



**Marine Data Store planning with data flows and connection standards for the distributed and federated system interfaced with ITINERIS central hub**



Deliverable number:	D5.2
Work package:	WP5 – Marine Domain
Intermediate Objective:	IO5.2 – (B3): (a) Marine Data Center Management plan and governance for the distributed and federated system; (b) Implementation plans for technological assets acquisitions at all OU; (c) first personnel hired; (d) WP progress meeting.
Deliverable type:	<input checked="" type="checkbox"/> Document, report
	<input type="checkbox"/> Websites, patent filings, videos, etc.
	<input type="checkbox"/> Other: please specify .....
Dissemination level:	<input checked="" type="checkbox"/> Public
	<input type="checkbox"/> Restricted
Estimated delivery (bimester):	B3
Actual delivery date:	30/04/2023
Author(s) (Partner-OU):	Mauro Caccavale (CNR-ISMAR),
Reviewed by:	ITINERIS Executive Board
Note:	

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System - CUP B53C22002150006 (D.D. n. 130/2022)  
 Funded by EU - Next Generation EU  
 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”

## Table of contents

1.	<i>INTRODUCTION</i> .....	4
1.1	<b>Purpose of the document</b> .....	4
1.2	<b>DEFINITIONS, ACRONYMS AND ABBREVIATIONS</b> .....	4
2.	ITINERIS Marine Domain .....	5
3.	<i>INTRODUCTION</i> .....	5
4.	<i>MARINE DOMAIN</i> .....	6
5.	<i>RESEARCH INFRASTRUCTURES INVOLVED</i> .....	7
6.	Marine Data Store Central Hub .....	9
7.	<i>DATA CENTRE DESCRIPTION</i> .....	9
8.	<i>DATA TYPE AND EXPECTED SIZE</i> .....	10
9.	<i>DATA ALLOCATION AND SECURITY</i> .....	11
10.	<i>DATA FLOW</i> .....	12
11.	Marine data store “External Interfaces” .....	18
12.	<i>INTERFACE WITH THE ITINERIS CENTRAL HUB WP2</i> .....	18
13.	<i>INTERFACE WITH THE ITINERIS VIRTUAL RESEARCH ENVIRONMENTAL WP8</i> .....	18
14.	<i>INTERFACE WITH THE ITINERIS TERRESTRIAL AND BIOSPHERE HUB WP6</i> .....	18
15.	<i>INTERFACE WITH THE COI/IODE/OBIS</i> .....	19
16.	FAIR data .....	19
17.	Data Policy .....	20

## 1. INTRODUCTION

### 1.1 Purpose of the document

This document identifies the aspects to be managed in the creation of the ITINERIS - Marine Domain Data Centre and its data access portal.

The document was developed in parallel with the D5.1 Marine Data Centre Management Plan of which it can be considered an extension and supplement.

The ITINERIS - Marine Data Store planning document is the key element of the marine domain data/metadata lifecycle management and describes the mechanisms for collecting, processing and generating data and data products.

The objective of the ITINERIS - Marine Data Store is to describe the current situation of the marine domain, the management of data/metadata between the various RIs in the short and long term by limiting the possibility of damage to the data/metadata and the optimization of the flows to be implemented and maintained. Some RIs (and/or partners) already have procedures and interfaces in place devoted to the external exposure of data, while others still need to implement certain functionalities and thus needs more time to align within the aims of the marine domain infrastructure.

It should be underlined that this document should be regarded as a “living document” that will evolve through subsequent updates.

### 1.2 Definitions, Acronyms and Abbreviations

CNR	Consiglio Nazionale delle Ricerche
COI	Italian Oceanographic Commission
DISSCO	Distributed System of Scientific Collections
EMODnet	European Marine Observation and Data Network
EOOS	Reuropean Oceanographic Observing System
ERIC	European Research Infrastructure Consortium
ESFRI	European Strategic Forum on Research Infrastructures
FAIR	Findable, Accessible, Interoperable, Reusable
GOOS	Global Oceanographic Observing System
INFN	Istituto Nazionale di Fisica Nucleare
INGV	Istituto Nazionale di Geofisica e Vulcanologia
IBBR	Istituto di Bioscienze e Bio-risorse
IODE	Oceanographic Information and Data Exchange
IOOS	Italian Integrated Oceanographic Observing System

ISMAR	Istituto di Scienze MARine
ISPRA	Istituto Superiore per la Protezione
ITINERIS	Italian Integrated Environmental Research Infrastructures System
LNS	Laboratori Nazionali del Sud
NetCDF	Network Common Data Form
NRT	Near Real Time
OBIS	Oceanographic Biodiversity Information System
OGS	Istituto Nazionale di Oceanografia e di Geofisica Sperimentale
RI	Research Infrastructures
SPARQL	SPARQL Protocol and RDF Query Language
WP	Work Package

## 2. *ITINERIS Marine Domain*

### 3. INTRODUCTION

The Italian Integrated Environmental Research Infrastructures System (ITINERIS) is the PNRR project dedicated to create and consolidate a networking between the currently involved 22 Research Infrastructures (RIs) (ref. section 2.2 for more details).

The continuing environmental challenges facing our society and the scientific community stress for a general adjustment of study strategies and technologies for the next decade.

The complexity of the Earth system, the multi-disciplinarity of research fields, the growing demands of civil society in terms of green solutions, and the availability of ever-increasing amounts of data, represent not only a challenge but also an opportunity for defining future scientific research infrastructures and models.

