



# IDEATION

## Reference architecture and Interoperability Guidelines (KER3) V1

Deliverable 4.1

WP4: Reference Architecture for interoperable inland water  
Digital Twin

Author: Mirko Gallone and Caterina Sarno

Date: 31 May 2025



Funded by  
the European Union



<b>GRANT AGREEMENT NUMBER</b>	101136799		
<b>ACRONYM / FULL TITLE</b>	IDEATION: InlanD watErs in the digitAl TwIn OceaN		
<b>START DATE</b>	01/06/2024	<b>DURATION</b>	24 months
<b>END DATE</b>	31/05/2026		
<b>PROJECT URL</b>	<a href="http://www.ideation-euproject.eu">www.ideation-euproject.eu</a>		
<b>DELIVERABLE</b>	D4.1		
<b>WORK PACKAGE</b>	4		
<b>CONTRACTUAL DATE OF DELIVERY</b>	31 May 2025		
<b>ACTUAL DATE OF DELIVERY</b>	31 May 2025		
<b>NATURE</b>	Other	<b>DISSEMINATION LEVEL</b>	Public
<b>LEAD BENEFICIARY</b>	ENG		
<b>RESPONSIBLE AUTHOR</b>	Mirko Gallone and Caterina Sarno		
<b>CONTRIBUTIONS FROM</b>	Natalia Zamora (BSC), Georgina Díez (BSC), Aberto Abella (FIWARE)		
<b>ABSTRACT</b>	This document is a report accompanying the first version of the Reference Architecture for an interoperable Inland Water Digital Twin. It begins by analyzing the structure of existing Digital Twins, focusing on the protocols and communication standards they use. The report then explores the key principles required to ensure interoperability. Finally, it presents the initial version of the proposed architecture.		

## Disclaimer

Any dissemination of results reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

## Copyright message

© IDEATION Consortium, 2025

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both. Reproduction is authorised provided the source is acknowledged.

## TABLE OF CONTENTS

<b>LIST OF ACRONYMS</b> .....	<b>5</b>
<b>EXECUTIVE SUMMARY</b> .....	<b>7</b>
<b>1. INTRODUCTION</b> .....	<b>10</b>
<b>2. DIGITAL TWINS</b> .....	<b>15</b>
2.1 Definition .....	15
2.2 Architecture analysis: other Digital Twins.....	16
2.3.4 EDITO-Infra .....	16
2.2.2 EDITO-Model Lab.....	18
2.3 State of the art data formats and protocols .....	21
2.3.1 Supported data formats for integration .....	21
2.3.2 Communication protocols and Data access methods.....	22
2.3.3 Guidelines for encoding and decoding mechanisms.....	22
2.3.4 Summary.....	22
2.4 State of the art definition of Integration layer .....	23
2.4.1 EDITO-Model Lab.....	23
2.4.2 EDITO Infrastructure.....	23
<b>3. INTEROPERABILITY PRACTICES AND FAIR PRINCIPLES</b> .....	<b>24</b>
3.1 Principles analysed .....	24
3.1.1 OpenML .....	24
3.1.2 DCAT/DCAT-AP .....	25
3.1.3 WaterML.....	27
3.1.4 MELODA5.....	29
3.1.5 FAIR principles .....	32
3.2 Shared semantics.....	35
3.3 Shared principles for Digital Twin .....	36
3.3.1 Semantic principles .....	36
3.3.2 Access to Data Principles.....	37
3.3.3 Organizational Principles.....	38
3.3.4 FAIR Principles .....	38



<b>4. REFERENCE ARCHITECTURE .....</b>	<b>41</b>
4.1 Main components.....	41
4.2 First version of IDEATION's Reference Architecture .....	43
4.2.1 Introduction.....	43
4.2.2 Data layer.....	44
4.2.3 DT Core .....	47
4.2.4 Interoperability hub.....	49
4.2.5 DT app.....	50
4.2.6 Infrastructure.....	52
4.3 Interoperability and compliance with the DTO.....	54
<b>5. CONCLUSIONS .....</b>	<b>57</b>
<b>REFERENCES.....</b>	<b>59</b>



## List of Acronyms

Short name / Acronym	Full name / Description
AI	Artificial Intelligence
API	Application programming Interface
ARCO	Analysis Reay Cloud Optimized
CF	Climate and Forecast
CI/CD	Continuous Integration and Deployment
CSW	Catalog Web Service
DCAT	Data Catalog Vocabulary
DCAT-AP	Data Catalog Vocabulary-Application Profile
DestinE	Destination Earth
DT	Digital Twin
DTO	Digital Twin Ocean
EDITO	European Dlgital Twin Ocean
EDITO-Infra	European Dlgital Twin Ocean - Infrastructure
EDITO-ML	European Dlgital Twin Ocean - Model Lab
EU	European Union
FAIR	Findable Accessible Interoperable Reusable
GUI	Graphical User Interface
HPC	High-Performance Computing
MELIODA5	MEtric for the evaLUation of Open DATA
ML	Machine Learning



MSF	Multi-Stakeholder Forum
O&M	Observation and Measurement
OGC	Open Geospatial Consortium
OpenML	Open Machine Learning
RAG	Retrieval Augmented Generation
RDF	Resource Description Framework
STAC	Spatio Temporal Asset Catalog
W3C	World Wide Web Consortium
WaterML	Water Markup Language
WFS	Web Feature Service
WMS	Web Map Service
WP	Work Package

## EXECUTIVE SUMMARY

This document represents the first tangible outcome of Work Package 4 (WP4) of the IDEATION project, which is dedicated to defining a reference architecture and interoperability guidelines for the future Digital Twin of inland waters. The overall objective of WP4 is to ensure that the inland waters Digital Twin is fully compatible and integrable with the Digital Twin Ocean (DTO), and eventually with other European Digital Twins. However, this deliverable does not aim to cover the entire scope of WP4; instead, it focuses on the results of Tasks T4.1, T4.2, and T4.4, namely, the technical specifications, interoperability practices, and architectural definition. The contents foreseen in Task T4.3, which focuses on the development of guidelines to support the accuracy, reliability, and applicability of models and algorithms for the creation of robust Digital Twins, are not included here. That task will be addressed in Deliverable *D4.3 Guidelines for the design and validation of physical simulation models and AI algorithms*.

To provide a clear foundation, the document introduces the concept of the Digital Twin, defined as a high-fidelity, continuously updated digital replica of a physical system. In the context of inland waters, the Digital Twin aims to support monitoring, simulation, decision-making, and scenario testing by integrating real-time data, physical models, and machine learning algorithms. The Digital Twin Ocean, currently under development within the EDITO framework, serves as a key reference model. It offers a modular, scalable infrastructure for simulating ocean conditions, accessing environmental data, and supporting user-facing applications.

One of the foundational steps in the development of the IDEATION Reference Architecture involved a detailed analysis of the architecture and components of the DTO, which currently represents the reference for European Digital Twins in the water domain. The DTO is being developed under the broader EDITO initiative, which is composed of two key components: EDITO-Infra and EDITO-Model Lab.

EDITO-Infra provides the entire infrastructure that supports the DTO. It includes a data lake, capable of storing large volumes of heterogeneous data, as well as data ingestion and access mechanisms. It also provides core services for data management, security, user authentication, and data sharing, enabling both machine-to-machine communication and human interaction through APIs and user interfaces. The infrastructure is designed to be scalable and modular, supporting high-performance computing (HPC) and cloud-based deployment models to serve diverse simulation and visualization needs. EDITO-Model Lab, on the other hand, focuses on the development, execution, and orchestration of environmental models within the DTO ecosystem. It provides tools and services for model integration, scenario testing, and simulation workflows, offering researchers and stakeholders a flexible platform for deploying and combining models. Its



architecture supports containerization, versioning, provenance tracking, and the definition of experiment pipelines, which are key to reproducibility and transparency. As part of this analysis, the main components of both EDITO-Infra and EDITO-Model Lab were identified and mapped in terms of their roles and technical functions. This mapping exercise provided a comprehensive understanding of how data and models are handled within the DTO framework. These insights directly informed the design decisions for the IDEATION Reference Architecture, ensuring that the inland waters Digital Twin can be developed in alignment with, and fully compatible with, the DTO environment.

A central concept addressed in this deliverable is interoperability, which is essential for the success of any Digital Twin initiative operating in a multi-system, multi-stakeholder European environment. Interoperability, in this context, is not limited to the mere exchange of data between systems, it encompasses a broader set of capabilities that ensure information can be understood, interpreted, and reused across diverse technological platforms, organizational boundaries, and domain-specific applications. In line with this definition, the IDEATION project, through Tasks T4.1 and T4.2, has conducted a systematic review and selection of widely accepted principles, models, and standards to ensure a high level of interoperability between the inland waters Digital Twin and external systems.

To support semantic alignment and data harmonization, the deliverable adopts several reference frameworks, most notably: the FAIR principles (Findable, Accessible, Interoperable, Reusable), the DCAT-AP metadata standard, WaterML, OpenML and MELODA5.

The culmination of the work presented in this deliverable is the first version of the IDEATION Reference Architecture, which defines the high-level structure, components, and integration strategies for the inland waters Digital Twin. This architecture was developed based on the benchmarking of the DTO (especially EDITO-Infra and EDITO-Model Lab), and the design principles consolidated through Tasks T4.1 and T4.2. It is explicitly conceived not as a rigid or monolithic system, but as a modular, scalable, and microservice-based framework that can adapt to different implementation contexts and evolve alongside the broader European Digital Twin ecosystem. The IDEATION Reference Architecture is organized into a set of conceptual layers, each designed to fulfill a specific role in the functioning of the Digital Twin. At its foundation lies the Data Layer, responsible for the ingestion, storage, and cataloging of diverse datasets while ensuring data integrity and accessibility. Built on top of this, the Interoperability Layer acts as the connective interface between the Digital Twin and external systems, such as the DTO, by managing shared catalogs and supporting standardized communication protocols. The DT Core encompasses the internal logic and processing pipelines that execute analytics, simulations, and forecasts, while the DT App layer provides user-facing services and tools to interact with the system. Finally, the entire architecture



[www.ideation-project.eu](http://www.ideation-project.eu)

runs on a flexible and scalable infrastructure layer, capable of supporting cloud-native deployment and high-performance computing when needed.

The next phase will see an even further development of IDEATION's Digital Twin, which will be accomplished by studying a broader suite of technologies and by introducing additional Digital Twins coming from other EU's projects like Destination Earth (DestinE). Moreover, the roadmap will remain dynamic, adapting to the concrete requirements and innovation gaps identified during future Multi-Stakeholder Forums (MSFs), ensuring that each new capability is always capable of supporting stakeholders's priorities and real-world use cases.

## 1. Introduction

This deliverable is dedicated to laying the foundational framework for the architecture to be defined along the WP4 of the IDEATION project. Particularly aiming to define the reference architecture and interoperability guidelines. In this document the outputs of three main tasks are reported thoroughly. The related tasks (T) are: T4.1 – Technical specifications and data models for seamless integration with the Digital Twin Ocean<sup>1</sup> (DTO) and Destination Earth (DestinE) Digital Twins; T4.2 – Collection of data interoperability good practices and FAIR principles (Wilkinson et al. 2016); and T4.4 – Design of a high-level reference architecture for an optimal inland water Digital Twin.

During this first project phase, one of the key activities concerned the identification, analysis, and evolution of the main components of the system architecture, with particular reference to the overall structure of the Digital Twin. This work did not take place in an abstract manner or disconnected from the existing landscape but was based on a careful and thorough reconnaissance of the architectures already present, both academic and industrial, with particular emphasis on the Digital Twin Ocean (DTO), whose architectural proposal emerged as a solid and internationally recognised reference point. Through a comparative analysis, the recurring patterns and technological components common to many already published implementations of Digital Twin in both urban and industrial settings were identified. This phase allowed us not only to understand the historical evolution of these architectures, but also to identify the most frequent criticalities (e.g. data fragmentation, poor interoperability between modules, lack of user-friendly tools for interaction) and the strengths on which to base our proposal.

In all Digital Twins, regardless of the application domain, there are some fundamental requirements that drive their design from the earliest stages. As illustrated in Figure 1, these requirements can be grouped into five main categories, which reflect the core functional and structural needs that a Digital Twin is expected to fulfill.

The first foundational pillar in the development of a Digital Twin is Data Collection and Management. This component plays a crucial role in enabling the system to ingest, organize, and preserve the vast amount of information that constitutes the basis for simulation, analysis, and decision-making. In a typical Digital Twin architecture, data may originate from a wide range of heterogeneous sources, often varying in structure, update frequency, format, and semantics. These sources can include both historical datasets and real-time data streams. The ability to harmonize and manage such diversity is essential to ensure consistency and reliability across the digital representation. Equally important is the implementation of advanced metadata cataloging strategies. Without a well-structured

---

<sup>1</sup> Mercator Ocean International. (n.d.). Digital Twin Ocean Overview. Retrieved from <https://digitaltwinoccean.mercator-ocean.eu/>

system for annotating and indexing datasets, it becomes impossible to ensure the traceability, interoperability, and reusability of data throughout the lifecycle of the Digital Twin. In this sense, metadata management is not a peripheral concern but a central element that directly impacts the quality and sustainability of the digital ecosystem.

