used to monitor the **data sharing monitoring landscape:** data originator/distributor, instrument type, platform type, temporal distribution, data access restrictions

https://cdi.seadatanet.org/search

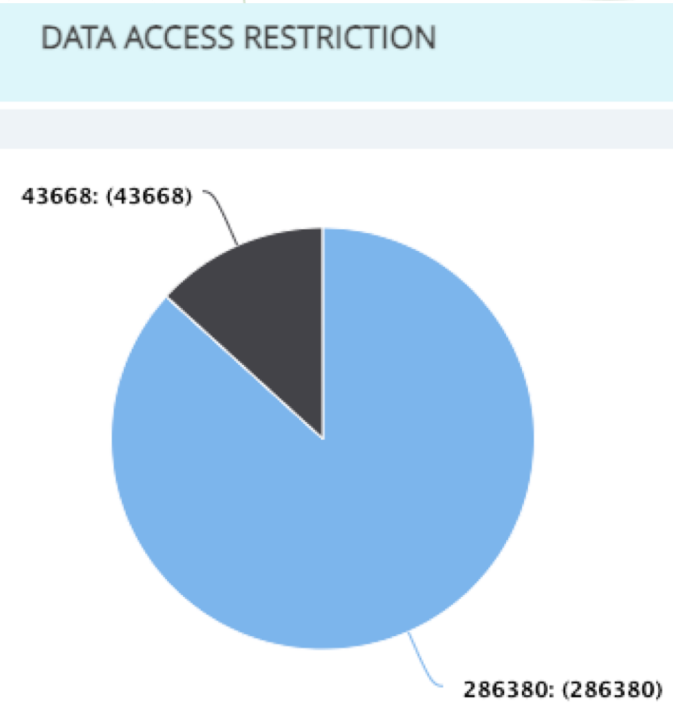
The Power of Metadata

https://cdi.seadatanet.org/search

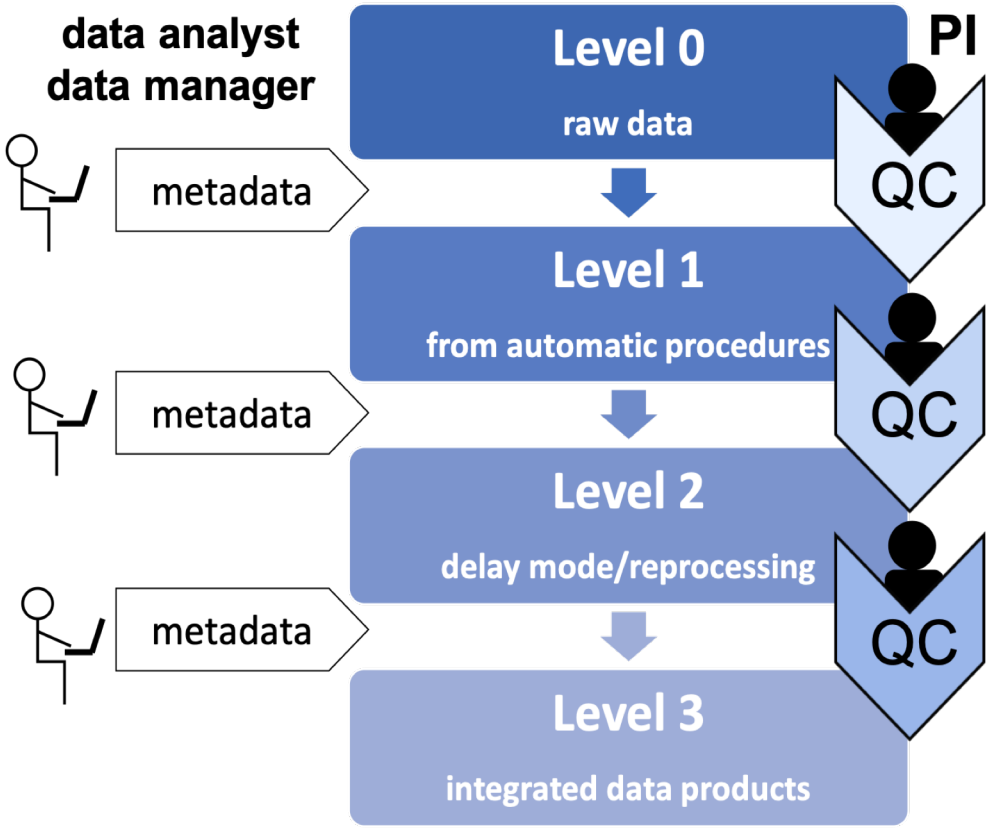




Metadata analysis from SeaDataNet data access portal on the specific data query results (i.e selected space-time domain, parameter)



Data Processing Level (DPL)



Data Latency

Real Time

Near Real Time

Delay Mode

Purpose

- ocean prediction
- early warning

- ocean state assessment
- climate studies

This could generate issues related to different data versions and duplicates within data infrastructures