The ITINERIS project will coordinate a network of 22 research infrastructures distributed throughout Italy and investigating the 4 domains into which the study of the Earth's environment is usually divided (atmosphere, hydrosphere, biosphere and geosphere).

The most important challenge will be to develop and structure all the inter-connections between the different environmental subdomains with a view to sharing, accessibility, and interoperability of data and resources.

The harmonization of the activities of the different research infrastructures participating in ITINERIS will enable the creation of the integrated national system that will be the future European reference in environmental sciences.

A first step toward establishing a national and European reference for the next 10 years will be to standardize the maturity level of the different RIs involved by defining and supporting the sharing of

best practices for the complementary and synergistic development of skills, resources and services of individual research infrastructures and between the different hubs for environmental study.

A first purpose of harmonization and uniformity among different hubs and among different RIs will be the creation of a system of Virtual Research Environments (VREs). VREs will provide a first tangible product of the harmonization activity of ITINERIS structures/infrastructure as well as data/metadata to end-users and stakeholders. They will also enable the assumption of leadership in addressing complex and multidisciplinary environmental challenges.

The project will create the ITINERIS HUB, where all users will have access to data and services, with an appropriate access management system and a comprehensive catalogue of data and services. ITINERIS will not develop a new Data Center, but will harmonize existing ones, which will be strengthened by the project activities. Most of the data-related activities will be devoted to achieving the most appropriate level of FAIRness for all RIs involved.

Within the ITINERIS project, the 22 RIs are organized into 4 hubs (corresponding to the 4 classic domains of subdivision of the environmental research scope) Fig.1.

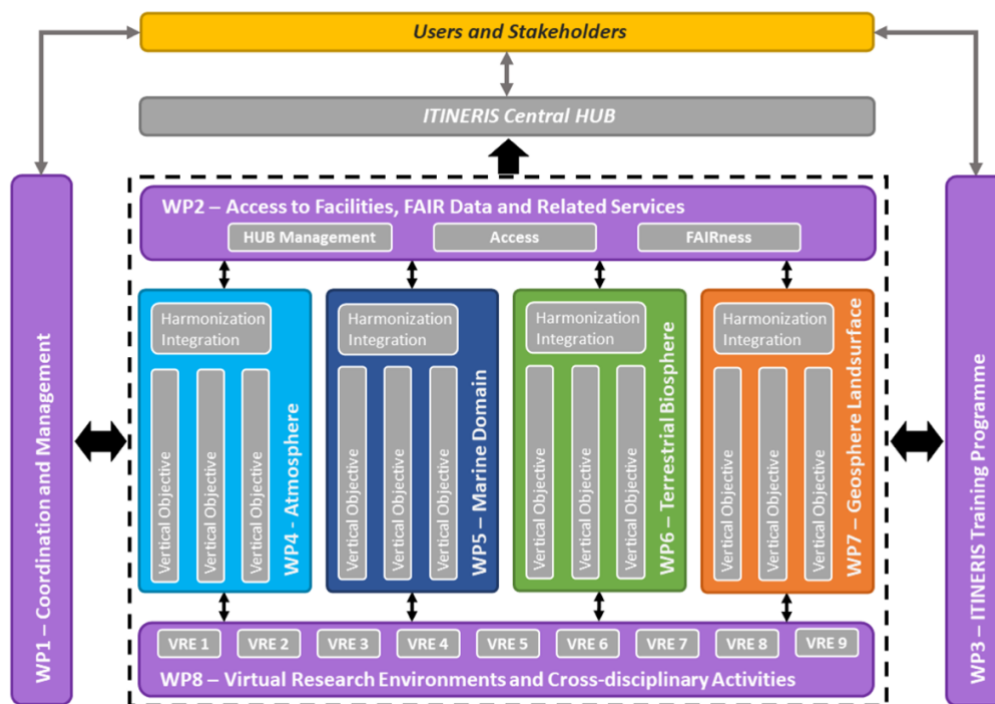


Fig.1 ITINERIS project Work Package organization.

#### 4. MARINE DOMAIN

The Marine Domain hub (WP5) consists of 11 RIs: DANUBIUS, eLTER, EMSO ERIC, EURO-ARGO ERIC, EUOFLEETS, ICOS ERIC, JERICO, Laboratori Nazionali del Sud, N/R Laura Bassi, SIOS, GeoSciences.

The Marine Domain hub (WP5) will constitute the Italian Integrated Oceanographic Observing System (IOOS), which will contribute at the European and International level (EOOS and GOOS) with the aim to meet the challenges of the "decade of the sea" and for the Mediterranean Sea.

WP5 will enhance the digital automatization of Italian Research Vessels and IOOS will improve quality and interoperability of ocean sensors, infrastructures and data at the national level for three critical themes: climate, operational services, and marine ecosystem health. WP5 will also show how data integration from RIs can improve monitoring of coastal waters (including river discharge) and the 4D view of the ocean state, thus building the digital twin of the ocean.

WP5 Marine Domain aims to become the Italian reference point for easy, fast and open access to research and data acquired through the 11 RIs dealing with the marine sphere and, at the same time, will include existing data/metadata and research products.

The implementation of a single marine data collector, integrated and cooperating with the global collector of the ITINERIS project (WP2), will be the basis of an infrastructure on which to implement the adoption of FAIR principles.