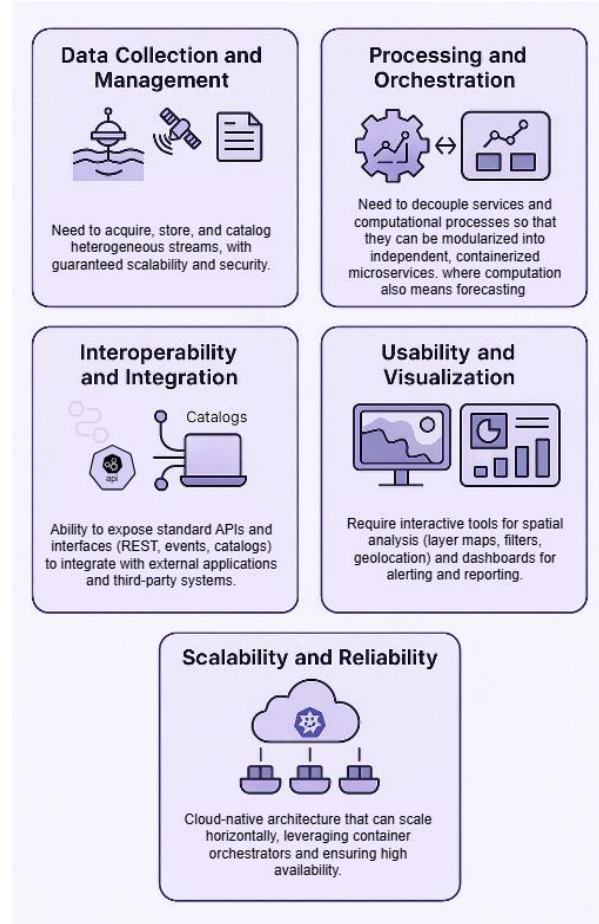


Figure 1 - Main features of the Digital Twin

A second fundamental building block of any Digital Twin platform is Interoperability and Integration. In a technological landscape that is increasingly distributed, federated, and collaborative, the ability to interoperate with external systems, platforms, and data sources is not merely a desirable feature—it is a critical requirement. Digital Twins must be capable of exposing and consuming standard interfaces, such as RESTful APIs, event-driven communication protocols, and shared catalogs for data, services, and processes. These mechanisms enable seamless integration between different actors and infrastructures, promoting the reuse of existing assets and supporting the construction of modular, scalable, and vendor-neutral solutions. In the context of the *Ideation* project, this requirement



becomes even more prominent. Given the necessity to ensure high levels of interoperability with the DTO, it was essential to design a reference architecture capable of enabling robust and efficient integration. To this end, the system's logical structure must be based on components that can operate in federated environments, publish and retrieve information through standardized and cataloged interfaces, and adapt dynamically to evolving interoperability scenarios.

The third key building block in a Digital Twin architecture is Processing and Orchestration. This component is responsible for transforming raw data into meaningful insights through analytical models, simulations, forecasting algorithms, and logic workflows that support automated or semi-automated decision-making processes. The DT must therefore provide for the possibility of coordinating services that are also very heterogeneous in terms of purpose, frequency and computational logic. These services must be able to interact via standard APIs and messaging protocols, making the results of processing accessible to other modules of the system or to external parties in a transparent and consistent manner. It is fundamental that the processing module is not a closed or static unit but, on the contrary, represents an intelligent and flexible node within a dynamic and interoperable environment, capable of evolving with the application domain and user needs.

The Usability and Visualization layer plays a critical role in determining the success and operational impact of a Digital Twin. It is not sufficient for data to be accurately collected and processed—what truly empowers users is the ability to explore, understand, and act upon the resulting information through intuitive and responsive interfaces. In this regard, the user interface is not a superficial front-end, but a full-fledged environment for interacting with the system. It must be designed to serve both technical and non-technical users, offering interaction paradigms that allow for clear interpretation and decision-making based on evidence. These tools should make it easy to navigate complex datasets, compare modeled scenarios and real observations, and access contextualized information in real time. At the same time, more analytical modules such as dashboards, alerting panels, and reporting tools enable users to monitor the evolution of phenomena, receive automatic notifications when critical conditions arise, and generate customized analyses to share with other stakeholders. To support this, a number of key non-functional requirements come into play, most notably system reliability and fault tolerance. These requirements are essential to ensure that the platform remains responsive and operational even under high load or in the presence of partial failures, particularly in distributed environments. In the context of the IDEATION project, the functional and non-functional requirements identified so far are detailed in Deliverable *D2.3: MSF results and functional and non-functional requirements (KER1) V1*.

Ultimately, it is essential that all these requirements are fulfilled within a system environment that is both scalable and reliable.

The key conceptual blocks identified as common requirements for Digital Twins, act not only guide the technical design choices but also provide the conceptual guideline across the document's sections.

The structure of the deliverable reflects this layered approach, progressing from theoretical foundations to the architecture definition:

- Chapter 2 lays the groundwork by introducing the concept of the Digital Twin and positioning the IDEATION initiative within the broader European context. It provides a detailed architectural analysis of prominent reference frameworks, specifically EDITO-Infra and EDITO-Model Lab, which underpin the Digital Twin Ocean. These analyses highlight technological components, design patterns, and integration strategies that have informed the development of IDEATION's architecture. The chapter then transitions into a technical discussion of data formats, communication protocols, and encoding/decoding mechanisms, establishing a technical vocabulary that supports interoperability.
- Chapter 3 expands on the notion of interoperability and introduces the standards and principles necessary to implement it in practice. It reviews prominent frameworks such as OpenML, WaterML, MELODA5, and the FAIR principles, each evaluated in terms of their relevance for cross-domain Digital Twin integration. The chapter concludes by formalizing a set of shared semantics and principles for data access, organizational governance, and reusability, thereby offering a conceptual foundation that supports technical integration.
- Chapter 4 builds on this foundation to present the first version of the IDEATION Reference Architecture. This section defines the main architectural components, *Data Layer*, *DT Core*, *Interoperability Hub*, *DT App*, and *Infrastructure*, each designed to meet the analyses derived from the previous chapter. The architecture follows a microservice-based model, enabling modularity, resilience, and scalability. Special emphasis is placed on how this architecture ensures compliance and compatibility with the DTO, particularly through shared interfaces, standardized APIs, and common metadata catalogs.
- Chapter 5 provides conclusions and a forward-looking roadmap, highlighting how the presented work will support future development stages. It links the current results to subsequent deliverables.

In this deliverable, several key elements have been addressed to shape a comprehensive, efficient, and needs-driven architecture. Notably, the requirements outlined in the deliverable D2.3: MSF results and functional and non-functional requirements (KER1) V1 were considered, serving as a foundation for both functional and structural decisions to ensure alignment with the operational and technical needs identified during the initial project phases. One of the components that has been integrated is the OpenKIWAS, as



[www.ideation-project.eu](http://www.ideation-project.eu)

described in deliverable D3.1: OpenKIWAS (KER2) V1. This integration not only builds upon typical components to retrieve and process data but also introduces a structured set of initial information to support further development.



**IDEATION - D4.1: Reference architecture and Interoperability  
Guidelines (KER3) V1**

## 2. Digital Twins

### 2.1 Definition

A Digital Twin, in the context of the European Union and initiatives such as Destination Earth<sup>2</sup>, is a high-accurate, continuously updated digital replica of a physical system (such as the ocean, land, infrastructure, or ecosystems). It integrates real-time data, physical models, artificial intelligence, and semantic frameworks to monitor, simulate, and predict the behavior of its real-world counterpart, ultimately supporting informed decision-making and policy implementation<sup>3</sup>.

Several initiatives exist that are related to IDEATION. For instance, the Digital Twin Ocean DTO (and the European DTO (EDITO))<sup>4</sup> is a data-driven simulation model that combines real-time and historical data from various sources, including satellites, ocean sensors, research vessels, and underwater drones. The DTO is designed to mirror the physical characteristics, processes, and dynamics of the actual ocean environment, allowing scientists, researchers, and policymakers to gain valuable insights and make informed decisions regarding ocean health and management. It incorporates cutting-edge technologies, such as artificial intelligence, machine learning, and big data analytics to process and analyse vast amounts of information continually. This enables, for instance, to provide accurate and up-to-date information on ocean temperatures, currents, salinity, marine life distributions, and other essential parameters that impact the Earth's climate and biodiversity. Following such goals, the EDITO initiative will feature a central core DTO that serves as the foundation, offering a vast repository of data, general ocean models, and AI processing toolkits. Built on this foundation, a wide range of customized applications or "local twins" can be integrated<sup>5</sup>.

In addition, the DestinE, a flagship initiative of the European Commission aimed at developing a highly precise Digital Twin of the Earth. This ambitious effort integrates advanced digital modelling, real-time data, and High-Performance Computing (HPC) to simulate and monitor natural processes, environmental hazards, and human activities. The resulting digital environment empowers users to devise targeted adaptation strategies and effective mitigation measures based on accurate, data-driven insights. DestinE represents a

---

<sup>2</sup> <https://destination-earth.eu/>

<sup>3</sup> <https://destination-earth.eu/glossary/>

<sup>4</sup> <https://www.edito.eu/>

<sup>5</sup> [https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/eu-missions-horizon-europe/restore-our-ocean-and-waters/european-digital-twin-ocean-european-dto\\_en#what-can-we-use-the-digital-twin-ocean-for](https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/eu-missions-horizon-europe/restore-our-ocean-and-waters/european-digital-twin-ocean-european-dto_en#what-can-we-use-the-digital-twin-ocean-for)

paradigm shift in the way we interact with Earth system data—offering unprecedented levels of accuracy, spatial and temporal detail, and interactive access to information. It leverages the full capabilities of Europe’s supercomputing infrastructure to support real-time analysis and forecasting.

As a cornerstone of the EU’s Green Deal and Digital Strategy, DestinE also provides a reference architecture for the development of interoperable, domain-specific Digital Twins—such as those for inland waters, cities, or biodiversity. This architecture defines common principles, data models, and technical frameworks to ensure consistency, scalability, and seamless integration across the broader Digital Twin ecosystem in Europe. Examples of the impact of e.g. EDITO project can have and the need to leverage such requirements can be found in their website<sup>6</sup>. As for DestinE aims for further enhancement of the DestinE system by 2026 and a full replica of the Earth by 2030<sup>7</sup> integration of additional Digital Twin and related services. As the DestinE project evolves, and further information is released, an update will be given in D4.2 "Reference architecture and Interoperability Guidelines (KER3) V2" that will be submitted on M21 (February 2026).

Particularly, IDEATION aims to prepare the development of the Digital Twin of the inland waters (urban, water, coastal waters, groundwater, rivers, lakes, reservoirs, wetlands, and snow and ice) addressing activities to be developed and to make it integrated and interoperable with the DTO for a unified Digital Twin of ocean and waters, addressing the hydrosphere as a whole. It will be achieved by following the main conceptual framework of Digital Twins.

## *2.2 Architecture analysis: other Digital Twins*

We present some core elements of different Digital Twins that will serve as foundation for the architecture defined in the next sections.

### 2.3.4 EDITO-Infra

The EDITO-Infra (European Digital Twin of the Ocean - Infrastructure)<sup>8</sup> is a state-of-the-art infrastructure under development designed to support ocean science by facilitating advanced modeling, data sharing, and collaboration. It serves as a cornerstone for the broader DTO initiative, which aims to create a real-time, high-resolution virtual model of the ocean by integrating data from sensors, satellites, and advanced models. This virtual representation enhances the monitoring and management of ocean environments,

---

<sup>6</sup> <https://destination-earth.eu/news/destine-joins-forces-with-the-national-observatory-of-athens/>

<sup>7</sup> <https://destination-earth.eu/>

<sup>8</sup> <https://edito-infra.eu/>

providing crucial insights into marine biodiversity, ocean health, and climate change adaptation.

The EDITO-Infra project effectively combines cutting-edge and proven technologies to create an accessible, FAIR-compliant ecosystem for oceanographic science. This scalable and secure infrastructure supports both researchers and policymakers, enabling transformative insights into ocean systems.

### 2.2.1.1 Core Components

- **EDITO Data Lake.** The EDITO Data Lake is a centralized storage aggregating public and private datasets, such as Copernicus Marine Service<sup>9</sup> and EMODnet<sup>10</sup> data. It supports ARCO (Analysis Ready Cloud Optimized)<sup>11</sup> formats and RESTful STAC APIs<sup>12</sup> for efficient data access. Acts as a federated platform, allowing global and local data contributions.
- **EDITO Engine.** It hosts computational functions and the Process Registry, enabling simulations and data generation through standardized Open Geospatial Consortium (OGC) APIs. Integrates AI-driven ocean modelling for enhanced analysis and predictive capabilities.
- **EDITO Self-Services.** It provides user-customizable workspaces with tools for data processing and model validation. Leverages near-data computing to reduce data transfer, optimizing real-time analytics for marine governance, biodiversity conservation, and disaster-risk management.

### 2.2.1.2 Design and Deployment

- **Cloud and Orchestration.** The EDITO-Infra platform uses OpenStack for cloud infrastructure and Kubernetes for container orchestration. The infrastructure is managed using Terraform for scalability and flexibility across different cloud providers, and docker containerization ensures consistency and portability.
- **Monitoring and Backup.** Monitoring is carried out using Prometheus and visualization via Grafana. The platform has implemented a backup system that uses Restic for secure data archiving on external storage.
- **Development and CI/CD.** EDITO uses a git-based version control system from GitLab, whose pipeline's components support continuous integration and deployment (CI/CD), ensuring iterative development and testing in staging and production environments.

---

<sup>9</sup> <https://marine.copernicus.eu/>

<sup>10</sup> <https://emodnet.ec.europa.eu/en>

<sup>11</sup> <https://doi.org/10.3389/fclim.2021.782909>

<sup>12</sup> <https://github.com/radiantearth/stac-api-spec>

### 2.2.1.3 Security

Security access control is done via Keycloak and OpenID Connect for robust authentication. The data within the platform is encrypted, and isolated namespaces are created for user activity tracking in Kubernetes.

### 2.2.1.4 Technical and Methodological Principles

The implemented infrastructure has adopted "boring technology" for reliability and stability over cutting-edge but unproven tools, incorporating the "12-factor app" methodology for cloud-native development, giving emphasis on scalability, configuration management, and separation of environments.

### 2.2.1.5 Key Technologies

- Data Management: ZARR format for multidimensional arrays, MinIO for private object storage.
- Geospatial Standards: STAC for metadata, OGC API Processes for interoperable geospatial processing.
- Collaboration Tools: Onyxia enables self-service computational environments tailored to scientists' needs.

### 2.2.1.6 Challenges and Innovations

The infrastructure addresses the increasing complexity of ocean data by enabling "what-if" scenarios and near-data processing. Its modular design fosters community contributions to data, models, and services, enhancing collaborative research.

## 2.2.2 EDITO-Model Lab

The EDITO-Model Lab (EDITO-ML) is a key initiative within the European DTO framework, designed to advance ocean numerical modeling and foster collaboration among researchers, policymakers, and industry stakeholders. Its primary objective is to enhance understanding, monitoring, and sustainable management of marine environments by developing high-resolution ocean simulations that integrate complex physical, chemical, biological, and ecological processes.