Dataflow

Data Provider

data analyst
data manager

PI

Level 0
raw data

QC

metadata

Level 1
from automatic procedures

QC

metadata

Level 2
delay mode/reprocessing

QC

metadata

Level 3
integrated data products

DOI
dataset paper +
scientific paper

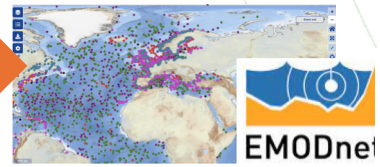
users

sharing

sharing

sharing

Blue Data Infrastructures



Euro Argo

emso ERIC

Findable Accessible Interoperable Reusable

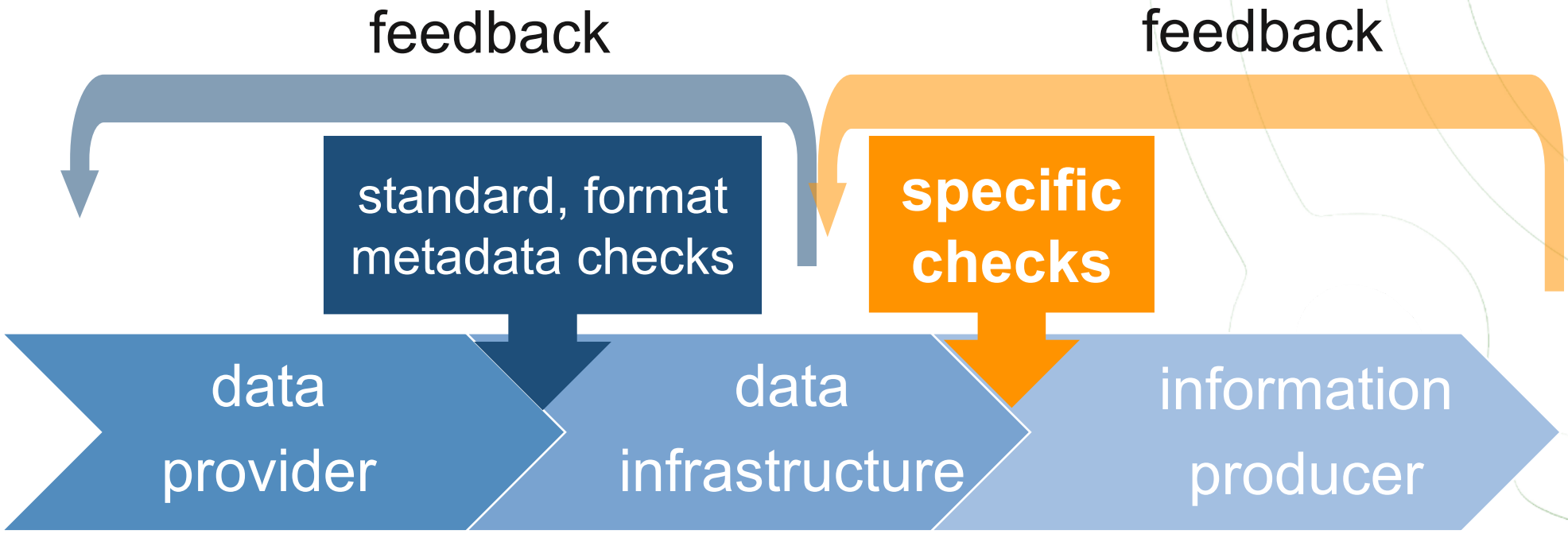
data integration &
metadata harmonization

PRODUCTS

end users

Findable Accessible Interoperable Reusable

Data Quality Control



- **FAIRness**
- **integrity**
- **consistency**
- **completenes**

→ several actors apply QA/QC procedures along with the data value chain

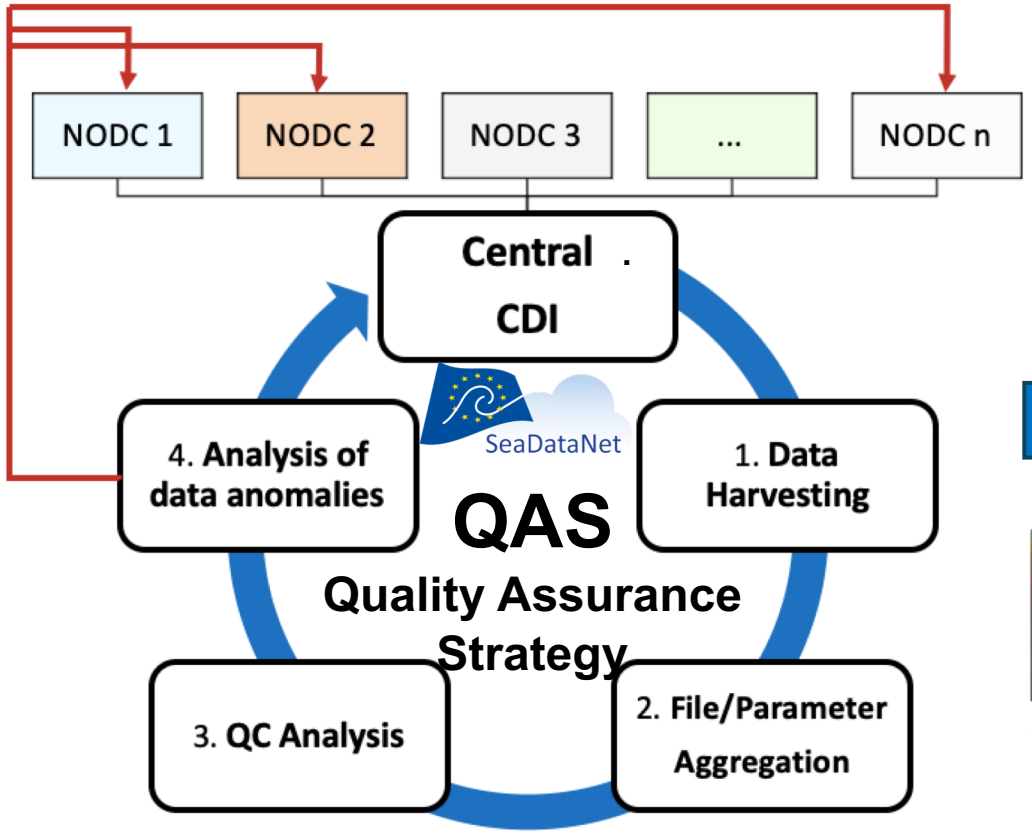
→ **provenance** and **lineage information** are key elements to preserve