This objective will be realized by close collaboration between WP5 and WP2, as well as with other stakeholders (see sections 4). At hub level specific actions are promoting and coordinating to ensure that data and other research products, including the services and archives needed to manage and list the various resources, are made FAIR. This will be done by establishing all necessary technical mechanisms to link resources and collaborative mechanisms for alignment and agreement on specifications and standards.

## 5. RESEARCH INFRASTRUCTURES INVOLVED

The marine domain of the ITINERIS project (WP5), which will form the basis of the Italian Integrated Ocean Observing System (IOOS), aims to coordinate national activities and observations from research infrastructures distributed throughout the country (Fig.2) in order to improve the quality and interoperability of ocean data.

There are 11 RIs involved in WP5, 5 developed under the European Strategic Forum on Research Infrastructures (ESFRI), with intra- and extra-EU connections, and 3 large national infrastructures. The management of these infrastructures involves 5 different entities (Fig.2): CNR, INGV, INFN, ISPRA and OGS.

### WP5: Participants in the WP

CNR-ISMAR-Venezia  
CNR-ISMAR-Bologna  
CNR-ISMAR-Lerici  
CNR-ISMAR-Roma  
CNR-ISMAR-Napoli  
CNR-IAS-Genova  
CNR-IRSA-Taranto  
CNR-ISP-Messina  
OGS\_OCE-Trieste  
OGS\_CGN-Trieste  
INGV-BO -Bologna  
INGV-WIS - Catania  
INFN-LNS -Catania  
ISPRA - Roma



Fig.2. List of ITINERIS Marine Domain partners.

The 5 RIs involved in ITINERIS WP5 are:

**Euro-Argo ERIC** allows active coordination and strengthening of the European contribution to the international Argo programme to explore the open ocean.

**EMSO ERIC** aims to observe the oceans in the long term, better understand the phenomena in and under them and their interactions and roles in terrestrial systems through a multidisciplinary approach.

**DANUBIUS-RI** mission is to facilitate and contribute excellent science on the continuum from river source to sea; to offer state-of-the art research infrastructure; and to provide the integrated knowledge required to sustainably manage and protect River-Sea Systems.

The **ICOS ERIC** Marine component currently coordinates twenty-nine ocean stations from seven countries monitoring carbon uptake and fluxes. ICOS ERIC ensures high quality measurements that are independent, transparent and reliable, to quantify fluxes and to assess their variability and drivers.

**JERICO-RI** is an integrated pan-European multidisciplinary and multi-platform research infrastructure dedicated to a holistic appraisal of coastal marine system changes.

**eLTER-RI** is a pan-European in-situ research infrastructure whose mission is to study long-term ecological changes in terrestrial, freshwater and transitional ecosystems through a holistic 'whole system' approach.

The **Eurofleets+ project** facilitates open access to an integrated and advanced research vessel fleet, designed to meet the evolving and challenging needs of the user community.

**SIOS** is a collaborative effort to develop and maintain a regional observational system for long-term measurements in and around Svalbard Islands (Arctic Ocean), addressing Earth System Science questions related to Global Change.

The 3 large national infrastructures are represented by:

“**Laboratori Nazionali del Sud**” **LNS** is a national RI based in Sicily, hosting unique facilities for astroparticle, nuclear and applied physics and marine science. LNS owns and operates the largest scientific subsea cabled network in Europe.

“**Laura Bassi**” is an icebreaker class research vessel. It operates in polar areas carrying out research and logistic activities for the Italian polar projects.

**GeoScience** is a network of Italian Regions and Regional Environmental Agencies providing access to data, service and scientific knowledge to carry out monitoring and environmental impact and strategic assessment procedures.

For each RIs, a number of contact peoples have been identified among the different partners who can contribute to the realization of the objectives of the different tasks. During the project, as a result of a still ongoing internal discussion, managers/referents will be defined to refer to for the data/metadata produced by the respective RIs. Similarly, the referents of each partner will be identified so that they can coordinate the development and implementation activities of the marine domain (or IOOS) both at the level of RIs and of the Research Institutions/partners involved (Fig.3).



Fig.3 List of ITINERIS Marine Domain RIs and referent peoples and institutions.

## 6. Marine Data Store Central Hub

## 7. DATA CENTRE DESCRIPTION

The data centre at the ISMAR (located in Naples) represents the main structure of the marine domain of the ITINERIS project and provides access to the data/metadata collected by the RIs involved. The infrastructure, endowed with considerable computing capacity and storage space, guarantees the possibility of providing primary data storage (or a backup copy of the data of the RIs or partners who request it). Specific services can also be provided at the request of the RIs according to future needs.

The infrastructure guarantees access to the marine domain data based on the FAIR principles and in coordination with the ITINERIS central HUB access portal to all data of the ITINERIS project (in the process of being defined for the technical aspects - WP2).

For the provision of these services and the management of a network of geographically distributed RIs (and partners), the Data Centre is structured so that all its components (in terms of redundant hardware and software) can guarantee the efficiency, stability and operational continuity of the services and security of the stored data/metadata.

The computing power, the data/metadata storage capacity and the overall performance of the system guarantee the possibility of adopting different approaches of "data flows" within the domain (intra-RIs) and with the outside (inter-RIs) each based on specific interconnections and data exchange.