By leveraging cutting-edge modeling techniques, data assimilation, and artificial intelligence, EDITO-Model Lab enables the scientific community to simulate various environmental scenarios, such as the impact of climate change, pollution, and mitigation strategies. These "what-if" simulations provide critical insights for decision-makers, supporting marine governance, biodiversity protection, and disaster risk management.

## 2.2.2.1 Core Components

**Virtual Ocean Model Lab (VOML).** It is a comprehensive platform for developing and testing DTO software components. It combines a collaborative development environment with advanced computational infrastructure.

### Relocatable Ocean Modeling Platforms

- **SURF:** Uses structured and unstructured grids for high-resolution ocean forecasts. Integrates seamlessly with large-scale models like NEMO for global and regional domains and SHYFEM for near-shore and estuarial applications.
- **HBM (HBMos):** A two-way nested ocean-ice model supporting dynamic setups for high-resolution (up to tens of meters) simulations in the Baltic-North Sea region.

**Autosubmit Workflow Manager.** This tool helps in orchestrating experiments and integrates cloud and HPC resources for efficient workflow management. Autosubmit uses YAML-based configurations and SSH connections for secure job execution across multiple platforms, such as EuroHPC systems (Leonardo and MareNostrum).

## 2.2.2.2 Computational Infrastructure

EDITO-Model Lab harnesses the power of EuroHPC<sup>13</sup> platforms to drive its high-performance simulations. Leonardo<sup>14</sup>, a pre-exascale supercomputer equipped with NVIDIA Tensor Core GPUs and Intel Sapphire Rapids CPUs, provides the necessary computational power for advanced ocean modeling. MareNostrum5<sup>15</sup>, another pre-exascale supercomputer located in Barcelona Supercomputing Center (BSC-CNS) facilities, complements this infrastructure by offering general-purpose and accelerated partitions optimized for AI and HPC workloads. Autosubmit ensures seamless interoperability between these HPC systems and cloud environments, facilitating scalable and robust ocean simulations.

To support collaborative development, GitLab is employed as a shared version control system, enabling co-development, continuous integration/deployment (CI/CD), and automated deployments. Hardened GitLab runners further enhance security for code execution on HPC clusters, reinforcing the principles of open and reproducible science.

## 2.2.2.3 User Interfaces

User interaction with EDITO-Model Lab's infrastructure is facilitated through a suite of Graphical User Interfaces (GUIs).

- The Autosubmit GUI provides real-time monitoring and log visualization for workflows running on HPC and cloud systems.

---

<sup>13</sup> [https://eurohpc-ju.europa.eu/about\\_en](https://eurohpc-ju.europa.eu/about_en)

<sup>14</sup> <https://leonardo-supercomputer.cineca.eu/>

<sup>15</sup> <https://www.bsc.es/marenostrum/marenostrum-5>

- The SURF GUI allows users to configure, manage, and visualize high-resolution ocean models intuitively.
- While the HBM GUI is designed for on-demand modeling, supporting user-defined subdomains with automated input generation and post-processing tools.

#### 2.2.2.4 Methodologies and Tools

EDITO-Model Lab employs relocatable modeling methodologies that enable dynamic downscaling from global to coastal models. The SURF and HBM platforms support multiple nesting levels to ensure high-resolution outputs, leveraging data sources such as Copernicus Marine.

In terms of data and workflow management, integration with EDITO-Infra ensures seamless access to data lakes, computational functions, and APIs for efficient data sharing and visualization. Autosubmit further enhances reproducibility and efficiency by automating experiment execution and monitoring processes.

#### 2.2.2.5 Focus Applications and What-if Scenarios

As an outcome of the EDITO-Model Lab project, the Focus Applications (FAs)<sup>16</sup> and What-if Scenarios (WiSs)<sup>17</sup> represent the primary demonstrators of the platform's capability to address real-world ocean challenges.

The Focus Applications are designed as concrete use cases that highlight the platform's added value across diverse maritime domains such as coastal resilience, biodiversity, and maritime spatial planning. Each application integrates advanced digital twin capabilities, coupling multi-source data with high-resolution models. Complementing these, the What-if Scenarios explore hypothetical but plausible futures, simulating the impacts of specific decisions or external drivers—such as regulatory changes or climate stressors—on the marine environment.

Together, the FAs and WiSs form the core narrative of EDITO-ModelLab's ambition to support informed, data-driven decision-making in marine policy and management.

#### 2.2.2.6 Challenges and Innovations

EDITO-Model Lab is designed to be both flexible and accessible, catering to diverse user needs from global-scale analysis to highly localized modeling. By providing intuitive GUIs and APIs, it simplifies interactions with advanced computational resources, making high-performance ocean modeling more accessible to a broad range of stakeholders.

Collaboration is a fundamental aspect of the lab's approach, as it fosters interdisciplinary cooperation between oceanographers, climate scientists, and data analysts. By adhering to

<sup>16</sup> <https://www.edito-modellab.eu/news/what-is-the-added-value-of-the-edito-model-lab-focus-applications-nbsp>

<sup>17</sup> <https://www.edito-modellab.eu/news/what-can-you-do-with-the-edito-model-lab-what-if-scenarios>

FAIR data standards, EDITO-Model Lab promotes open science and ensures that its datasets and models can be widely utilized and shared.

Scalability is another key innovation, as the lab is designed to accommodate growing computational and data demands while maintaining robust performance across diverse platforms. This adaptability ensures that EDITO-Model Lab remains at the forefront of ocean modeling and Digital Twin technology.

### 2.2.2.7 Final remarks

EDITO-Model Lab plays a crucial role in the European DTO initiative by developing next-generation ocean numerical models that drive innovation in oceanographic research and decision-making. Its integration of high-resolution simulations, AI-driven analysis, and HPC infrastructure ensures that stakeholders have access to state-of-the-art tools for monitoring and managing marine environments. By bridging the gap between scientific research and real-world applications, EDITO-Model Lab contributes significantly to sustainable ocean management and informed policymaking in the face of climate change and environmental challenges.

### *2.3 State of the art data formats and protocols*

Based on the information for the EDITO-Infra and EDITO-Model Lab, the relevant information regarding data formats, communication protocols, and encoding/decoding mechanisms is as follows.

#### 2.3.1 Supported data formats for integration

For structured data, metadata is primarily managed using JSON via the SpatioTemporal Asset Catalog (STAC) API, which supports semantic, metadata-driven queries. Scientific datasets, especially those involving multi-dimensional arrays such as ocean modeling outputs, are commonly stored in the ZARR format. This aligns with prevailing oceanographic and geospatial standards, including the CF (Climate and Forecast) Conventions and the ARCO (Analysis Ready Cloud Optimized) specification, both of which enhance compatibility with cloud-native workflows.

The EDITO-Infra Data Lake integrates data from major providers like Copernicus Marine Services and EMODnet, supporting widely accepted formats such as NetCDF and TIFF. These formats ensure that both gridded and raster datasets can be incorporated without the need for complex conversion.

## 2.3.2 Communication protocols and Data access methods

Communication within the system relies heavily on RESTful APIs. The STAC API serves as a primary interface for accessing spatiotemporal datasets, while the OGC API - Processes standard enables web-based interaction with geospatial processing services. Tools like Autosubmit employ HTTP for front-end to back-end communication, while SSH is used to securely manage remote jobs on HPC systems. Background tasks are handled using mechanisms such as Cron jobs and HTTP workers, ensuring efficient execution of recurring processes.

Data access and workflow management are facilitated through Autosubmit's API, local SQLite databases, and object storage systems like MinIO, which offers S3 compatibility. These components work together to manage configurations, track metadata, and store user data effectively.

## 2.3.3 Guidelines for encoding and decoding mechanisms

Regarding encoding and decoding, JSON remains the dominant format for API communication and metadata exchange. YAML is used specifically for workflow configuration within Autosubmit.

Responses from APIs are structured in JSON, promoting compatibility with common web and data processing tools. Numerical outputs from HPC model runs are typically handled in NetCDF, which integrates well with Python-based tools such as Xarray and Matplotlib for analysis and visualization.

Interoperability is further supported by Python libraries like Flask, which are used for handling API endpoints, and by specific mechanisms for managing nested model data. For example, in relocatable ocean models like SURF and HBM, boundary conditions and grid structures are encoded to be fully compatible with parent systems such as Copernicus Marine models.

## 2.3.4 Summary

The EDITO systems prioritize interoperability and use:

- JSON and YAML for metadata and configurations.
- REST APIs for communication.
- Standard geospatial and scientific formats (STAC, ZARR, CF Conventions) for data exchange (see description above). These standards ensure robust integration of diverse datasets and tools in a unified digital ocean modeling environment.

## 2.4 State of the art definition of Integration layer

Interoperability within the EDITO - Infra and EDITO - Model Lab projects refers to the ability of systems, data, and services to work together seamlessly across platforms and stakeholders. It is supported by standard data models, open interfaces, harmonized vocabularies, and shared protocols, enabling integration, discoverability, and collaboration within a unified digital ocean framework.

Both projects emphasize a modular, open architecture with standardization at its core, and addressing interoperability as described in the following subsections.

### 2.4.1 EDITO-Model Lab

- Advocates co-design of models and interfaces to ensure alignment between data providers and model developers.
- Focuses on developing reference use cases that test interoperability across services and data inputs.
- Promotes the use of FAIR principles as a guided methodology.
- Supports standardized metadata schemas and semantic alignment for better model-data interoperability.

### 2.4.2 EDITO Infrastructure

- Proposes a layered architecture (data ingestion, transformation, analytics, visualization) that supports interoperable APIs.
- Introduces a catalog service built on standard schemas to allow consistent service discovery.
- Uses containerization and orchestration tools (e.g., Docker, Kubernetes) to ensure interoperability at the infrastructure and deployment level.
- Leverages EuroHPC and EOSC<sup>18</sup> infrastructure to align with broader European standards for interoperability.

---

<sup>18</sup> [https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/open-science/european-open-science-cloud-eosc\\_en](https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/open-science/european-open-science-cloud-eosc_en)

## 3. Interoperability practices and FAIR principles

In the development of complex systems such as Digital Twins, data interoperability is not merely a desirable goal, but a fundamental requirement. As Digital Twins integrate heterogeneous data sources and interact with external platforms, it becomes essential to ensure that data can be exchanged, interpreted, and reused in a seamless and consistent manner. This chapter explores the key principles and best practices that support technical and semantic interoperability within a Digital Twin. The work carried out focused on the analysis of shared principles, such as those defined by the FAIR framework (Findable, Accessible, Interoperable, Reusable), as well as established standards including DCAT-AP, MELODA5, WaterML, and OpenML. These references provide a solid foundation for designing interoperable architectures and establishing sustainable and reusable data flows across multiple application domains.

Through the identification and comparison of these approaches, the chapter outlines an initial set of shared principles and operational recommendations that will drive the evolution of the interoperable and semantically consistent Digital Twin with external systems.

### 3.1 Principles analysed

#### 3.1.1 OpenML

OpenML<sup>19</sup> is an open-science platform designed to streamline and scale machine-learning research by turning datasets, algorithms, and experimental results into shareable objects. It works as a collection centre where researchers can upload their own data, models, and experiments, as well as explore and reuse the work of others. With powerful APIs, a searchable web interface and native support in popular languages such as Python, R, Java, and Julia, OpenML reduces the overhead of reproducing results and facilitates collaboration across the community. To make studies from different authors easy to compare and understand, OpenML created a well-defined structure that any experiment must follow: datasets, tasks, flows and runs.

The platform is structured around a clear and consistent object model that makes it possible to compare studies from different authors directly. At its foundation lies the dataset, which includes not only the raw data, but also extensive metadata automatically extracted during the upload. This metadata includes statistics such as the number of instances, features, class balance, presence of missing values and so on. These attributes are very useful since they enable sophisticated filtering and meta-learning without the need to manually inspect or process each dataset.

---

<sup>19</sup> <https://www.openml.org/>

Once a dataset is available, researchers can define one or more tasks that specify what kind of problem the dataset represents (like classification, regression, clustering and so on) and how the problem should be evaluated. A task defines the target variable, the evaluation metric (e.g., accuracy, RMSE, precision) and a reproducible estimation procedure like 10-fold cross-validation. By embedding these details into the task definition itself, OpenML ensures that any model evaluated on that task uses exactly the same methodology, enabling fair comparisons across different studies.

To solve a task, a researcher provides a flow, which is a versioned representation of the machine-learning algorithm or pipeline. A flow includes the full model structure, its hyperparameters, the software environment it depends on and optionally links to the source code or GitHub repository. Unlike traditional platforms where models are often treated as black boxes, OpenML makes flows transparent and reproducible. It's important to notice that the flows are executed on the user's own machine, allowing full control over compute resources and compliance with privacy or security constraints. Only the metadata, configuration and optionally the trained model are uploaded to OpenML.

The result of applying a flow to a task is called a run. This object captures all outputs of the training and evaluation process, including predictions on test folds, performance metrics, runtime statistics and details about the hardware environment. Runs are automatically linked to the corresponding flow, task and dataset, creating a traceable lineage that connects raw data to final results. This architecture transforms even long and complex training jobs into easily ready-to-use research assets. It allows others to explore existing results, reproduce them effortlessly or use them as a baseline in new experiments.

Together, datasets, tasks, flows and runs form a cohesive, version-controlled ecosystem for machine-learning experimentation. This structured approach promotes transparency, encourages reuse and significantly lowers the barrier to conducting large-scale, reproducible research. By making these objects searchable and accessible through both web and code interfaces, OpenML allows researchers to spend less time setting up infrastructure and more time pushing the boundaries of machine learning.

### 3.1.2 DCAT/DCAT-AP

In the context of data interoperability, a central role is played by the ability to describe and index information resources in a standardized way, making them easily searchable, accessible and reusable by both people and automated systems. To this end, the World Wide Web Consortium (W3C) has developed the Data Catalog Vocabulary (DCAT), a Resource Description Framework (RDF) vocabulary designed to facilitate interoperability among data catalogs published on the Web. It is based on semantic principles that allow information to be linked, query-able, and understandable by different systems.