Simoncelli et al. (2022) https://doi.org/10.1016/B978-0-12-823427-3.00001-3

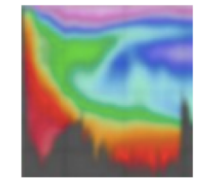
Data Products (SeaDataCloud example)

SeaDataCloud

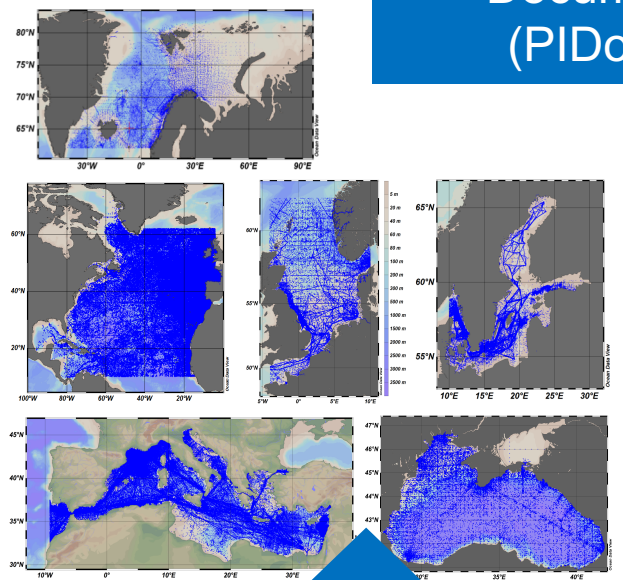
network of SeaDataNet data centers



T&S aggregated dataset



Ocean Data View



Copernicus Marine Service

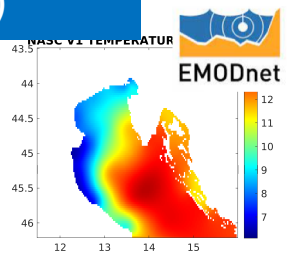
NOAA

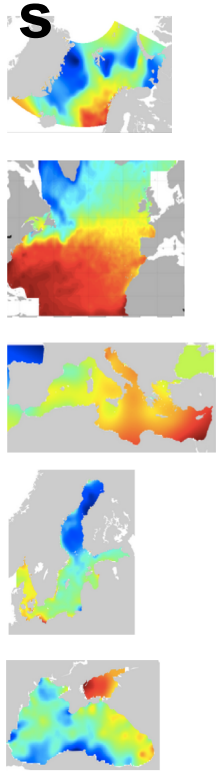
merging data from other BDIs NEW

SDN catalog (DOI) + Product Information Document (PIDoc)