The Data Centre provides all information and communication technology (ICT) services for all users and the public. The Data Centre services should be adapted and will evolve according to the requirements of the RIs and the project. However, some basic services can be identified:

- Storage and access to RIs data/metadata for all users;
- Providing the possibility of a security copy of the data;
- Providing the possibility of a long-term off-line copy for data;
- Provide access and interchange of data/metadata with other WPs of the project and with other infrastructures (IODE, COI, ...)

- Other services that will be highlighted in the continuation of the project

The main components of the Data Centre are: the repository, the ITINERIS data and metadata Catalogue, the metadata registry, the service catalogue and the portal.

## 8. DATA TYPE AND EXPECTED SIZE

A survey was conducted to determine the state of the art in data management of each RI. Content and results of the survey are reported in D5.1 Marine Data Centre Management Plan; management needs, service requirements, governance for the distributed and federated system, currently, not all RIs replicate at national level the general structure defined at international level, in terms of access portals, data management plans, data storage strategies, code/software repositories and/or guidelines.

This creates difficulties in adopting a geographically widespread and federated approach at RIs level for the management, storage and dissemination of the data produced by individual RIs.

This difficulty is also often encountered at the level of the partners, who manage the RIs, thus making it necessary and desirable to define shared strategies at national level to guarantee the correct data workflow and availability over time and secure storage.

The harmonization of data/metadata and of the axis procedures applied by the different RIs of the marine domain will contribute to the rapid establishment of the Italian Integrated Ocean Observation System (IOOS) as a system of distributed and cooperating systems to ensure continuity of data/metadata and services.

A fair degree of uniformity among the different RIs of the marine domain is found in the use of a shared vocabulary (SeaDataNet <https://www.seadatanet.org/Standards/Common-Vocabularies>) that will facilitate data interoperability within the same hub.

With regard to the data formats used, there is no uniformity among the different RIs in the marine domain. The prevailing trend is to provide the same data in different formats. The adoption of different standards, even within the same RI, encourages a shared and non-exclusive approach to be adopted both within the marine domain and in synergy with WP2, which is currently being defined. Particular attention is paid to the possibility/need to guarantee different data output formats only on demand. This solution makes it possible to diversify the output format for the end user without necessarily duplicating the stored data.

Almost all of the RIs in the marine domain adopt a CC-BY 4.0 data license and give their data a unique PID, although there is some disparity as to type.

Searching for data using a SPARQL query language (SPARQL Protocol and RDF Query Language), which is currently not widely used among the Marine Domain RIs, is the subject of internal discussion in WP5 and with the co-ordination of WP2 to evaluate its implementation.

In the current situation, while the infrastructure of the marine domain of the ITINERIS project is being developed and implemented, it is not possible to have a precise and detailed overview of every aspect related to the resources needed to store the data produced by the different RIs. Neither it is possible to have a detailed overview of the costs of running such an infrastructure.

However, the distributed and federated structure that is being set up has been designed in such a way as to guarantee high efficiency in terms of data/metadata security, efficiency in its dissemination and interoperability, and the necessary computing and storage resources for current and future long-term data.

By way of example only, it should be noted that, as a major value, a single observation node of an RIs can produce more than 100 TB of data per year.

As will be shown in the following section, the marine domain infrastructure will be distributed (both in terms of hardware and data/metadata copies) and federated through secure interconnections for information exchange and data/metadata retrieval.

## 9. DATA ALLOCATION AND SECURITY

One of the priorities of the ITINERIS project's marine domain infrastructure is to safeguard data. The value of the scientific data produced by RIs is considered invaluable to the understanding of the ocean system and once lost they are no longer recoverable or reproducible. Extreme care is being taken to ensure that the data is always protected from any errors or malicious actions that could compromise its integrity.

A backup copy of the data stored at the Data Centre will replicate at an off-site location regularly, ensuring availability regardless of possible risks and a full recovery of any potential lost data.

As this is a distributed facility, efforts are made to establish secure connections between the partners/RIs and the end-users of the data, as between the other WPs of the project and the nodes connected with the infrastructure, to ensure the security of what is transferred and used at all times. Certified public and virtual private encrypted connections are favoured, as point-to-point (P2P) connections cannot be guaranteed.

Currently, the quality of data and associated metadata check is a responsibility of the RIs and their contact persons/institutions.

The security requirement of the data/metadata storage is demanded to the RIs internal organization and their contact persons/institutions, who undertake to ensure the security of the data and its restoration in the event of a disaster.

The suggested strategy is to adopt the so-called 3-2-1 rule: there should be at least three copies of data, two different backup formats, and one backup stored off-site.

Although the 3-2-1 backup approach was developed about 20 years ago for the effective and long-term preservation of photo archives, the basic concepts are still the benchmark for data protection today.

In short, the 3-2-1 backup approach provides:

### **Three copies of the data**

Copies of all critical data (RAW) must be made at regular intervals. There should be at least three copies: original data and two backups.

### **Two types of storage**

Copies should be kept on two different types of storage: different machines to prevent the breakage of a single machine from affecting all copies of the data. The types of storage devices can be either in-house or cloud-based.

### **An external location**

At least one copy of the data must be stored on a storage medium located in a geographically different location from the other copies. This copy may be a replica of the primary storage site or an off-line (tape library type) copy of the data.

The 3-2-1 approach is a first step in defining a standardised procedure for secure data storage within the marine domain. A discussion is currently underway to define the technical details and some improved strategies to the approach to adapt it to an infrastructure with multiple geographically distributed nodes.