In the context of data management and publication, one of the fundamental elements introduced by DCAT is the concept of a data catalog. A catalog, in this sense, is a structured tool for organizing, describing, and making accessible collections of datasets and, more generally, publishable information resources. Although the catalog concept could extend to other types of resources, DCAT focuses exclusively on datasets and information services to maintain a simple and interoperable standard.

Vocabulary used is not simply a metadata schema, it is a formal ontology, developed according to semantic web principles and represented in the RDF language. In the context of knowledge representation, an ontology is a structured conceptual model that describes a domain of interest, in this case, data catalogs. DCAT is based around seven main classes, including<sup>20</sup>:

- `dcat:Catalog` represents a catalog, which is a dataset in which each individual item is a metadata record describing some resource; the scope of `dcat:Catalog` is collections of metadata about datasets, data services, or other resource types;
- `dcat:Resource` represents a dataset, a data service or any other resource that may be described by a metadata record in a catalog.
- `dcat:Dataset` represents a collection of data, published or curated by a single agent or identifiable community.
- `dcat:Distribution` represents an accessible form of a dataset such as a downloadable file.
- `dcat:DataService` represents a collection of operations accessible through an interface (API) that provide access to one or more datasets or data processing functions.
- `dcat:DatasetSeries` is a dataset that represents a collection of datasets that are published separately, but share some characteristics that group them.
- `dcat:CatalogRecord` represents a metadata record in the catalog, primarily concerning the registration information, such as who added the record and when.

In the context of publishing and sharing open data, especially in public and institutional contexts, it is essential to have a common structure to describe datasets in a consistent and interoperable way. To address this need, the European Union has promoted the development of DCAT-AP, an application profile of DCAT, the W3C's standard ontology for representing data catalogs. DCAT-AP thus stems from the adaptation of an international standard (DCAT) to the specific needs of the European context, with the aim of harmonizing the description of data published in national and thematic portals, thus improving data discovery, access and reuse across borders and application domains.

---

<sup>20</sup> <https://www.w3.org/TR/vocab-dcat-3/#dcat-scope>

DCAT-AP could be employed in various scenarios within the project, especially where the federation and organization of European open data are crucial:

- **Data cataloging and discovery:** DCAT-AP would provide a standardized model to describe datasets from multiple sources, facilitating their search and access by different users and systems.
- **Integration of heterogeneous data:** Thanks to uniform metadata, DCAT-AP would help integrate data from different domains and countries, enabling a more coherent and complete Digital Twin.
- **Improvement of data quality and transparency:** Standardizing data descriptions would support monitoring the quality and completeness of available information, facilitating targeted improvement efforts.
- **Facilitation of data reuse:** With clear and consistent metadata, DCAT-AP would make it easier for various stakeholders to find and reuse reliable datasets for analysis, simulations, and decision-making.
- **Support for regulatory compliance:** Lastly, adopting DCAT-AP would help comply with European open data directives, promoting an interoperable and federated ecosystem.

### 3.1.3 WaterML

WaterML<sup>21</sup> is an OGC markup standard for the management and exchange of hydrological and environmental data. Its XML-based model removes the need for costly ad-hoc translations by giving agencies, research institutes and software tools a common syntax and a shared vocabulary.

Within the OGC catalogue two broad families of standards exist: *interface standards* that describe how systems talk and *semantic standards* that describe what the data mean. WaterML is the most recent one. By supplying a formal language for water observations, it satisfies the FAIR<sup>22</sup> principle I1 (“(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation”) and, through rich domain attributes, supports FAIR re-usability principles R1 (“(Meta)data are richly described with a plurality of accurate and relevant attributes”) and R1.3 (“(Meta)data meet domain-relevant community standards”). The same FAIR ethos is embedded in the OGC mission to make location information “findable, accessible, interoperable and reusable”.

Interoperability is provided on two complementary planes. Syntactically information is encoded as XML, ensuring that any compliant parser can read it. Semantically every variable is defined in a uniform way, so its meaning survives when records move between platforms.

---

<sup>21</sup> <https://www.ogc.org/standards/waterml/>

<sup>22</sup> <https://www.go-fair.org/fair-principles/>

The schema itself is anchored in the Observation and Measurement (O&M) conceptual model v 2.0 and Geography Markup Language 3.2, mirroring all O&M components while introducing a Collections element that lets users bundle related time series under shared metadata.

Although best known for hydrological time-series exchange, WaterML 2.0 also supplies a broader feature set that streamlines access, sharing and interpretation of complex data. Its principal capabilities are:

- Semantic self-description, for clear context and meaning accompany every value.
- Explicit time-series structure, so that each data point can carry accuracy, uncertainty, method and boundary conditions.
- Flexible exchange schema, which means that multiple identifiers and naming conventions coexist, so a single payload meets diverse organisational practices.
- Transport-agnostic delivery, so data can travel via FTP, web services or other network protocols.
- Extensibility, so that external schemas and soft typing allow new elements without breaking conformance.
- Provenance tracking, to record origin, processing history and derived “data products”, ensuring transparency and reproducibility.

The standard is published in four complementary parts that extend its reach beyond raw measurements:

1. Time Series. It supplies a conceptual model in which every observation (timestamp, location, value) is bundled with rich metadata (measurement method, accuracy, quality) so parameters such as flow, level or rainfall carry exactly the same meaning across platforms, organisations and end-users.
2. Ratings, Gauging and Sections. It captures the full lifecycle of rating curves (linking easy-to-measure stage to harder-to-measure discharge), the field measurements used to build those curves and the channel geometry that underpins them, enabling organisations to share calibration histories and quantify uncertainty for modelling, forecasting and risk analysis.
3. Surface Water Features (conceptual model). A conceptual UML model, published in 2018, which provides internationally recognised terminology (from the WMO/UNESCO glossary) and class hierarchy for rivers, lakes, canals, basins, impoundments, lagoons and estuaries. It describes the relationships between these water bodies but leaves time-varying behaviour and a concrete transfer format to future specifications, serving instead as a domain-vocabulary building block.
4. Groundwater. Conceptual and logical model (current v 2.2) for groundwater data. It supplies GML/XML schemas that encode five core components (hydrogeological units, fluid bodies, voids, fluid flow and wells) capturing properties such as porosity, permeability, recharge, discharge and monitoring-site details. Version 2.2 adopts

TimeseriesML in place of WaterML 2.0 Part 1 and can “bridge” between existing groundwater schemas, maintaining data consistency and interoperability across systems.

Through this layered architecture WaterML delivers a coherent, extensible language for water data that remains adaptable to evolving scientific and operational needs.

### 3.1.4 MELODA5

Metric for the evaluation of Open Data (MELODA) 5 (Abella et al., 2019) is a comprehensive metric designed to evaluate the reusability of data, mainly open data, particularly focusing on how datasets can be effectively reused by different users and systems. It is focused on the professional reuse of data and helps to qualify data sources based on 8 dimensions. These dimensions analyse the datasets content, the methods for accessing and the publication entity.

Every dimension defines a set of levels (from 3 to 5) in which the data source can be qualified. Every dimension is equally weighted.

#### 3.1.4.1 Aim of MELODA 5

MELODA 5 was born to allow the actual reuse of data under the open data paradigm but also when data is shared privately between organizations. Making available is not enough for a fruitful data sharing because some elements could include friction to data reuse. MELODA 5 is based on actual experiences, and it has been assessed by groups of international experts on data sharing. Finally, the metric is limited to dimensions that can be, to some extent, objective qualified and eventually automatically reviewed.

Main uses in IDEATION can include

- Rank and compare datasets by their reuse potential.
- Identify specific areas for improvement.
- Track changes in data reusability over time across portals.

#### 3.1.4.2 Dimensions and levels

MELODA 5 evaluates eight dimensions. Each dimension has discrete levels, scored from 1 up to a dimension-specific maximum, as shown in



# IDEATION

[www.ideation-project.eu](http://www.ideation-project.eu)

Table 1.

**IDEATION - D4.1: Reference architecture and Interoperability  
Guidelines (KER3) V1**

Table 1 - MELODA5's dimensions and levels

Dimension	Max	Level 1	Level 2	Level 3	Level 4	Level 5
<b>License</b>	6	1: Private use only	2: Non-commercial reuse only	3: Commercial reuse or no restrictions	—	—
<b>Access Mechanism</b>	6	1: Single dataset via web/URL	2: Single record via web interface	3: API or query language	—	—
<b>Technical Standard</b>	6	1: Closed or non-reusable open standard	2: Reusable open standard	3: Open standard with per-record metadata	—	—
<b>Data Model Standardization</b>	10	1: Proprietary model	2: Published ad hoc harmonization	3: Local-level standardization	4: Global standardization	—
<b>Geolocation</b>	6	1: No geographical information	2: Simple or complex text field	3: Complete coordinates or geospatial data	—	—
<b>Update Frequency</b>	15	1: >1 month intervals	2: Between 1 day and 1 month	3: Between 1 hour and 1 day	4: Between 1 min and 1 hour	5: <1 min intervals
<b>Diffusion</b>	6	1: No systematic communication	2: Available update channels	3: Proactive/push	—	—



			(e.g., social feeds)	dissemination		
<b>Reputation</b>	6	1: No information on data source reputation	2: User-feedback-based stats/reports	3: Rankings or indicators based on data source reputation	—	—

The dimension Licensing belongs to the dataset same as geolocation, update frequency, data model standardization and technical standard. The dimension access mechanism belongs to the tool to make the datasets available and the dimensions diffusion and reputation belongs to the entity which is publishing the data.

### 3.1.4.3 Scoring Methodology

1. **Assign Levels:** For each dimension, assign the level that matches the dataset's characteristics. Levels have point values equal to their numeric label (e.g., Level 1 = 1 point, ... Level 5 = 5 points).
2. **Sum Points:** Calculate the **Total Score** by summing points across all eight dimensions.
3. **Interpretation Range:**
  - **Inadequate:** 8–23 points
  - **Basic:** 24–47 points
  - **Advanced:** 48–61 points

### 3.1.4.4 Potential uses of MELODA 5 in IDEATION

MELODA 5 can be employed in a variety of scenarios where a thorough evaluation of dataset quality is crucial:

- **Dataset Comparison and Selection:** When choosing between multiple datasets for a specific task, MELODA 5 provides a standardized way to compare their overall reusability and identify the most suitable option.
- **Data Quality Monitoring:** Over time, the reusability of a dataset can degrade due to various factors (e.g., entity reputation, access methods, licensing, etc). Regularly calculating the MELODA 5 score can help monitor these changes and trigger alerts when quality falls below acceptable thresholds.

- **Identifying Areas for Improvement:** The individual dimension scores within MELODA 5 highlight specific aspects of the dataset that need attention. This allows data engineers and scientists to focus their efforts on targeted data cleaning and enhancement strategies.
- **Communicating Data Reusability:** A single, aggregated MELODA 5 score can effectively communicate the overall reusability of a dataset to stakeholders who may not have the technical expertise to interpret multiple individual metrics.
- **Benchmarking Datasets:** Within an organization or across different entities, MELODA 5 can serve as a benchmark for data reusability, allowing for comparisons and the identification of best practices.

### 3.1.5 FAIR principles

The FAIR Principles<sup>23,24</sup> are a foundational framework designed to improve the stewardship and usability of digital data assets for both humans and computational systems. Developed in response to the challenges posed by the increasing volume and complexity of digital data.

Acronym FAIR stands for:

- **Findable (F):** This principle emphasizes that data and its associated metadata should be easy to locate for both human users and computational systems. Key aspects include assigning globally unique and eternally persistent identifiers (PIDs) to data and metadata (F1), describing data with rich metadata (F2), ensuring metadata clearly includes the identifier of the data it describes (F3), and registering or indexing (meta)data in a searchable resource (F4). Effective discovery relies heavily on machine-readable metadata.
- **Accessible (A):** Once data is found, this principle dictates that users must know how the data can be retrieved, including any necessary authentication and authorization procedures. This involves making (meta)data retrievable by their identifier using a standardized communications protocol (A1), ensuring the protocol is open, free, and universally implementable (A1.1), and allowing for authentication and authorization where necessary (A1.2). A crucial point is that FAIR data does not require all data to be openly accessible, but the access mechanism should be clear and standard. Metadata should also remain accessible even when the data are no longer available (A2).
- **Interoperable (I):** This principle addresses the need for data to be readily integrated with other datasets and to work effectively with various applications or analytical workflows. It involves using a formal, accessible, shared, and broadly applicable

---

<sup>23</sup> <https://www.go-fair.org/> and [https://commission.europa.eu/sites/default/files/turning\\_fair\\_into\\_reality\\_1.pdf](https://commission.europa.eu/sites/default/files/turning_fair_into_reality_1.pdf)

<sup>24</sup> <https://www.nature.com/articles/sdata201618>



language for knowledge representation (I1), such as standard data formats and metadata schemas. It also requires using vocabularies that follow FAIR principles (I2), employing controlled vocabularies, ontologies, and thesauri to describe data elements consistently. Furthermore, (meta)data should include qualified references to other (meta)data (I3), creating explicit, machine-readable links between datasets and related resources

- **Reusable (R):** The ultimate aim of the FAIR principles is to maximize the potential for data to be reused in future research, for different analytical purposes, or to inform new applications. This requires that (meta)data are richly described with a plurality of accurate and relevant attributes (R1), including details about data quality, processing history, and methodologies. Data and metadata should be released with a clear and accessible data usage license (R1.1) and be associated with detailed provenance information (R1.2). Additionally, (meta)data should meet domain-relevant community standards (R1.3).

**Error! Reference source not found.** details each of the FAIR sub-principles and their core requirements.

Table 2 - FAIR's sub-principles and their core requirements

Principle	ID	Core Requirement (Description)
<b>Findable</b>	F1	(Meta)data are assigned a globally unique and eternally persistent identifier.
	F2	Data are described with rich metadata.
	F3	Metadata clearly and explicitly include the identifier of the data they describe.
	F4	(Meta)data are registered or indexed in a searchable resource.
<b>Accessible</b>	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.



<b>Interoperable</b>	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.
<b>Reusable</b>	R1	(Meta)data are richly described with 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 detailed provenance.
	R1.3	(Meta)data meet domain-relevant community standards.