DIVA

T&S gridded climatologies

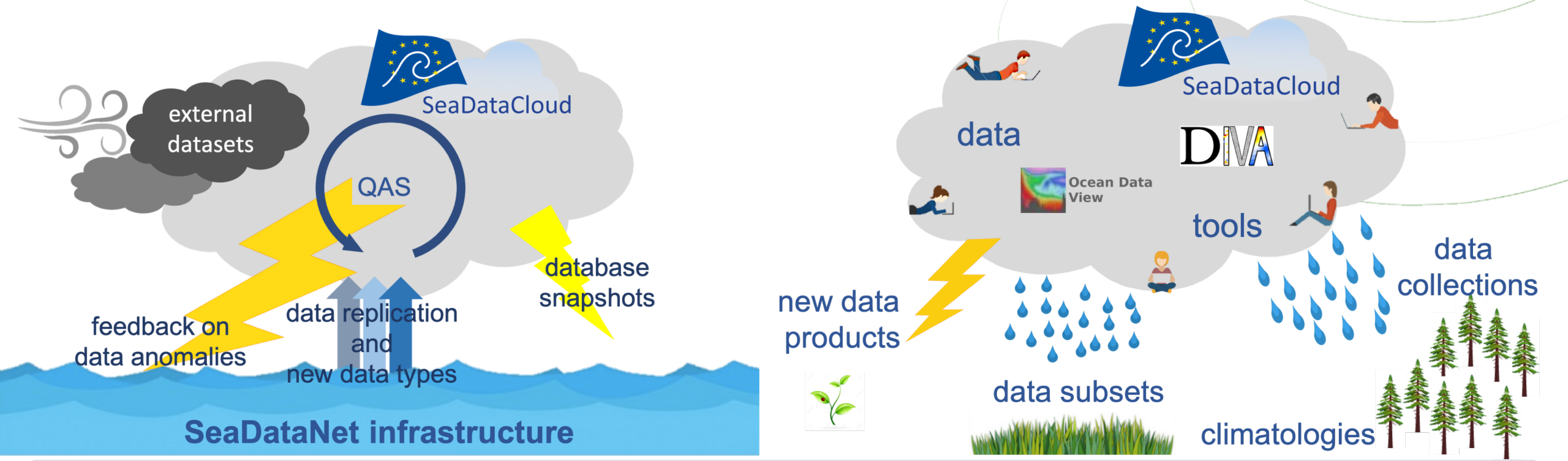




+ OHC, MLD, density, stratification

SeaDataCloud pilot

A **cloud environment** and a **VRE** were implemented to have data replication for faster data access and a co-working environment with shared tools to derive products

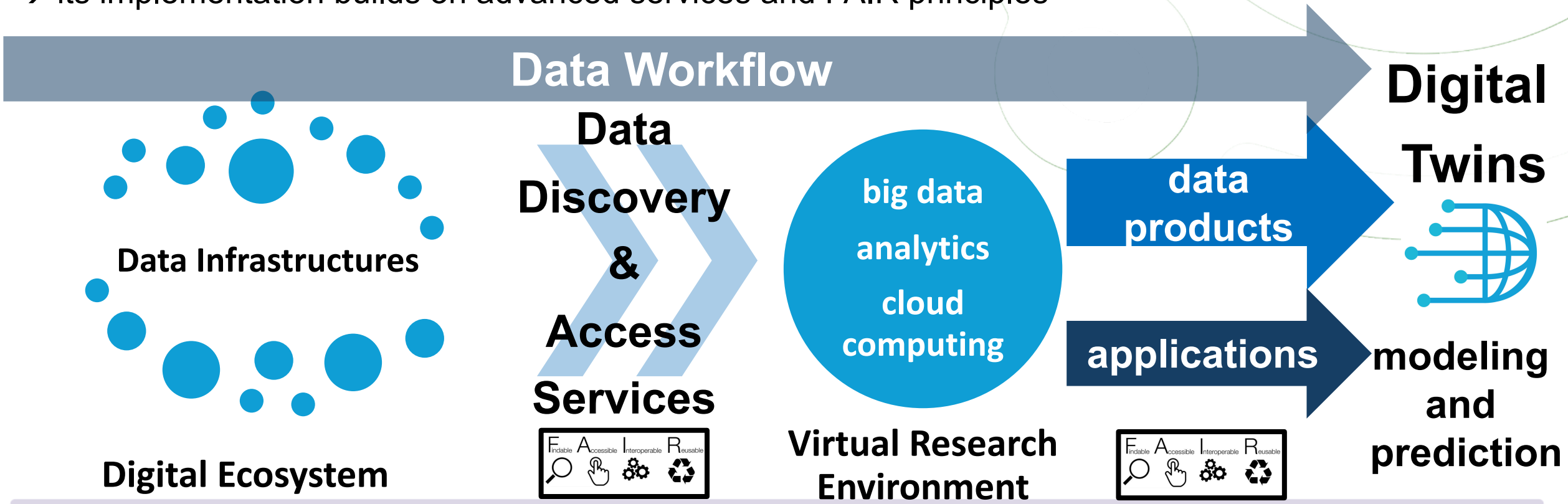


Operational Workflows

Workflow → a series of activities/processes that are necessary to complete a task

→ its **automation and managment process** increase efficiency, optimize the results, allows its replicability and monitoring

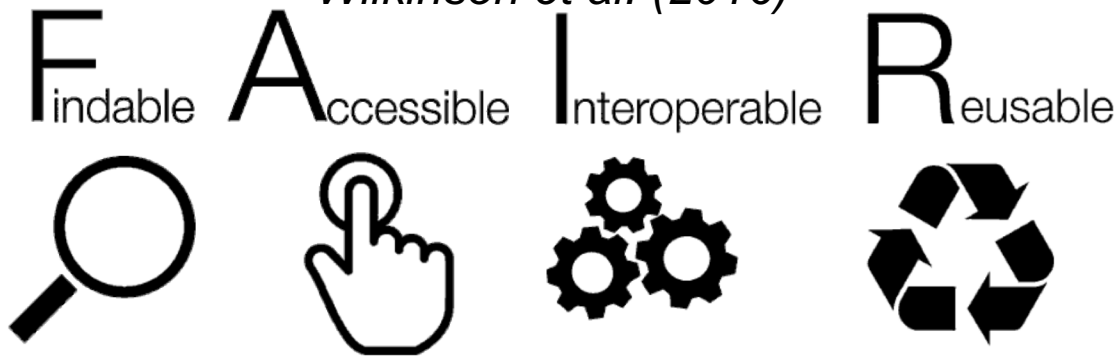
→ its implementation builds on advanced services and FAIR principles