The actions taken and being defined ensure, as far as possible, that there are no single points of failure for the data.

All nodes of the distributed and federated system will be requested to prove that 3-2-1 approach for data secure and storage is adopted.

## **10.DATA FLOW**

The data/metadata produced by the ITINERIS marine domain RIs will be accessible through the data centre located at the ISMAR site in Naples and the access portal of the marine domain, as through the overall project HUB portal (coordination with WP2 will define the technical aspects of the interoperability of the two portals).

Through the marine domain portal, it will be possible to access the download of data and services of the marine domain (e.g.: data/metadata conversation or organization), through the ITINERIS metadata catalogue search interface.

Deliverable D5.1 details the ITINERIS Data and Metadata catalogues, their content, the rules and timing for their updating. Accordingly, when the system is operational the information (data/metadata catalogue) stored at the infrastructure will be updated as soon as recorded (Near Real Time – NRT), at regular intervals (e.g.: hours or daily) and/or according to the specific needs/requirements of the RIs in order to always guarantee the maximum efficiency of the system for data/metadata identification and retrieval.

All data/metadata present and/or accessible through the marine domain portal will follow FAIR principles and will be structured to ensure their interoperability according to the standards defined by the project and easily identifiable as products within the ITINERIS Project.

Where possible, data will be shared immediately after acquisition and/or production. Data quality assurance and control is one of the fundamental requirements for the proper use of the infrastructure by end users. At this early stage, control over the quality and validity of the data produced is left to the individual RIs and the partners who manage them.

At this stage it is not possible to assess the costs of making the data/metadata available and compliant with the FAIR principles. An assessment can be made once the long-term data/metadata retention policies and the precise flow of data/metadata have been established.

The main objective of the Data Centre is to provide access to services (metadata, data, ...), following FAIR principles, for all users, including the public and research infrastructure partners, and to collect data from data providers.

Following the second scenario, the most common one, reported in the D5.1 (Fig. 4a) a more detailed representation of the Intra- and Inter- Marine Domain Data Flow is presented (Fig.4b).

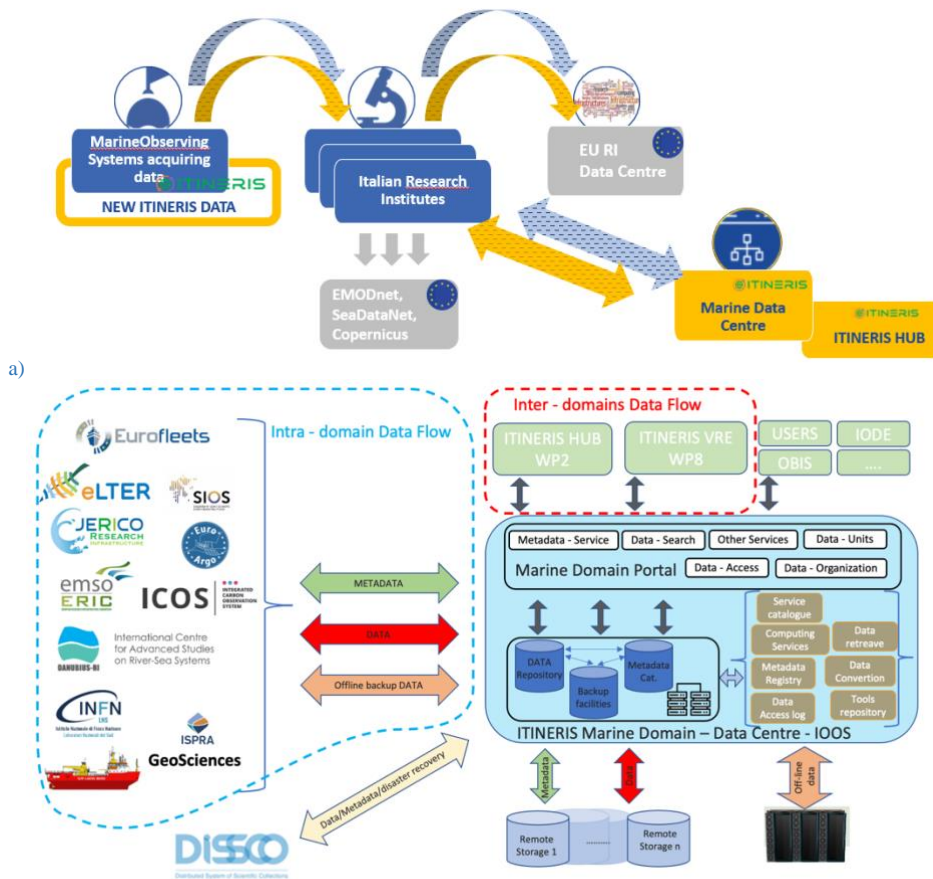
Two possible general configurations for the Intra- and Inter- Marine Domain Data Flow, and three External data retrieve flow, is presented to define the solution constrain for each RIs.

The scenarios proposed by D5.1 and D5.2 are completely equivalent when the Marine Data Center and the Marine Data Store represents the same technological and logical infrastructure (to considering equivalent to an Italian research Institute).

### 10.1.1 Intra-Marine Domain Data flow

Within a geographically dispersed and federated infrastructure, involving 11 different RIs (5 different research organisations) and different European projects with different degrees of progress, the flow of exchanged data is not always easy to define and is susceptible to limitations imposed by technical and policy constraints.