### 3.1.5.1 Aim of the FAIR principles

The core objective of the FAIR principles is to "optimize the reuse of data". This is achieved by making research data and other digital objects more discoverable and usable, not only by human researchers but, critically, by computational systems. A key emphasis is placed on "machine-actionability," which is the capacity of computational systems to find, access, interoperate, and reuse data with minimal or no human intervention.

### 3.1.5.2 Potential uses of FAIR principles in IDEATION

Once agreed, initially, on the shared semantic, a potential implementation of these principles would take these steps:

- **Assign Persistent Identifiers (PIDs):** Assign globally unique and persistent identifiers to datasets and other critical entities (entities defined in the semantics shared).
- **Create Rich and Standardized Metadata:** Develop comprehensive metadata using standardized schemas relevant to the data type (e.g., ISO 19115 for geospatial data, Dublin Core for general resources, or DCAT-AP as a general rule). For water data, this includes detailing location, time, parameters, units, and methods.
- **Define Data Access Protocols:** Clearly state the mechanisms for accessing the data, utilizing standard, open protocols like NGSI-LD or other Digital Twins services. Ensure metadata is accessible even if data access is restricted.
- **Utilize Interoperable Data Formats and Vocabularies:** Provide data in open, widely used, and machine-readable formats (e.g., CSV, JSON, JSON-LD, etc.). Employ

shared, FAIR vocabularies for consistent description of parameters, units, and methods. If vocabularies are not available new data models at Smart Data Models can be defined.

- **Implement Clear Data Usage Licenses:** Assign clear and preferably standardized licenses (e.g., Creative Commons, Open Data Commons) to define reuse terms.
- **Ensure Comprehensive Provenance Documentation:** Document the data's origin, collection methods, processing steps, and quality control. If the object is compliant with SDM some of these needs are created automatically.
- **Use a FAIR compliant:** Publish the FAIRified data, metadata, and license in the chosen repository or platform to ensure discoverability and access.

### 3.2 *Shared semantics*

Shared semantics form the basis for ensuring that data exchanged between different systems, organizations and applications are interpreted correctly and consistently. Without a common language or shared vocabularies, data risks losing meaning, generating ambiguity, or being unusable in complex integration contexts such as those of a Digital Twin. A set of initial recommendations regarding semantics for the future development of the digital twin have been obtained. This initial set of recommendations, obtained through an internal discussion with experts in the consortium and an analysis of the tools analyzed, aims to define an initial common framework that facilitates terminological and semantic alignment among the different actors involved, thus enabling effective cross-border collaboration and interoperability.

This is an initial set of recommendations for the future development of the Digital Twin:

- 1. Adopt a Common Conceptual Model:** Establish a shared understanding of the key entities, concepts, and relationships of the Digital Twins involved and the applications and systems that are going to use the data. It could be based on existing semantics, however it is unlikely that this will cover all the needs and an agile approach for extending the semantics should be used.
- 2. Utilize Standardized Vocabularies and Ontologies:** A comprehensive initial list of vocabularies and ontologies should be defined. They should be, initially, capable of allocating the concepts defined in the previous point.
- 3. Ensure Semantic Interoperability:** it is needed to go beyond data exchange formats (syntactic interoperability) to ensure that the meaning of the data is preserved and understood consistently across all connected Digital Twins and applications. Unique URI could help in this direction.
- 4. Combine FAIR Principles and MELODA 5 reusability practices:** Assess the data sharing platforms (Digital Twins, applications) against these principles and try to set up a plan for improving those elements with a poor performance according to these frameworks.

5. **Establish a Data Governance and Stewardship:** Define clear roles and responsibilities for managing and maintaining the shared semantic models and the data it describes. This includes processes for updating vocabularies and ontologies, ensuring data quality, and managing access rights.
6. **Document Semantics Thoroughly:** Provide comprehensive documentation for the shared conceptual model, vocabularies, and ontologies. If using SDM most of this job is done automatically and it could save resources of the project. It could include the mapping of existing vocabularies and ontologies to profit from this automation.
7. **Include Extensibility and Evolution:** Anticipate that the semantic model will need to evolve as IDEATION's Digital Twin grows across project execution. SDM already has this built-in characteristic, but other semantic sources could have some limitations.
8. **Foster Collaboration and Stakeholder Engagement:** Developing a shared semantic understanding requires active collaboration among all stakeholders, so a specific group must be created across members of the project for addressing this collaboration.
9. **Care about Data Quality and Provenance:** Whenever possible, and based on the available resources, care (set up some KPI about quality and provenance) about data quality and provenance

### 3.3 *Shared principles for Digital Twin*

A structured and collaborative reasoning process was conducted, combining comparative analysis of existing standards, and in-depth discussions among the project partners. This joint effort led to the definition of an initial set of shared principles to guide the development of a robust and interoperable Digital Twin architecture.

These principles synthesize best practices and concrete operational requirements, addressing both technical priorities such as modular architecture, real-time data integration, and data quality assurance, and broader values such as openness, usability, and collaborative development. The goal is to provide a practical and actionable framework that supports the implementation of FAIR principles and promotes long-term sustainability, scalability, and interoperability across Digital Twin domains.

The following is a summary of the key principles identified through this process. Each principle is accompanied by a brief explanation and rationale highlighting its relevance to IDEATION.

#### 3.3.1 Semantic principles

##### 3.3.1.1 Shared data

This principle advocates for the use of open licensed data models to facilitate data sharing among different Digital Twins and with external sources. The adoption of open standards

allows for potential synergies and simplifies the connection and exchange of information between various data sources and Digital Twins.

**Relevance for IDEATION:** For Digital Twins of inland waters –which might involve diverse datasets related to waterways, infrastructure, environmental conditions, and logistics– shared data models are crucial for integrating information from different sources (e.g., sensor data, historical records, weather forecasts). This interoperability is essential for a holistic view and effective management of inland systems.

### 3.3.1.2 Data Quality Assurance

This principle emphasizes the importance of ensuring accurate and reliable information feeding into Digital Twins. It requires implementing processes such as data cleaning, validation, and normalization to guarantee the use of high-quality data.

**Relevance for IDEATION:** The accuracy of Digital Twins of inland waters heavily relies on the quality of collected data. Poor data quality can lead to inaccurate simulations, especially relevant with highly non-linear systems, analyses, and predictions, impacting decision-making related to navigation, water management, and infrastructure maintenance.

## 3.3.2 Access to Data Principles

### 3.3.2.1 Geolocated data

This principle states that whenever possible, all data managed by a Digital Twin should be geolocated. This means that data describing different assets should include coordinates and hopefully these coordinates in compatible formats.

**Relevance for IDEATION:** Systems covering inland waters are inherently spatial. Geolocated data is fundamental for representing the physical location of assets like rivers, canals, locks, bridges, and vessels within a Digital Twin. This allows for spatial analysis, visualization on maps, and integration with geographical information systems (GIS), which is critical for navigation, spatial planning, and emergency response.

### 3.3.2.2 Open API

This principle suggests that the API for accessing and retrieving information from the Digital Twin should ideally have an open and well-documented specification. Having an open API allows for automated information retrieval and enables others to adopt the API or integrate other tools with the Digital Twin. If possible, the API should be extensively tested in real and diverse scenarios and if possible open-source implementation should be available.

**Relevance for IDEATION:** An open API promotes interoperability and allows various applications and services to access and utilize the Digital Twin's data and functionalities. It

also can save resources when open-source implementations are available if they can be customized and with affordable integrations

### 3.3.2.3 Real-time integration

Digital twins should be designed to integrate and visualize data in real time to ensure timely decisions and continuous updates. This involves implementing data streaming technologies to capture and update information from sensors and IoT devices in real time, and using real-time monitoring systems to detect critical events.

**Relevance for IDEATION:** Real-time data integration is crucial for some aspects of the dynamic inland systems. It allows for monitoring current conditions like water levels and weather, enabling timely responses to events such as floods, accidents, or changes in navigation conditions.

## 3.3.3 Organizational Principles

### 3.3.3.1 Security and Privacy

This principle highlights the critical need to ensure that data processed by Digital Twins are protected from unauthorized access and comply with privacy regulations such as GDPR, to avoid the risk of data breaches.

**Relevance for IDEATION:** Digital twins of inland waters could eventually handle sensitive data, including infrastructure details, actuation on infrastructures, operational information, and potentially personal data related to users or personnel. Robust security measures are paramount to protect this information from cyber threats and ensure compliance with data protection laws, maintaining trust and system integrity.

### 3.3.3.2 Modular and Scalable Architecture

This principle emphasizes that to ensure flexibility and adaptability, Digital Twins must be designed in a modular way, using a microservice architecture.

**Relevance for IDEATION:** Digital twins of inland waters can vary significantly in scope and complexity. A modular and scalable architecture allows for building systems that can adapt to different needs, from small-scale river sections to entire waterway networks. This design enables easier updates, maintenance, and the integration of new functionalities as requirements evolve.

## 3.3.4 FAIR Principles

### 3.3.4.1 FAIR (accessible). User Interfaces

User interfaces, such as GUIs, provide user-friendly access points for data exploration, simulation management, and workflow monitoring.

**Relevance for IDEATION:** Accessible user interfaces are essential for enabling various stakeholders, including waterway managers, navigators, and researchers, to interact with the Digital Twin. Intuitive interfaces facilitate data visualization, simulation analysis, and operational control, making the Digital Twin a practical tool for decision-making and management.

### 3.3.4.2 FAIR (interoperable). Multi-format Support

The system should support widely used scientific formats ensuring data compatibility across diverse applications. Eventually it should include conversion systems. Supporting such interoperable formats facilitates the exchange of information between different tools and platforms, allowing users to work with their preferred tools while maintaining consistency and accessibility in their data management processes.

**Relevance for IDEATION:** Waterway data related to inland waters will integrate data from different systems that come in various formats from different sources. Supporting multiple standard formats ensures that data from sensors, models, and external databases can be easily integrated and used within the Digital Twin, enhancing its ability to interact with the broader data ecosystem.

### 3.3.4.3 FAIR (interoperable). Workflow manager

Workflow managers can facilitate cross-platform interoperability by integrating HPC resources with cloud systems. This capability enables efficient coordination and execution of complex computational tasks across different environments, maximizing resource utilization and scalability.

**Relevance for IDEATION:** Digital twins of inland waters often require complex simulations and data processing that may utilize various computing resources. A workflow manager can orchestrate these tasks across different platforms (e.g., local servers, cloud computing, high-performance computing), ensuring efficient and scalable operation of the Digital Twin.

### 3.3.4.4 FAIR (reusable). Collaborative Development

Version-controlled development environments, such as GitLab CI/CD pipelines, allow users to share and update models, simulations, and workflows. Modular platforms (see previous principle) can provide reusable components for generating simulations.



**Relevance for IDEATION:** Collaborative development environments support teams of developers, researchers, and domain experts working together on Digital Twin projects on inland waters. Especially when using open-source components, version control and reusable components facilitate efficient development, model sharing, and the creation of complex simulations relevant to inland waterway dynamics.

#### **3.3.4.5 FAIR (reusable). Documentation and Tutorials**

Comprehensive tutorials and GUIs help reusability by ensuring that datasets, processes, and tools can be easily reused by new users with minimal additional effort.

**Relevance for IDEATION:** Good documentation and tutorials are crucial for making Digital Twin of inland waters components and datasets accessible and understandable to a wider audience not only within the project but for external users, even after IDEATION's end. This promotes the reuse of existing resources, reduces the learning curve for new users, and fosters the growth of the inland Digital Twin community.

## 4. Reference Architecture

### 4.1 *Main components*

In the introductory chapter (Chapter 1), the main components common to various Digital Twins were analyzed and described. This understanding was derived from a thorough review of the literature and discussions within the IDEATION consortium, which fostered continuous exchange of opinions, experiences, and knowledge among its members. However, in order to effectively implement all these components, their functionalities, and the resulting requirements, it is equally essential to carefully consider the type of architecture to adopt. Specifically, to adequately address the demands of a Digital Twin, an architecture must be capable of evolving and adapting to diverse application scenarios, managing variable data volumes, and supporting future modifications without disrupting the underlying core structure. Only by designing an architecture that ensures flexibility, scalability, and reliability can the fundamental characteristics required by such a system be met.

Therefore, it is not feasible to employ a traditional monolithic architectural model. Although monolithic systems may appear simpler during initial design phases, they tend to become rigid and inefficient in dynamic and heterogeneous environments typical of Digital Twins. In these contexts, where rapid adaptability and independent evolution of components are critical, adopting a more modular and flexible architecture is imperative.

To overcome these limitations, IDEATION adopts a microservice-based architecture. This approach decomposes the system into loosely coupled, independently deployable services, each responsible for a specific function, such as data ingestion, processing, orchestration, or visualization. By decoupling components, it becomes possible to upgrade or scale parts of the system individually, enable parallel development, and integrate new features through continuous delivery pipelines. The result is an infrastructure that is both resilient and adaptive: capable of reacting to changing data streams, incorporating new models, and supporting evolving user needs.

Building on this foundation, the following section provides a detailed breakdown of the core components that make up the IDEATION Reference Architecture. Each layer has been designed not only to perform its technical role efficiently, but also to align with the interoperability and modularity principles that underpin the entire system.

The architecture is structured into the following key layers:

- Data layer: this layer is responsible for integrating data into the architecture;
- DT Core: is the heart of the architecture and handles containerization and execution of all operations such as simulation, analysis, and other computation;



- Interoperability Hub: this layer facilitates interaction between internal components and external platforms by sharing the full set of services and processes made available;
- DT App: this layer includes the application services and user interfaces that enable visualization, control, and management of Digital Twin functionality.
- Infrastructure: the underlying infrastructure layer provides the necessary computational, storage and network resources. It ensures the scalability, reliability of the entire architecture in the cloud.

Together, these components form a robust and future-ready architecture capable of integrating with external platforms such as the DTO, and supporting a wide variety of real-world scenarios.