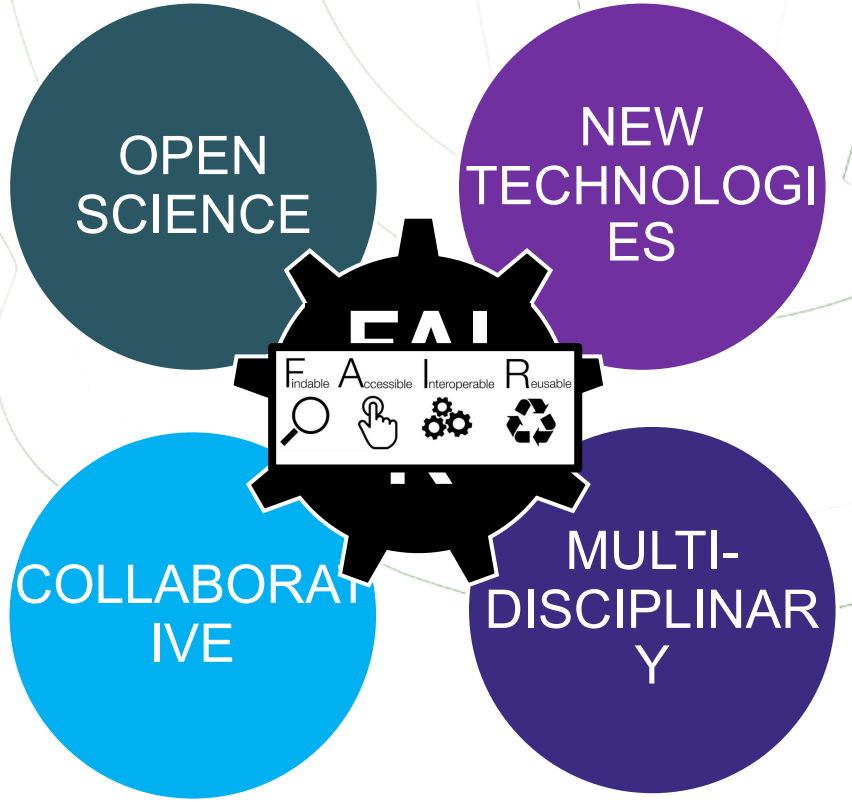
FAIR principles

Societal benefit of science can be enhanced through a community effort to collect, manage and share the data acquired with a specific purpose for further re-use → data driven science

Wilkinson et al. (2016)



Findable: easy to find by humans and computers thanks to metadata and unique persistent identifiers
Accessible: stored for easy access and downloading
Interoperable: ready to be combined with other datasets by humans and computer systems
Re-usable: ready for reuse thanks to detailed, accurate documentation and clear usage license



FAIR principles applied to data and digital artefacts (software, services) are key enablers of the expected ocean science revolution, speeding up information and knowledge generation process

Blue-Cloud open science platform

Data Federation
interoperability, discovery, access and sharing

Collaborative Research
data preparation, dana analysis and publication

Virtual Laboratories

VLab1 Coastal oceans observations along Europe

VLab2 Coastal currents from observations

VLab3 Carbon-Plankton dynamics

VLab4 Marine Environmental Indicators

VLab5 Global Fisheries Atlas

Products

Technology

Data series, data products

Copernicus Marine Service

EcoTaxa 2.4

ENA European Nucleotide Archive

EMODnet European Marine Observation and Data Network

Euro-Argo

SOCAT Surface Ocean CO2 Atlas

EurOBIS European Ocean Biogeographic Information System

ICOS Integrated Carbon Observation System

SeaDataNet

SIOS Svalbard Integrated Arctic Earth Observing System

emso ERIC

EMBRC European Marine Biological Resource Centre

WEkEO by Copernicus

Products

Technology

eosc | Blue-Cloud2026

Research method Scientific applications

Collaboration · Sharing · Reuse · Reproducibility

VRE
Blue-Cloud Core Services
EOSC Core Services

EOVs data collections

EDITO European Digital Twin Ocean

Data Workbenches
data extraction, data cleaning data processing and deployment

Products

Technology

physical workbench for temperature, salinity

chemical workbench, linked to eutrophication: nutrients, chlorophyll, oxygen

ecosystem workbench for plankton biomass and diversity

Funded by the European Union

Connecting European community, assets, science and technnology

To ensure an operational Digital Twin Ocean core infrastructure by 2024, the building blocks of the initiative are underway through the European-funded EDITO projects, namely EDITO-Infra and EDITO-Model Lab.

Together, these two projects will build the Digital Twin Ocean platform, incorporating a data lake, processing engine and virtual simulation environment.