The marine domain project adopts a combination of two general configurations; a hybrid or circular solution will be implemented depending on particular requirements.



b) Fig.4 a) Scenario number two from D5.1 b) Intra-Marine Domain Data Flow. Scheme 1

The first scenario (Fig.4) provides that all the RIs involved in the Marine Domain (Intra-domain) exchange a copy of the data and metadata with the IOOS Marine Data centre and then the IOOS

Marine Domain takes care of the off-line backup copy of the RIs' data. This is the "trivial" solution in which the search, organization, use of data for analysis is simplified and centralised. It will be combined with the following configurations, shown below.

Should it be required, and necessary, the Marine Domain Data Centre may host part of the data (or provide a disaster recovery copy of the data) of RIs or data providers not involved in WP5.

The marine domain data centre guarantees connections with remote and offline storage; all services for downloading and visualization as for cataloguing and searching data/metadata, data access and request logs. In addition, it manages all the services of the marine data access portal that can be consulted by end-users and the direct connections with the general project portal (WP2) and with the VRE of WP8 (Inter- domain), as well as the global and pan-European data infrastructures (IODE system of NODCs, OBIS, SeaDataNet, EMODnet, Copernicus Marine Services).

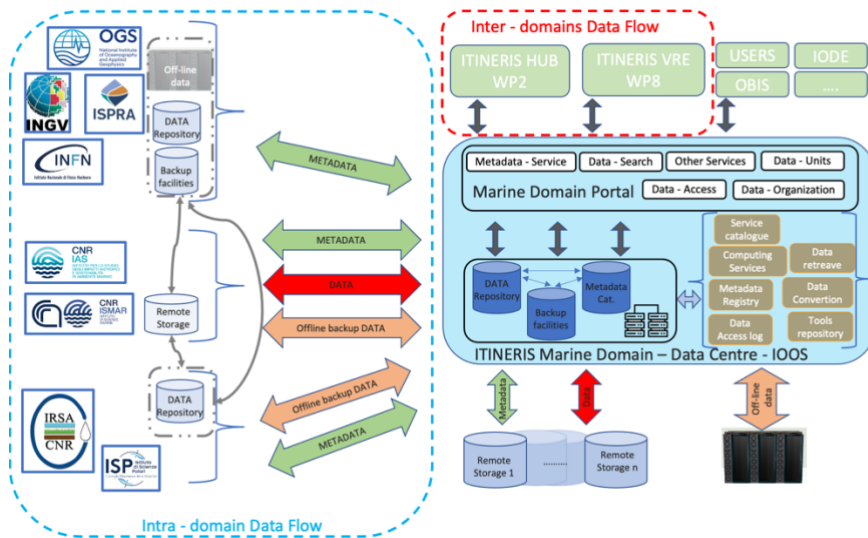


Fig.5 Intra-Marine Domain Data Flow. Scheme 2

The second scenario (Fig5.) differs from the previous one for the Intra-domain part. In this case, instead of the 11 RIs, the 5 different research institutions involved appear.

In this configuration, while all partners send the metadata to the data centre following the guidelines developed provided by the Marine Domains, only a few also send the data and make use of the possibility of off-line copying of the data.

All partners also provide access to a local copy of the data and possibly also to the off-line one, connecting all these systems with each other and with the data centre.

10.1.2 External data retrieve flows

Data requests, received through the Marine Domain data access portal, are handled by the data centre in different ways depending on the location of the requested data.

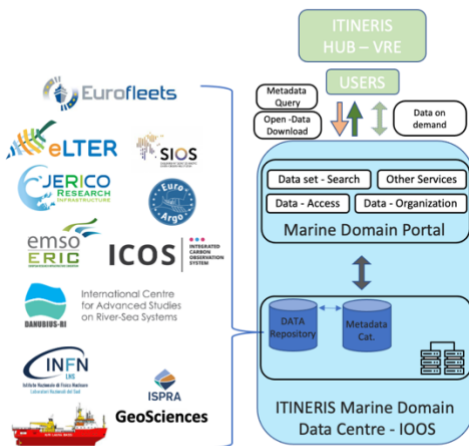


Fig.6 External data retrieve flow. Scheme 1

In the first theoretical case (Fig.6), all data are contained within the data centre's own repository. This is the simplest and quickest situation for the system to respond. The request, in fact, arrives at the data centre via a query on the portal, and the portal makes the requested data available for download after retrieval, without the need to query or retrieve data at the distributed nodes.

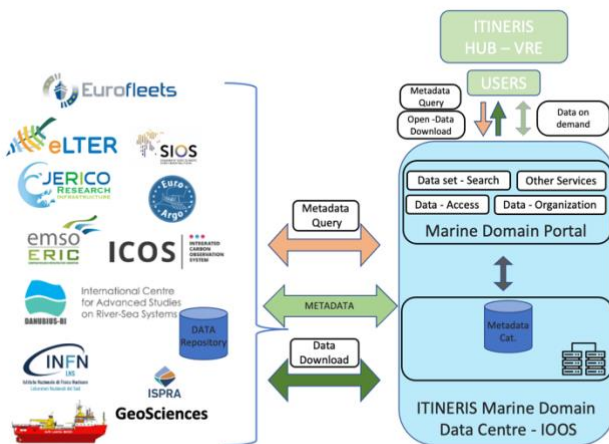


Fig.7 External data retrieve flow. Scheme 2

In Fig. 7 is reported the second case when part of the data resides on external repositories. These are consulted by the data centre, via a specific query, to temporarily acquire the data required to satisfy the end user's initial request.