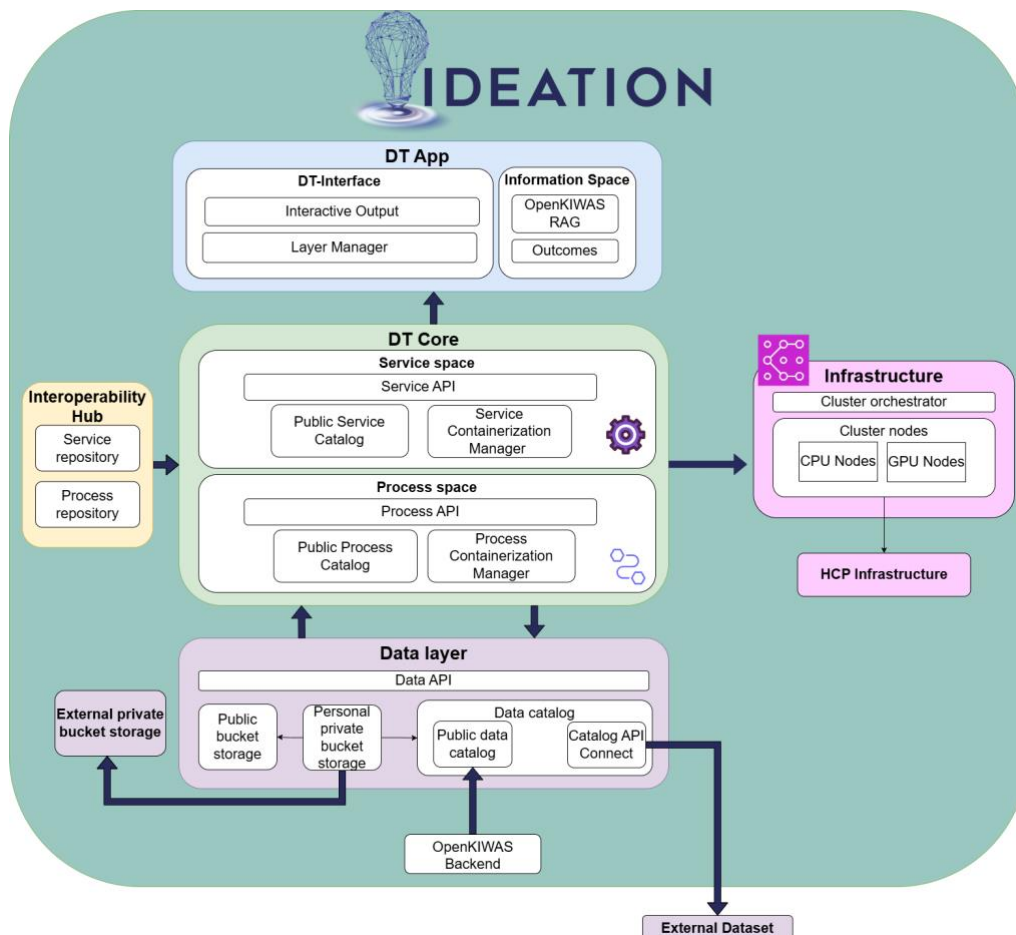


Figure 2 - - First version of IDEATION's Reference Architecture

## IDEATION - D4.1: Reference architecture and Interoperability Guidelines (KER3) V1

## 4.2 *First version of IDEATION's Reference Architecture*

### 4.2.1 Introduction

One of the main goals of IDEATION is to define a solid, flexible, and reusable reference architecture for the development of a Digital Twin of inland waters. From the earliest stages of the project, several versions of the architecture have been created, each one an evolutionary step forward from the previous, through an iterative process that has progressively refined the conceptual framework, leading to the current solution (see Figure 2) that will be described in detail in the following subsections.

The architecture has been deliberately designed at a logical level, aiming to provide an abstract yet clear view of the system, capable of supporting a modular, structured, and scalable approach to the inherent complexity of a Digital Twin.

The proposed architecture in this first version is organized into multiple layers and subdivided into functional blocks that represent microservices, independent components, each with a specific purpose. Every block or layer interacts with the others by exchanging data or providing services, forming a flexible, distributed, and highly modular system. In the architectural diagram, every arrow does not merely indicate a one-way data flow; instead, it represents a dynamic and bidirectional exchange of information, underscoring the interactive and collaborative nature of the system. This ensures that each block maintains its defined function while seamlessly interacting with others.

Adopting a layered architecture brings several important advantages to a Digital Twin. This approach ensures that each level has a specific role and set of responsibilities. By clearly separating concerns, redundancy and confusion in the way components interact is avoided, and as a result, the system becomes easier to design and scale. In addition, isolating functionality into distinct layers simplifies troubleshooting and maintenance. Furthermore, this structured design greatly facilitates, especially in the second version of the architecture, the future addition, modification, or removal of components without compromising the overall system stability or coherence.

Secondly, a layered architectural model strongly supports scalability, which is a crucial factor for a Digital Twin designed to evolve over time. Each layer can be independently updated, improved, or extended, allowing the system to flexibly adapt to increasing data volumes, higher computational demands, or greater analytical complexity.

Another key factor that enhances the scalability of the proposed architecture is its strong cloud-based design. From the outset, the solution has been conceived with cloud infrastructure in mind, allowing for more flexible resource management and computational

scalability. This choice is also in line with one of the core goals of the project, namely to ensure the highest possible level of interoperability.

In the context of a complex and distributed Digital Twin, interoperability is a fundamental requirement to enable seamless integration with external data sources and third-party systems. The following subsections will explore in detail how this is achieved, highlighting the essential role of connectors to external datasets, the necessity of maintaining a catalog to expose and publish data, and the implementation of a shared repository where users can access and reuse the services provided. These components are critical in making the platform not only more scalable, but also more open, collaborative, and capable of evolving over time.

By following a microservices-based approach, the resulting architecture will provide a solid and flexible foundation for building a robust and versatile Digital Twin. It will be equipped with advanced tools that can be easily adapted and reused across various use cases, ensuring efficiency, scalability, and interoperability.

## 4.2.2 Data layer

The *Data Layer* (see Figure 3) represents the foundation of the entire architecture, both conceptually and infrastructurally. Its function is to govern the acquisition, storage, standardisation and publication of data in an integrated manner, guaranteeing accessibility, scalability and above all interoperability with other environments, such as the DTO.

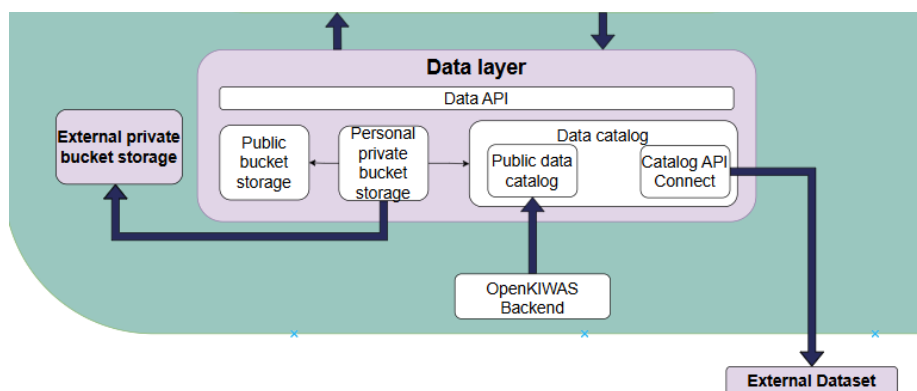


Figure 3 - Data layer section into the reference architecture

This level can be logically subdivided into two main sections: one dedicated to physical data storage, the other to management and semantic federation via the catalogue.

The storage will be structured predictably in two spaces: a private one, associated with each user, and a public one, accessible to the whole community. The private space will be

automatically allocated during profile registration, with capacity limits to ensure sustainable use of resources. Within it, users will be able to upload and organise their own data, analyses, model outputs or any other type of content useful for the life cycle of the Digital Twin. Storage management can be entrusted to object storage technologies, such as MinIO<sup>25</sup>, which offers APIs compatible with the S3 standard and allows heterogeneous data to be managed without constraints related to their structure or format. The use of this technology, or similar solutions, would make it possible to easily define private, public or shared spaces, simply by configuring dedicated buckets and access rules. Next to the private space, the architecture provides public storage, within which data published by users or uploaded by the platform are made available. These data are visible to all and can be federated in the semantic catalogue, thus enabling integration in the broader context of the Digital Twin. The distinction between private and public space is strategic, it allows users to experiment and test in isolation, before publishing content of general interest that contributes to the common knowledge base. An additional strength lies in the flexibility of the system, private space need not correspond to storage managed internally by IDEATION. In fact, an existing external bucket belonging to a third-party infrastructure or already in use by the user can be configured as personal storage, and in this way, unnecessary duplication of data is avoided, transfer times are reduced, and a federated and collaborative approach is encouraged. A research team, for example, can share a bucket with relevant data and make it accessible to developers or other Digital Twins without replicating or physically moving resources.

Parallel to storage, as anticipated, there is a fundamental section dedicated to federated data management (*Data catalog*), which has the task of organizing, describing and making datasets accessible within the IDEATION architecture, but also outward, with a view to federation and reuse of data among distributed Digital Twins. Within this functional block it is possible to distinguish two main components: on the one hand, the public and shared catalog, which will collect the resources explicitly published by users or made available by the platform, and on the other hand, a set of APIs that will allow the federation of data already present on the web or in other instances, without the need for import or physical duplication. The public catalog will then serve as a reference point for all internal components of the system, enabling the indexing and searching of datasets through standardized metadata and the exposure of related services through open protocols (such as OGC CSW<sup>26</sup>, WMS<sup>27</sup>, WFS<sup>28</sup>). This will ensure that any data published in public or federated storage from external sources can be natively integrated and exploited by the system, maintaining full semantic and operational interoperability. The federation APIs will be

---

<sup>25</sup> <https://min.io/>

<sup>26</sup> <https://www.ogc.org/standards/cat/>

<sup>27</sup> <https://www.ogc.org/it/standards/wms/>

<sup>28</sup> <https://www.ogc.org/it/standards/wfs/>



responsible for intercepting external resources, such as datasets exposed by other Digital Twins or public repositories and including them in the IDEATION catalog in a federated form. This process, which can also be done dynamically, will allow the platform to access and use data from external environments, without replicating them locally, while maintaining their integrity, provenance and traceability.

Positioned above the two primary subsystems of the *Data Layer*, the Data API serves as a horizontal, cross-cutting component essential to the overall architecture. This layer is responsible for exposing a consistent and standardized set of APIs through which the entire system, as well as any external components, can interact uniformly with data and related resources. The Data APIs are not limited to simply extracting or writing files, but enable complete operational management of the entire data ecosystem, allowing, for example, linking and mounting external storage buckets to the IDEATION architecture, or interactions with the catalog. Finally, through these interfaces it will be possible to directly access content stored in private or public storage, thus representing the connecting element between the raw data and the other layers of the architecture, facilitating interoperability within the Digital Twin.

Another significant result of the IDEATION project is OpenKIWAS, whose first version is thoroughly described in Deliverable *D3.1 - OpenKIWAS*. To effectively leverage and utilize the information gathered through OpenKIWAS within the architecture of the Digital Twin, it is essential that this data are first included in the project's data catalog. This step ensures that the information becomes accessible for provisioning purposes and supports its application across various use cases (as the Reference Use Cases under definition and that will be detailed in D2.5 "Co-created use cases description and visual representation" to be submitted at M18, November 2025). The work conducted within WP3 has led to the identification of a wide range of valuable data types, including: a catalog of EU and international policies and standards; a catalog of project results and information systems related to inland waters; a catalog of scientific research outputs; a catalog of data models, AI technologies, simulations, intelligent systems, and Digital Twins associated with inland waters; a catalog of open data and sensing sources; and a catalog of data and services from national meteorological organizations.

Data collected in the OpenKIWAS can be directly imported into the system's catalog. Once integrated, this data will become immediately available for retrieval and use across various architectural components. Adding this kind of information provides an important and reliable starting point to support the Digital Twin, enriched with technical metadata such as the indication of API protocols, where available, or URLs pointing to relevant specifications and implementation guidelines.

Within the architecture, this integration is represented by the block called 'OpenKIWAS Backend', which is directly connected to the data catalog. This connection enables the

federation of selected information within the system, guaranteeing a continuous and structured flow of data to the components that make use of it.

### 4.2.3 DT Core

This layer (see Figure 4) represents the beating heart of the Digital Twin, as this is where the computational and logical operations that give meaning and value to the collected data take place, making it usable through visualizations, analysis or transformations made available to end users. At its core, the layer is divided into two main macro-components: the service block (*Service space*) and the process block (*Process space*).