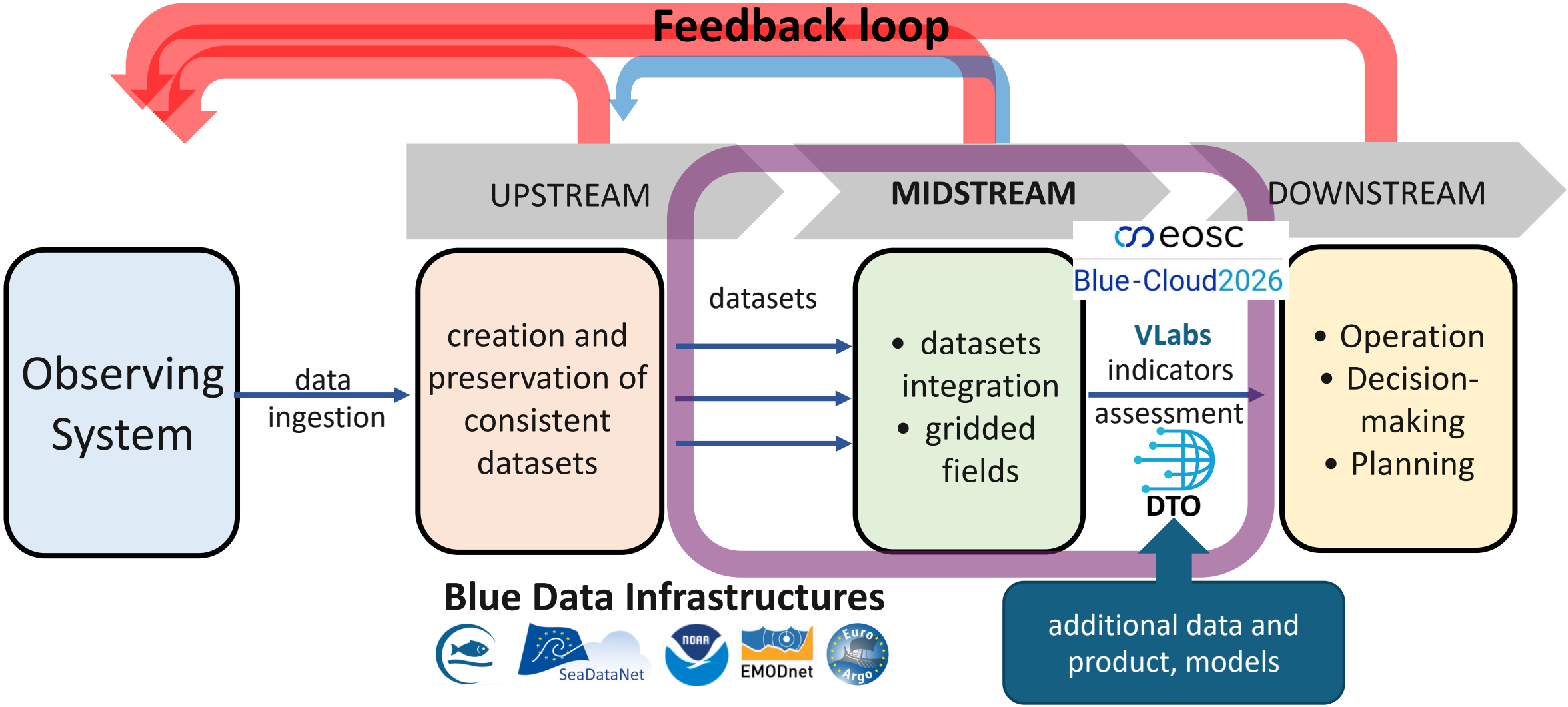
Fostering innovation through co-creation and collaboration, EDITO connects the European marine community, leveraging all observing, monitoring and data programmes, expertise and cutting-edge technology across the European ocean science landscape.

EDITOInfra

A 2-year project that will build the public infrastructure backbone for EDITO by integrating key data service components (among which Copernicus Marine Service and EMODnet), and by sharing cloud processing capabilities and software into a single digital framework

EDITOModelLab

A 3-year project aiming to develop the next generation of ocean models, combining artificial intelligence and high-performance computing, to be integrated into the EDITO public infrastructure, providing access to focus applications and simulations of different what-if scenarios



Blue-Cloud 2026 aims at evolving into a **federated EU ecosystem to deliver FAIR & Open data** and **analytical services** instrumental for deepening research of the ocean

about Quality

Quality Assurance	QA is about adoptinge good measuring practices (sampling strategies, data collection methodologies, sample pretreatment, transport and storage)
Quality Control	QC process of anomalous data detection that apply after data gathering
Quality Assessment	It consists of evaluating if data meet the necessary quality requirements or are adequate for an intended use. It influences the selection of a data product among others

Monitoring activity

data product L0-L4

downstream product

Data accuracy and uncertainty are key quality elements of data reliability→ data **uncertainty determination and its propagation along with the data value chain** is still a present challenge. It is very important to annotate data uncertainty in the metadata

Quality Control

- QC is vital for data reuse, and without it data from different sources cannot be combined to gain value
- Scientific, analytical and statistical evaluations must determine if data present adequate quality to support the intended data usage, resulting in labeling each numerical value with a Quality Flag (QF) and avoid modifying the original data record following a harmonized scheme of QFs
- QFs ensure that the quality of the data is apparent to the user, who holds sufficient information to decide the suitability for a specific task applying the proper data filtering

QC practices include:

- data integrity checks (e.g., format)
- data value checks (range checks, spikes and outliers checks, neighbor checks, climatology checks)
- QC highly depend on the data thematic, sensor type and the available amount of time for the analysis (RT vs DM)