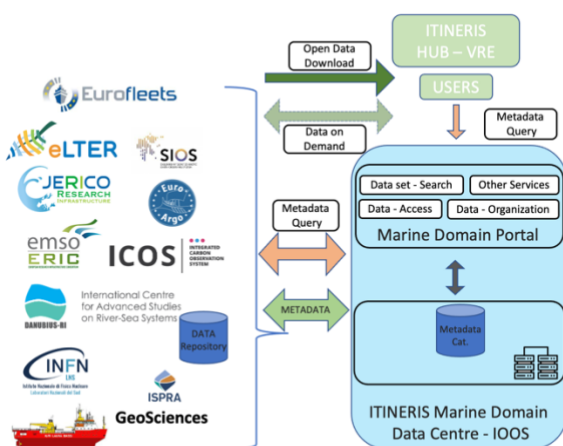


Fig.8 External data

retrieve flow. Scheme 3

In the third case (Fig.8), the data centre still queries remote nodes for the part of the data that is not available in its repository, differently from the previous case it no longer downloads any data locally. A set links to download the requested data directly from the external repository are send to the final user. This fully satisfies the Interoperability requirement of the FAIR principles.

The Data Centre will continuously work to fulfil the FAIR principles by offering data and data products that are retrievable, accessible, interoperable, and reusable. The constant testing and validation of new approaches to data/metadata transmission and preservation will enable it to serve as a national and European reference for FAIR initiatives at the European level. Given the wide range of areas covered by the different RIs in the marine domain, several metadata standards have been identified that are used and need to be harmonised, however, the commonly used ones are the ISO, EML, and OpenAire standards.

### 10.1.3 Exchange mechanisms for metadata

The ITINERIS project's marine domain data centre infrastructure stores and updates the metadata of all RIs and exposes it to the outside world using an interface capable of providing reachable and downloadable data information. The main interfaces with these features are OAI-PMH and OGC CSW. The former, in contrast to the others under evaluation, ensures ease of use and reflects the open-source principles of the project.

### 10.1.4 OAI-PMH

The Open Archives Initiatives Protocol for Metadata Harvesting (OAI-PMH) is the interface best suited to the needs of sharing metadata within marine domain nodes. The protocol is open, easily implemented, and well-supported by the scientific community with several additional tools.

#### 10.1.5 OGC CSW

The Open Geospatial Consortium Catalogue Services for the Web (OGC CSW) is a protocol capable of exposing the content of a catalogue using the eXtensible Markup Language (XML) and is more complex to implement than the OAI-PMH. However, it is the recommended exchange mechanism for metadata within the European INSPIRE framework.

#### 10.1.6 Exchange mechanisms for data

The infrastructure of the marine domain and its access portal is based on two key and indispensable aspects: interoperability at the metadata level and the ability to exchange (search and capture) data also in real time.

For data transfer, the FTP protocol is usually used regardless of the file format. However, the existence of different data transport protocols and the possibility of their coexistence does not necessarily favour one over the others.

#### 10.1.7 HTTP/FTP

Represents the simplest methodology for data exchange, but suffers from limitations due to the size of the data to be shared and the lack of precise standardization of formats.

#### 10.1.8 OPeNDAP

OPeNDAP (existing in different implementations e.g., THREDDS, ERDDAP, Hyrax, and pyDAP) is an open-source software considered extremely valid and versatile. It is capable of integrating flows from the various nodes of the distributed marine domain system if supported by appropriate data harmonisation. OPeNDAP also minimises overhead as no files are moved, the client connects to the data streams, reads the necessary data, and closes the connection.

#### 10.1.9 OGC WFS/WCS/WMS

OGC Web Feature Service (WFS), OGC Web Coverage Service (WCS), and OGC Web Mapping Service (WMS) are three protocols that, with different characteristics and peculiarities, allow the exchange and visualization of data

## *11. Marine data store “External Interfaces”*

### 12. INTERFACE WITH THE ITINERIS CENTRAL HUB WP2

The marine domain of the ITINERIS project (WP5) is the collector of all data/metadata from the 11 RIs involved in the project for the marine sphere. The data acquired by the different RIs, which ensure their quality and consistency, are harmonized and organized, according to the project standards and will be freely accessible to end users following the FAIR approach. The long-term organization and storage of marine variables will form the basis of the integrated national observation system and contribute to the EOOS and GOOS systems, meeting the main challenge of the United Nations Ocean Decade of Science for Sustainable Development.

In addition, WP5 is coordinating with WP2 (overall project hub), to ensure access to data/metadata from all RIs of marine interest. A discussion is underway between WP5 and WP2 to define the best strategy to proceed.

Several alternatives are currently being considered to ensure the availability of the sea domain data/metadata in coordination with the project coordination:

- 1) Access to the data/metadata will be primarily through the overall project hub (WP2) and a version of it, limited to marine data only, can be replicated to ensure autonomous access to the marine domain only
- 2) Data access by the overall hub (WP2) will be through the WP5 infrastructure that will ensure its metadata sharing and data access will

Other forms of shared data/metadata access are currently under discussion to define the most fruitful and technically sound approach

### 13. INTERFACE WITH THE ITINERIS VIRTUAL RESEARCH ENVIRONMENTAL WP8

The RIs and data acquired from different measurement systems, ships and autonomous mobile platforms (e.g.: Gliders, ...)

will be integrated and made visible by WP5 (establishing IOOS), which will interface with the central ITINERIS HUB and provide data for the development and implementation of services and tools for analysis, data visualisation and collaboration between researchers (VRE - WP8).