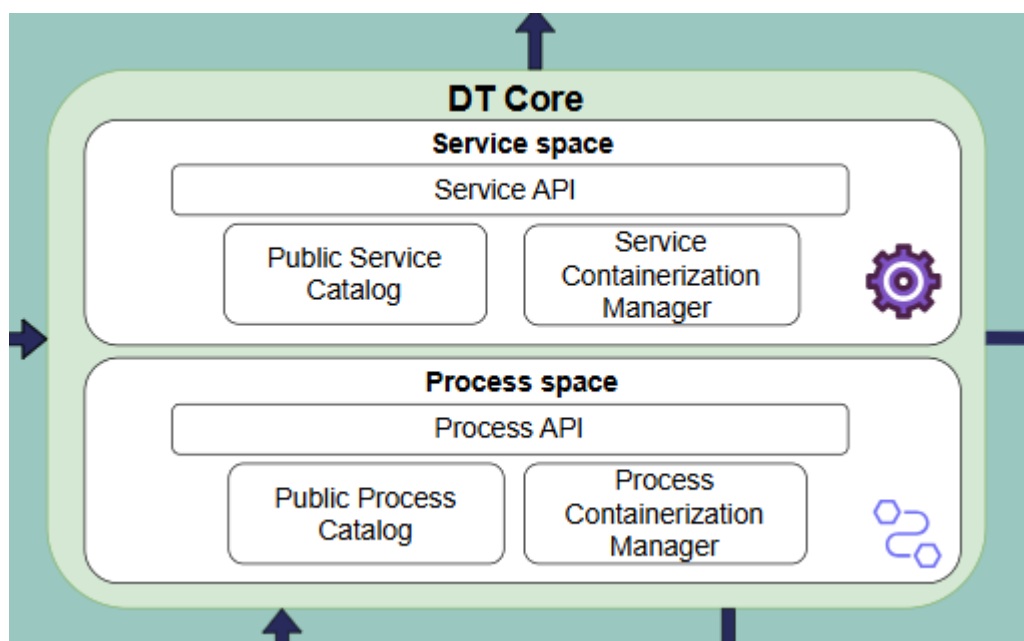


Figure 4 - DT Core section into the reference architecture

The *Public Service Catalog* refers to sets of functionalities that, once configured, can be activated as real applications. These applications, with or without a graphical interface, may include data science tools, analytic modules for specific users or microservices useful for data transformation and management. Nevertheless, since in the IDEATION context, at the moment, there is no provision for the direct and interactive development of these applications within the Digital Twin itself, the system adopts a dynamic configuration-based deployment logic. Each service is then described through a configuration file, which defines the functional characteristics and resources required to instantiate the container in which it will run; the file does not contain the application code itself, but describes its expected

behavior. Once defined, this file is read by the *Service Containerization Manager*, which is responsible for instantiating the service on an available infrastructure node, then making it operational and accessible to the user. To ensure reuse, interoperability, and transparency, application configuration is recorded within the service catalog, an area of the system dedicated to collecting officially validated and ready-to-use applications. In parallel, however, the entire management of configuration files, from their initial creation to publication, is handled in a separate environment called the *Interoperability Hub*. This component, designed as a collaborative repository, like GitHub, allows developers to upload new configurations, modify them, submit them for review, and finally propose them for integration into the official catalog. This clear separation of catalog and hub is crucial to maintaining a high level of control and quality; while the catalog collects only tested, validated, and containerization-ready services, the hub represents a shared workspace for experimentation and co-creation. Access to the final publication stage is then filtered by a controlled approval process, avoiding overloading the system with non-compliant or redundant proposals.

The same mechanism described for services also applies to process management in the *Process space*, with the substantial difference that in this case the deployed component does not represent an application in the strict sense, but a computational function, oriented toward data processing or transformation.

This distinction, also used in DTO, defines process and service as distinct but complementary concepts related to the execution and exposure of capabilities within the Digital Twin.

A process is a remote function that generates data, such as data transformation, pre/post-processing, reanalysis, forecasts, surveys, What-If scenarios, and quality checks. It can be piped, programmed or triggered on demand. Inputs and outputs can be configured at runtime. Unlike a service, it is not interactive during execution (it does not host a web server or user interface)<sup>29</sup>.

Also in this case, the catalog contains the configuration files, and only when they are launched does the *Process Containerization Manager* make the processes executable within containers. The processes available in the catalog, like the previously mentioned services, have passed all the required steps for publication in the *Process repository*.

The results obtained in this layer, whether images or results of analyses and predictions, will then be stored in the *Data Layer*, which is why you can see a double arrow in the diagram.

Above the processes and services, the architecture provides for the presence of two distinct layers of APIs, each dedicated to the management of and interaction with their respective components, which not only facilitate the connection between the internal modules of the system, going to invoke services and processes internally via APIs, but also a way to be able to export outside the architecture the use of these components. This layer of APIs thus acts

---

<sup>29</sup> [https://pub.pages.mercator-ocean.fr/edito-infra/edito-tutorials-content/#/\\_glossary?id=edito-process](https://pub.pages.mercator-ocean.fr/edito-infra/edito-tutorials-content/#/_glossary?id=edito-process)

as an abstraction layer that unifies the behavior of the different underlying elements, also allowing easier integration with external systems or federated environments such as other Digital Twins, where communication between architectures must take place according to common standards and shared interfaces.

#### 4.2.4 Interoperability hub

The *Interoperability Hub* (see Figure 5), as anticipated in the previous section, plays an important role within the IDEATION Reference Architecture, serving as a centralized and structured environment for the management, creation and validation of configuration files required to deploy services and processes in the Digital Twin.

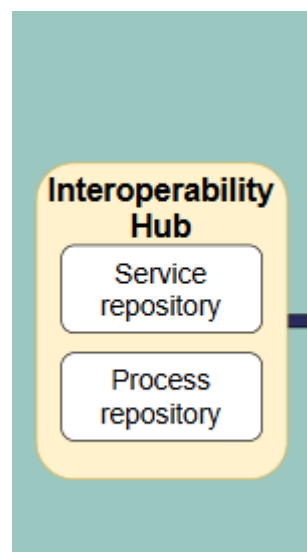


Figure 5 - Interoperability Hub section into the reference architecture

Within the Digital Twin, processes and services can be used directly by users in the *DT Core*, either by writing custom code and processing pipelines, performing analysis and computation on available data, or accessing applications provided as ready-to-use services. However, when a new service or process needs to be added and made available to the entire community, it must go through the *Interoperability Hub*. This layer provides a set of structured repositories where Digital Twin developers can collaborate on the development of new components, following a standardized workflow. This establishes a clear and controlled process for integrating new functionalities, ensuring community involvement, the possibility for review and quality control, and consistent governance before a service becomes publicly available.

The operation of the *Interoperability Hub* is inspired by more established collaborative versioning systems, such as those also adopted in the Digital Twin Ocean project.

Developers can work in an existing repository, modify its contents or begin development of a new service or process, following a logic based on branch, commit and merge request. Once a developer has completed their contribution, they can submit a merge request, which will be reviewed by a technical team or automated quality control systems. In this way, only those services and processes that pass this stage will be integrated and made available in the official catalog of distributable components.

In addition to managing configurations, APIs exposed by the various layers of the architecture, such as *Data API*, *Service API*, and *Process API*, fall into the *Interoperability Hub*. These APIs make components, features, and data accessible not only internally within the IDEATION ecosystem, but also externally, enabling integration with external platforms and applications.

Interoperability within the IDEATION architecture is not limited to data sharing and accessibility, but also extends in a fundamental way to services and processes. This means that not only can information be exchanged between systems, but also the computational components themselves, such as algorithms, analytical modules, pipelines and microservices, can be instantiated or run in different environments, both inside and outside the IDEATION platform. Thanks to this expanded view of interoperability, externally developed services and processes can also be integrated; any component, as long as it meets the integration specifications and follows the process defined within the *Interoperability Hub*, can be registered, versioned, and made available in the official catalog. This approach allows for continuous and open growth of the ecosystem, enabling adoption of solutions developed by other projects, institutions, or communities, while ensuring consistency, quality control, and integrability with other elements of the architecture.

This approach makes it possible to build an open but selective ecosystem, where it is possible to benefit from the collaboration of multiple developers and distributed actors, without compromising the consistency and reliability of the services offered within the Digital Twin. In addition, the possibility of versioning configurations makes it possible to maintain an evolutionary history of each component, facilitating reuse, debugging and documentation.

#### 4.2.5 DT app

The final layer of the architecture is dedicated to data visualization and exploration (see Figure 6), and it plays a crucial role in ensuring the accessibility and usability of the results generated by the system's internal services and processes. This layer is designed to present processed information clearly, intuitively, and effectively through an advanced, user-friendly graphical interface.

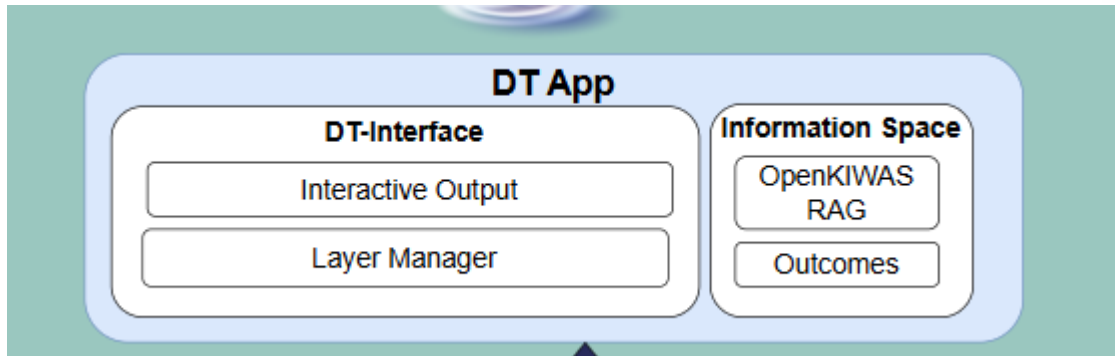


Figure 6 - DT App section into the reference architecture

To ensure that the results produced by the analyses and operations performed in the lower layers are truly usable and understandable, it is essential that this information can be represented and visualized in a clear and accessible manner. The visualisation layer was designed precisely for this purpose; within it, a dedicated block called *DT-interface* is responsible for hosting the most appropriate representations for different types of data, such as map views, globe views with various styles or other forms of interactive output. Additionally, a *Layer Manager* is included, tasked with dynamically handling more complex visualizations, including the overlay of multiple information layers and the selection of the most effective rendering methods depending on the operational context and user preferences. Specifically, the *Layer Manager* will be capable of handling and visualizing both the results generated and stored by the *DT-core* in the *Data layer*, and the data available in the *data catalog*. This dual capability allows users to explore, in an integrated manner, both the outputs from internal Digital Twin processes and federated external datasets accessible through the catalog. As such, the *Layer Manager* plays a central role in making large volumes of heterogeneous data accessible and explorable, ensuring consistent representation and offering tools to manage overlays, visualization styles, and customizable levels of detail according to operational needs.

Beside the *DT-Interface*, which allows the visualisation of data and results in an intuitive way, there is a need for more detailed information or specific results from complex activities or analytical processes, this need is fulfilled by the *Information Space*, which serves as a container for *Outcomes*. *Outcomes* represent the results of activities, services or processes performed within the system that may be the result of simulations, forecasts, or even continuous monitoring operations. These outcomes are fundamental for evaluating the performance and effectiveness of a given service over time and for making comparisons between different configurations or scenarios. For example, they may be useful to monitor the change of a parameter over time, to analyse the effect of different variables on an outcome or to compare the effectiveness of different policies or strategies. Depending on the use case, it might then be possible to activate a number of operations,

such as monitoring how these change over time in response to specific changes, or for example adding new inputs or changing existing parameters or configuring notifications to alert the user when certain values exceed predefined thresholds.

Finally, to further increase the added value of OpenKIWAS, a dedicated RAG, Retrieval-Augmented Generation, block is considered as part of the Reference Architecture. The RAG is an innovative system that combines the ability to retrieve information from a vast database with the ability to generate answers in natural language. In practice, it allows the OpenKIWAS to be queried using questions posed in natural language, receiving detailed and relevant answers directly from the system, without the need for advanced technical knowledge.

This approach makes interaction with the database much more fluid and natural for the user, who can formulate questions as he or she would in an everyday conversation, without worrying about the structure of the queries or the complexity of the system.

The RAG does not just provide pre-constructed answers, but has the ability to process information, reprocess it and present it in a coherent and easily understandable manner, adapting to complex queries. This makes the system highly versatile, capable of answering a wide range of questions efficiently and with an extremely intuitive user interface, now familiar from the widespread use of virtual assistants and chatbots. The adoption of this technology offers users significant added value, allowing them to explore the OpenKIWAS database without the need to manually navigate through catalogues and information, but obtaining immediate and precise answers to specific questions. In this way, users can access in-depth and detailed knowledge easily and directly, improving productivity and effective decision-making.

#### 4.2.6 Infrastructure

The *Infrastructure* of the IDEATION's Reference Architecture (Figure 7) must be conceived as a dynamic and scalable cloud-native ecosystem, designed to support the entire microservice architecture, guaranteeing flexibility, isolation, reliability and performance in the deployment and runtime operations of the components. Each time a new service or process is deployed, the system must be able to automatically allocate the necessary resources, creating the most suitable environment in terms of space, computational power, and technological compatibility.

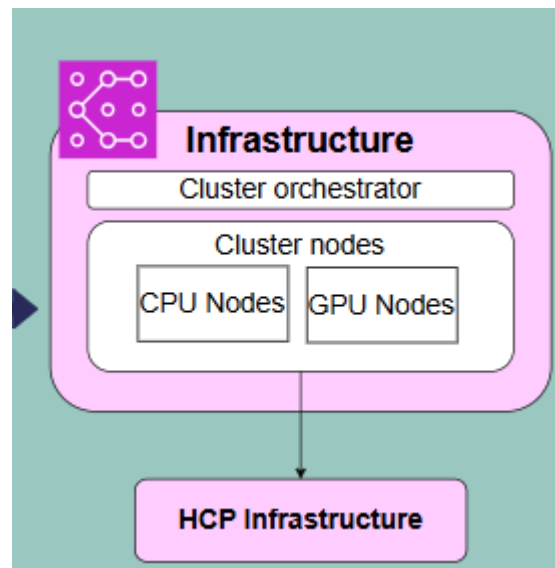


Figure 7 - Infrastructure section into the reference architecture

To meet these requirements, the *Infrastructure* must be based on a heterogeneous architecture, consisting of CPU and GPU nodes, in order to respond to both light and computationally intensive computing tasks, such as simulation or prediction tasks based on Machine Learning models that may be required as functionalities of the final Digital Twin of inland waters. GPUs, in particular, are crucial in parallel computing flows required by advanced predictive models, deep neural networks or high-resolution reanalysis algorithms. Managing the complexity of this architecture will be an *orchestrator*, whose job is to control the state of microservices, allocate resources, monitor load, dynamically scale components and ensure the resilience of the entire system. A tool such as Kubernetes<sup>30</sup> presents itself as an ideal candidate to fulfil this role, due to its maturity, large support community and compatibility with hybrid and distributed cloud environments. Kubernetes allows load balancing, auto-scaling, and resource management policies to be implemented for each pod and container, thus ensuring that each service or process is deployed with the correct resources and in the most efficient location in the *Infrastructure*. However, internal resources are not always sufficient, especially in computationally-intensive scenarios or during load peaks, which is why the infrastructure is designed to be open and integrable with external *High-Performance Computing* nodes, allowing it to temporarily or permanently extend its computational capacity. These nodes can belong to distributed computing centres, public cloud providers or institutional partners, and are reached via standard and secure interfaces, ensuring operational continuity and interoperability even on a federated scale.