Automatic QC (RT procedures) → algorithms

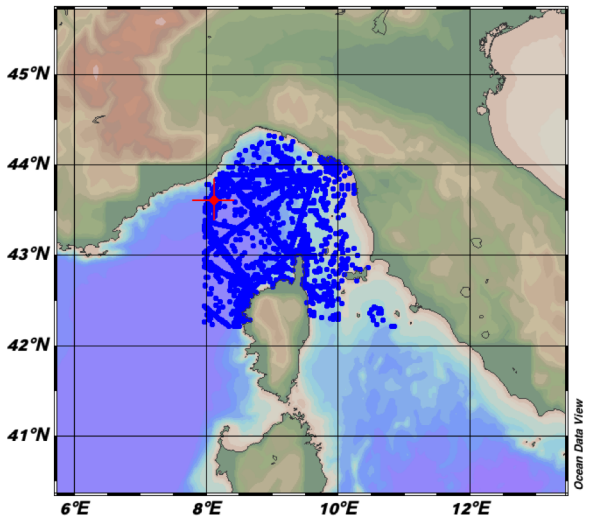
Visual QC (DM procedures) → visual tools (i.e. OceanDataView) and graphical interfaces

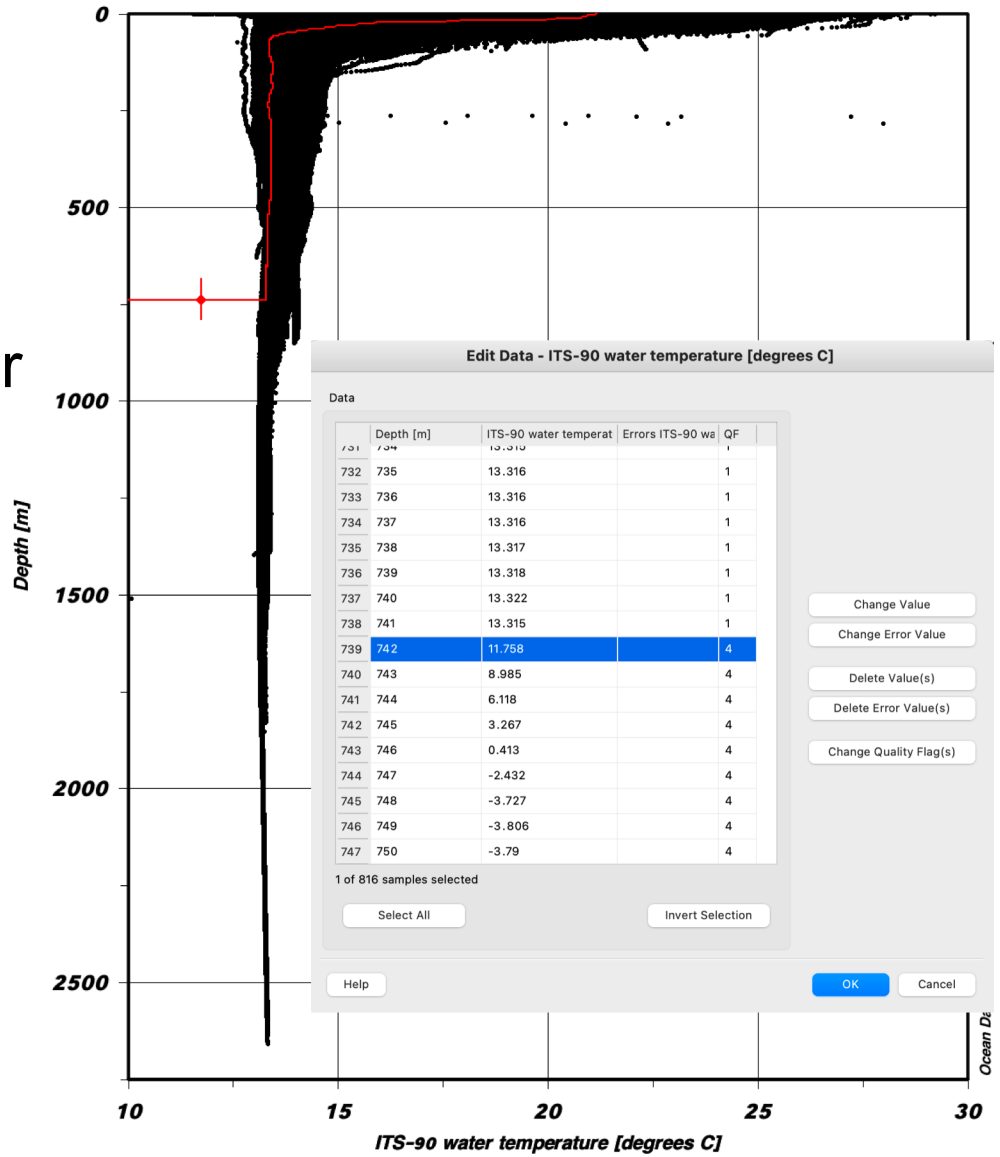
Quality Flags

- labels associated to each measurement that the user can use to filter data according to the needs
- SeaDataNet harmonised scheme of QC Flags to be used to label individual data values
- QC Flag scale is available in the SeaDataNet Common Vocabularies as list L20