### 14. INTERFACE WITH THE ITINERIS TERRESTRIAL AND BIOSPHERE HUB WP6

As part of the sharing of best practices, a discussion with the RI DISSCO (formally in WP6 of the ITINERIS project and involving CNR ISMAR and IBBR, and Firenze University as partners) was initiated to coordinate and agree on the best strategies for defining the data flow (from production to dissemination), its secure storage, and its profitable use.

The possibility that a backup copy of locally produced data from the RI DISSCO could be hosted and managed by the WP5 infrastructure is currently being discussed and evaluated.

## 15. INTERFACE WITH THE COI/IODE/OBIS

Another relevant aspect of WP5's central role in the national and international management of marine data is its central role in the forming Italian Integrated Ocean Observation System (IOOS). The IOOS aims to coordinate national observations to improve the quality and interoperability of ocean data (including freshwater river outflows) for three critical themes: climate, operational services, and marine ecosystem health.

As part of the IOOS formation, WP5 will interface with the Italian Oceanographic Commission (COI), which plays a key role internationally in promoting and enhancing the Italian contribution to the Global Oceanographic Observing System (GOOS).

Discussions are currently underway as to whether the system developed by WP5 should place itself on the same level as other pre-existing national systems, but with more limited content, or take the opportunity offered by the ITINERIS project to define a new standard for a federated system of all marine data.

At the same time, possible interactions and/or integrations of the data/metadata (both at a policy and technical level) with the various international institutions dealing with marine data/metadata access to which some of the ITINERIS RIs individually belong are also being assessed.

Examples include: International Oceanographic Information and Data Exchange (IODE); Oceanographic Biodiversity Information System (OBIS).

The harmonisation of data/metadata and the products of the different RIs will form the basis for building an integrated system that can guarantee the continuity of data and services and meet the needs of users and the environmental challenges of the next 10 years.

## 16. FAIR data

The ITINERIS project aims to establish the Italian hub of research infrastructures in the environmental science domain that, among other things, will provide easy, fast and open access to ITINERIS data and other research results already existing and made available by national RIs distributed or produced through the project.

Similarly, WP5 will develop a similar infrastructure for the marine domain called IOOS (Italian Integrated Oceanographic Observing).

In both cases the two infrastructures, which will be inter-operating and synergic, will manage the integration of large-scale data through a network of nodes distributed throughout the national territory and federated, making sure that all data are discoverable.

As mentioned above, all metadata produced by the RIs will be collected and homogenised consistently with the needs of the marine domain and in accordance with the other WPs.

At present there is no shared global view on the standard to be used, although in the marine domain, as mentioned above, the use of the approach defined by SeaDataNET is widespread. The discussion, within the marine domain RIs and with the other WPs of the project, will continue taking also into account the INSPIRE requirements for data and metadata with the aim of creating comprehensive metadata to increase the reusability of the data.

At the marine domain level, efforts will be made to ensure that the different RIs are coordinated and inter-operating in the exchange of data/metadata to always guarantee cataloguing, access and dissemination according to FAIR principles.

The same principles used at the marine domain level (inter-operability between the marine domain RIs) will also be valid at a more general level between the different hubs and between WP5 and WP2.

Within the marine domain, data traceability is ensured through the adoption of persistent identifiers, such that information on instrumental measurements, processing, software, provenance is available. This approach ensures that data can be uniquely and unambiguously identified and also supports the citation process.

In compliance with the FAIR principles, both at the general ITINERIS project level and at the Marine Domain level the following points will be followed:

**Findability and accessibility:** all data will be deposited under a CC-BY 4.0 licence, using recommended file formats, metadata standards and PIDs. Open-source software will be preferred. To support this principle, controlled vocabularies will be used and field specifications will be provided for the metadata templates.

The data/metadata provided by the marine domain RIs will be harmonised/mapped and synchronised with the overall project hub and others known to access information.

Although the technical aspects for information access and transfer are still under discussion. However, the security, stability and multiplicity of access will always be guaranteed until shared protocols and standards are defined.

**Interoperability and reusability:** The interoperability of the data produced by the different RIs participating in the ITINERIS project is fundamental for the proper implementation of IOOS and its operation in combination with the other environmental domains. It is therefore necessary to define a common basis on which to build shared procedures. The basis of this process is:

1. forming a strong network across the environmental sub-domains;
2. promoting the creation, alignment and use of FAIR terminology and actively integrating the terminology into the operational infrastructure;
3. creating ITINERIS TERMINOLOGY as a service that enables adaptation and integration with existing producer and consumer data/metadata discovery architectures. Interoperability of semantic data/metadata has a direct impact on data findability, accessibility and reusability. Richly described metadata with a plurality of attributes and accurate and relevant vocabularies (controlled vocabularies, thesauri and ontologies), the use of open licences and the implementation of provenance also improve reusability.

## *17. Data Policy*

Please refer to ITINERIS data policy and to in D5.1 Marine Data Centre Management Plan: management needs, service requirements, governance for the distributed and federated system. This should be also exchanged with WP2 to be considered in the definition of the ITINERIS data policy.