<sup>30</sup> <https://kubernetes.io/>

## 4.3 Interoperability and compliance with the DTO

One of the founding goals of the DTO is to ensure seamless interoperability between the various modules, external systems and the international scientific community, as well as alignment with industry standards defined for Digital Twins. Starting from the bottom, each layer modelled in the IDEATION Reference Architecture is designed to be compliant with the DTO, so that it will be possible to exchange data, services and methodologies used.

The *Data Layer* aligns perfectly with the DTO: through a single *Data API*, in fact, the architecture exposes both native storage buckets and external S3 buckets, exactly as in the DTO. This means that those already using a personal S3 bucket for the DTO can continue to use the same resource within IDEATION, without having to reconfigure or replicate data. For example, a research team could set up a MinIO instance on-premise or in the cloud, configure it as an external S3 endpoint and immediately see their own buckets available within IDEATION. All NetCDF and GRIB files are treated according to the CF-Conventions, while geospatial components are made available as OGC services (WMS, WFS, WMTS). Consequently, any module of the DTO can query, display and download data without the need to develop dedicated adapters or handle proprietary formats. In this way, the Data Layer not only inherits full compatibility with the repositories of the oceanographic community, but also ensures the same level of reliability and standardisation required by the DTO reference model.

Moving up to the *DT Core*, interoperability with the DTO is realized through the dual exposure of API interfaces, which are distinct but perfectly integrated with each other. On the one hand, the *Service API* is responsible for the invocation of application functionality and compute module, on the other, the *Process API*, designed for the management and execution of operational flows, whether batch or event-driven. A key aspect of this interoperability lies in the fact that both services and processes rely on dedicated repositories, a *Service Repository* and a *Process Repository* respectively, which are structured and populated according to the same logical schema provided in the DTO. This means that the way components are loaded, organized, and registered follows exactly the same process, each service and process is described with files containing all parameters and definitions to then be instantiated and containerized in the infrastructure section. In practical terms, a service or process already developed for the DTO environment can be imported into IDEATION's DT Core without any structural changes, simply by registering it in the corresponding repository. This mechanism allows not only direct reuse of existing components, but also complete consistency in the way they are described, discovered, and invoked, benefiting both technical integration and collaboration between environments. The *DT Core*, therefore, does not merely replicate the behavior of the DTO, but adopts its entire conceptual and operational framework, effectively becoming a fully interoperable node within the DTO ecosystem.



Moreover, with regard to the application layer, the architecture was designed with the objective of ensuring full interoperability with the DTO, focusing in particular on aspects related to data visualisation, exploration and use. In this layer, an interactive, user-oriented solution takes shape, capable of accessing both the information contained in the semantic catalogue and the data stored in the bucket storage connected to the system, regardless of whether these are local or external. An essential aspect is that the representation of data does not imply its physical transfer or downloading within the IDEATION environment. Thanks to the use of OGC standards for geospatial services, such as WMS for raster visualisation, WFS for feature vectors, coupled with the use of standard-compliant APIs for accessing temporal and sensory streams, the platform is able to consume content hosted on remote infrastructures in real time, without duplicating it or altering its origin. This ensures not only efficiency in resource utilisation, but also transparency and adherence to FAIR. This paradigm enables federated visual interoperability in that by connecting directly to the DTO catalogue, which exposes metadata according to the DCAT model and makes geospatial and thematic endpoints available, it is possible to explore, integrate and represent, within the IDEATION interface, dynamic content that was not originally present in the system. Similarly, the representations developed within the IDEATION environment, being described and distributed with the same technological and semantic standards as the DTO, are fully compatible and can therefore be integrated into any other Digital Twin system that adopts the same architectural model. This makes possible not only the sharing of data, but also the dissemination of locally developed representations, analytical insights and visual narratives.

In the context of the infrastructure, the architecture is designed to ensure scalability, portability and full integration with the operating modes already adopted by the DTO, i.e. that all application and service components are released as standard-compliant containers, orchestrated through orchestrators such as Kubernetes, and ready to be deployed on any compatible environment that can range from public clouds to on-premise infrastructures, to edge nodes or HPC clusters.

Apart from the infrastructure, the interoperability between the DTO and IDEATION manifests itself most concretely in two key domains: services and processes, and data catalogues (see Figure 8). These areas represent critical points of contact, enabling the two systems not only to coexist within a common ecosystem, but also to collaborate actively and share capabilities dynamically. Firstly, the fully cloud-native approach of both architectures allows for direct sharing of service and process repositories, which means that a containerised webapp, developed and registered within the DTO's service repository, can be easily retrieved, orchestrated or reused by IDEATION as well, without the need for adaptation or reconfiguration. This level of interoperability makes it possible to maximise the reuse of components developed by the community, avoiding duplication and fostering a modular and federated growth of the Digital Twin system. The second fundamental point



of contact is the data-centric level, where interoperability is ensured through the use of federated catalogues and standardised interfaces. This allows, for example, IDEATION to expose its catalogue and make it visible to other Digital Twins or to draw on content already published in the DTO, while maintaining its traceability, metadata and original structure.

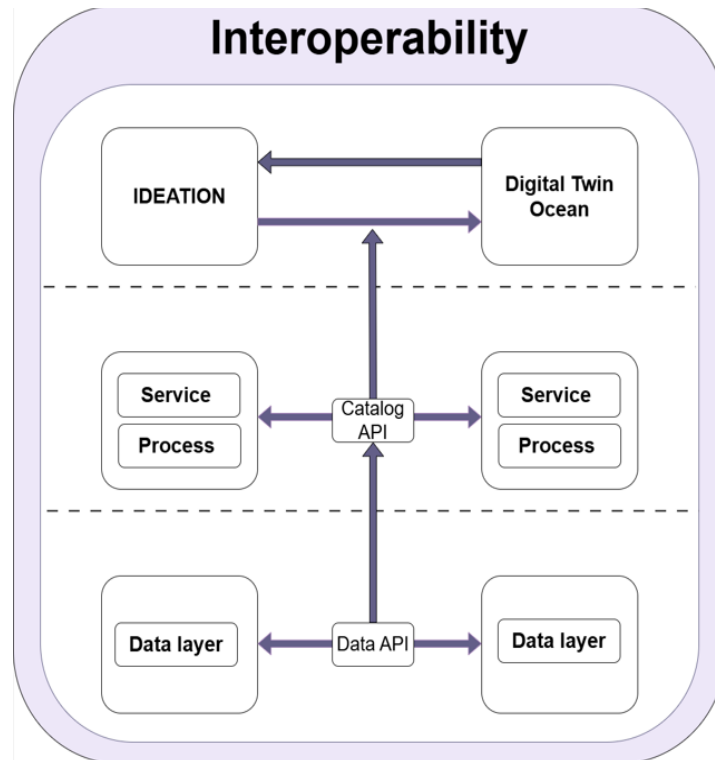


Figure 8 - Interoperability scheme with the DTO

## 5. CONCLUSIONS

The path traced in this document has made it possible to organically traverse the theoretical and practical foundations that guide the development of the Digital Twin Ocean (DTO) and the Destination Earth (DestinE) initiative, translating these principles into a concrete and operational architecture such as that proposed by IDEATION. This result was also achieved thanks to an in-depth investigation of the protocols, standards and APIs adopted by the main European Digital Twins, which provided a solid reference for the design of the platform. IDEATION is therefore configured not only as a system in line with European specifications and standards, but as an active and federated element within the DTO infrastructure. Each architectural level from the data layer to the application layer, passing through the management of services, processes and infrastructural resources, has been designed to guarantee openness, reuse and interconnection, according to a logic that favours the adoption of shared formats and protocols, the modularity of the components and the possibility of federating heterogeneous resources in a transparent manner. A decisive contribution to the structuring of the architecture also came from the formalisation of the principles for interoperability analysed in the course of the work, which made it possible to build a more solid and systematic vision of how each functional block should be structured and how they should communicate within a federated context. In this first version, some of the functional and non-functional requirements identified in task T5.1 have been considered, but in the final version of the architecture, these requirements will be analysed and mapped in their entirety, further consolidating consistency with the European reference frameworks.

This version of the architecture is an initial version, destined to be refined and enhanced in the second half of the project thanks to the inputs from the Reference Use Case (WP2, T2.3) and the mapping between functional and non-functional requirements and current solutions (WP5, T5.1) that will be reported at M18 (November 2025) and M12 (May 2025), respectively. Its role is to lay the foundations of a structural system that is consistent with European standards, but at the same time open to evolution and continuous improvement. In particular, the architecture will be progressively adapted and enriched on the basis of feedback from stakeholders, whose operational and design requirements represent a fundamental element in guaranteeing the effectiveness and adoption of the system. Similarly, the work of collecting and analysing functional and non-functional requirements will continue to be refined through further MultiStakeholder Forums, with the aim of providing as exhaustive and detailed a vision as possible, which will then be incorporated into the logical blocks. This activity will make it possible to consolidate the initial architectural choices, but also to identify any critical points, opportunities for extension or adaptation needs.

In the next phase, the architecture will also directly benefit from the results of T4.3, which is



[www.ideation-project.eu](http://www.ideation-project.eu)

dedicated to the definition of guidelines for the design and validation of physical models and artificial intelligence algorithms geared towards the Digital Twin of inland waters. The indications that will emerge from this work will be fundamental in ensuring the reliability, reproducibility, and applicability of the computational models integrated in the system, and will provide a structured framework for data preparation, model selection, performance verification and compliance with regulatory standards.

**IDEATION - D4.1: Reference architecture and Interoperability  
Guidelines (KER3) V1**



## REFERENCES

- A. Abella, M. Ortiz-de-Urbina-Criado, and C. De Pablos-Heredero, "Meloda 5: A metric to assess open data reusability," *El Profesional de la Información*, vol. 28, Jan. 2020, doi: 10.3145/epi.2019.nov.20.
- Abernathy, R., Hamman, J., Rocklin, D., Gentemann, C., & Hwang, L. (2021). Pangeo Forge: Crowdsourcing analysis-ready, cloud-optimized data production. *Frontiers in Climate*, 3, 782909. <https://doi.org/10.3389/fclim.2021.782909>
- Gos, Konrad & Zabierowski, Wojciech. (2020). The Comparison of Microservice and Monolithic Architecture. 150-153. 10.1109/MEMSTECH49584.2020.9109514
- Tapia, F.; Mora, M.Á.; Fuertes, W.; Aules, H.; Flores, E.; Toulkeridis, T. From Monolithic Systems to Microservices: A Comparative Study of Performance. *Appl. Sci.* 2020, 10, 5797. <https://doi.org/10.3390/app10175797>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>



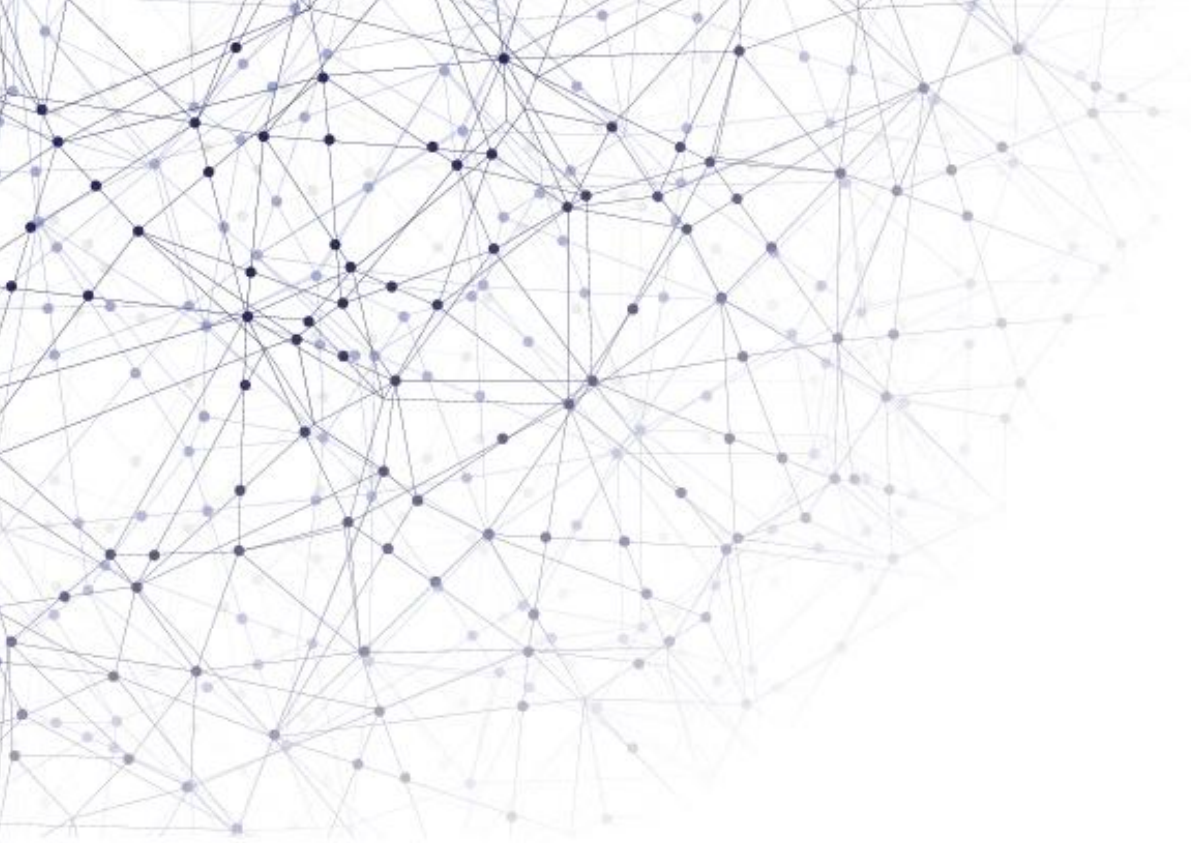
IDEATION to prepare the development of the digital twin of the inland waters (rivers, lakes, reservoirs, wetlands, snow, and ice) addressing activities to be developed and to make it integrated and interoperable with the DTO for a unified digital twin of ocean and waters.



# IDEATION



Funded by  
the European Union



# IDEATION



Funded by  
the European Union