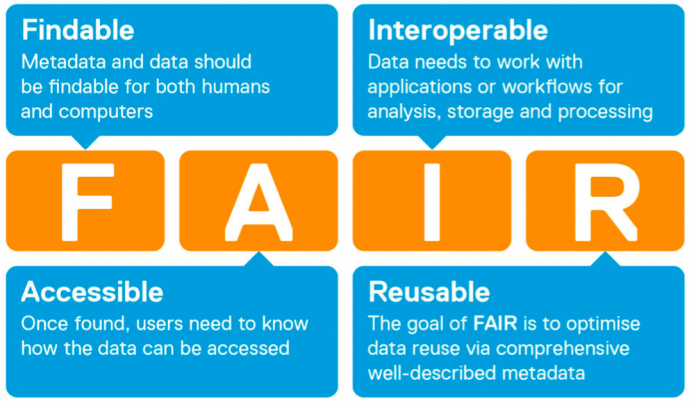
Key	Entry Term	Abbreviated term	Term definition
0	no quality control	none	No quality control procedures have been applied to the data value. This is the initial status for all data values entering the working archive.
1	good value	good	Good quality data value that is not part of any identified malfunction and has been verified as consistent with real phenomena during the quality control process.
2	probably good value	probably_good	Data value that is probably consistent with real phenomena but this is unconfirmed or data value forming part of a malfunction that is considered too small to affect the overall quality of the data object of which it is a part.
3	probably bad value	probably_bad	Data value recognised as unusual during quality control that forms part of a feature that is probably inconsistent with real phenomena.
4	bad value	bad	An obviously erroneous data value.
5	changed value	changed	Data value adjusted during quality control. Best practice strongly recommends that the value before the change be preserved in the data or its accompanying metadata.
6	value below detection	BD	The level of the measured phenomenon was too small to be quantified by the technique employed to measure it. The accompanying value is the detection limit for the technique or zero if that value is unknown.
7	value in excess	excess	The level of the measured phenomenon was too large to be quantified by the technique employed to measure it. The accompanying value is the measurement limit for the technique.
8	interpolated value	interpolated	This value has been derived by interpolation from other values in the data object.
9	missing value	missing	The data value is missing. Any accompanying value will be a magic number representing absent data.
A	value phenomenon uncertain	ID_uncertain	There is uncertainty in the description of the measured phenomenon associated with the value such as chemical species or biological entity.

Ocean Data View (ODV) software package for the interactive exploration, analysis and visualization of oceanographic and other geo-referenced profile, time-series, trajectory or sequence data





https://odv.awi.de/



FAIR proposed by the community

https://force11.org/info/the-fair-data-principles/

Findable	
F1	(meta)data are assigned a globally unique and eternally persistent identifier.
F2	data are described with rich metadata.
F3	(meta)data are registered or indexed in a searchable resource.
F4	metadata specify the data identifier.
Accessible	
A1	(meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1	the protocol is open, free, and universally implementable.
A1.2	the protocol allows for an authentication and authorization procedure, where necessary.
A2	metadata are accessible, even when the data are no longer available.

Interoperable	
I1	(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2	(meta)data use vocabularies that follow FAIR principles.
I3	(meta)data include qualified references to other (meta)data.
Re-usable	
R1	(meta)data have a plurality of accurate and relevant attributes.
R1.1	(meta)data are released with a clear and accessible data usage license.
R1.2	(meta)data are associated with their provenance.
R1.3	(meta)data meet domain-relevant community standards.

FAIRness Assessment

Useful for data providers/producers to learn and adopt FAIR principles during data generation or to implement data FAIRification on existing datasets

There exist subjective and objective assessment methods:

1. Data Maturity Model (self-evaluation) https://doi.org/10.15497/rda00050
2. F-UJI tool (automatic evaluation) https://www.f-uji.net/

FAIR principles and assessment is expanding to all «digital research objects» (software, semantic artifacts, VRE, ...) to facilitate Virtual Research Environment workflows.

Software FAIRness: indicators under development https://doi.org/10.15497/RDA00068

References

Simoncelli, S., Manzella, G.M.R., Storto, A., Pisano, A., Lipizer, M., Barth, A., Myroshnychenko, V., Boyer, T., Troupin, C., Coatanoan, C., Pititto, A., Schlitzer, R., Schaap, D.M.A., Diggs, S., 2022. A collaborative framework among data producers, managers, and users. https://doi.org/10.1016/B978-0-12-823427-3.00001-3 In: Manzella, G., Novellino, A. (Eds.), **Ocean Science Data: Collection, Management, Networking and Services**. Elsevier, pp. 197–280. ISBN: 9780128234273

UNESCO-IOC (2023). Ocean Decade Data & Information Strategy. Paris, UNESCO. (The Ocean Decade Series, 45)

Calewaert et al. (2024). Ocean Decade Vision 2030 White Papers – Challenge 8: Create a Digital Representation of the Ocean. Paris, UNESCO-IOC. (The Ocean Decade Series, 51.8.). https://doi.org/10.25607/bxhy-ra59

Miloslavich et al. (2024). Ocean Decade Vision 2030 White Papers – Challenge 7: Sustainably Expand the Global Ocean Observing System. Paris, UNESCO-IOC. (The Ocean Decade Series, 51.7.). https://doi.org/10.25607/brxb-kr45

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3:60018 https://doi.org/10.1038/sdata.2016.18

https://emodnet.ec.europa.eu/sites/emodnet.ec.europa.eu/files/public/PDF/EUInSituMarine_EMODnet_CMEMS_FINAL.pdf

European Commission: Directorate-General for Research and Innovation, The digital twin ocean – An interactive replica of the ocean for better decision-making, Publications Office of the European Union, 2022, https://data.europa.eu/doi/10.2777/343496

Locati & Cacciola (2024) Research Data Management https://istituto.ingv.it/ufficio-gestione-dati

THANKS!

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani