Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

multi-Risk sciEnce for resilienT commUnities undeR a changiNgclimate

Codice progetto MUR: **PE00000005** – CUP LEAD PARTNER C93C22005160002

Deliverable title: Design and prototyping of data collection, pre-processing and integration tools to support the execution of the models and to manage the output

Deliverable ID: 1.5.2

Due date: June 1st, 2025

Submission date: June 1st, 2025

AUTHORS

Tasso Alberto (Tasso A.); Paolo Campanella (Campanella P.); Menapace Marco (Menapace M.): Pintus Fabio (Pintus F.)

1. Technical references

Project Acronym	RETURN
Project Title	multi-Risk sciEnce for resilienT commUnities undeR a changiNg climate
Project Coordinator	Domenico Calcaterra UNIVERSITA DEGLI STUDI DI NAPOLI FEDERICO II domcalca@unina.it
Project Duration	December 2022 – November 2025 (36 months)

Deliverable No.	DV#.1.5. 2 - Design and prototyping of data collection, pre-processing and integration tools to support the execution of the models and to manage the output
Dissemination level*	
Work Package	WP#1.5
Task	T#.# - Task Title
Lead beneficiary	CIMA, POLIMI
Contributing beneficiary/ies	CIMA, POLIMI

* PU = Public

PP = Restricted to other programme participants (including the Commission Services)

RE = Restricted to a group specified by the consortium (including the Commission Services)

CO = Confidential, only for members of the consortium (including the Commission Services)

Document history

Version	Date	Lead contributor	Description
0.1	02/05/2025	Alberto Tasso (CIMA)	First draft
0.2	15/05/2025	Francesco Ballio (POLIMI)	Critical review and proofreading
0.3	20/05/2025	Fabio Pintus (CIMA)	Edits for approval
1.0	31/05/2025	Francesco Ballio (POLIMI)	Final version

2. ABSTRACT

This document describes the proposed solution for managing the data and services to be integrated into the Digital Ecosystem foreseen in the Spoke VS1 of the “Return” project.

By leveraging the WASDI platform's integration features, it will be possible to connect existing data sources and deploy existing models directly onto the platform. While this will address the main needs of the scientific community, the Digital Ecosystem also needs to provide dedicated features to enable seamless data integration and workflow execution.

The Digital Ecosystem will include a unified geospatial data service that will enable users to ingest geospatial data into the system if it is not available as an external service that can be federated directly as a WASDI data provider.

It will also include an integrated workflow engine designed to orchestrate the execution of WASDI processors directly from the Digital Ecosystem.

Finally, it will provide a set of basic operations to be performed on geospatial data, offering a foundation for data analysis.

3. Table of contents

List of Figures

3. Introduction

This document outlines a proposed solution for managing the data and services slated for integration into the Digital Ecosystem foreseen in Spoke VS1 of the “Return” project

We plan to leverage the WASDI platform's integration features to connect existing data sources and deploy pre-existing models directly onto the platform. While this approach will address the primary needs of the scientific community, the Digital Ecosystem also requires dedicated functionalities to ensure seamless data integration and efficient workflow execution.

The Digital Ecosystem will feature a unified geospatial data service. This service will enable users to ingest geospatial data directly into the system, particularly when it's not available as an external service that can be federated as a WASDI data provider. Furthermore, it will include an integrated workflow engine, designed to orchestrate the execution of WASDI processors directly from within the Digital Ecosystem. Finally, the system will provide a foundational set of basic operations for geospatial data analysis

4. Geospatial Data Service

In today's rapidly evolving digital landscape, efficiently managing geospatial information presents significant challenges. Users often struggle to integrate heterogeneous datasets, ranging from high-resolution raster data to intricate vector maps, and to make them easily discoverable and accessible through standardised interfaces. Traditional static catalogues or fragmented data repositories often prove inadequate for dynamic data updates and seamless interoperability across various platforms. This proposed service addresses these complexities directly by introducing a modern, flexible architecture that unifies data cataloguing, granular feature access and bespoke API integration into a single, cohesive solution.

Our approach centers on delivering a dynamic, user-centric geospatial data service that is built upon robust open standards. The service primarily leverages the SpatioTemporal Asset Catalog (STAC) specification for comprehensive data discovery and the Open Geospatial Consortium (OGC) API – Features standard for precise, feature-level access to vector datasets. A key feature of this solution is the inclusion of a specialised REST API layer that has been designed to integrate seamlessly as an external data provider within the WASDI platform. This ensures that the rich array of geospatial assets managed by our service can be consumed and utilised directly within WASDI's powerful processing and analysis environments.

4.1 Service Architecture

At the heart of this solution lies a web service designed to orchestrate the dynamic handling of geospatial data. It intelligently interacts with a central metadata repository and a robust geospatial feature store to provide a fully dynamic, standards-compliant geospatial data experience.

The service's comprehensive functionalities are organised into distinct yet interconnected layers: Data Ingestion and Processing; the STAC API for Discovery; the OGC API – Features for granular access; and a specialised WASDI Integration API.

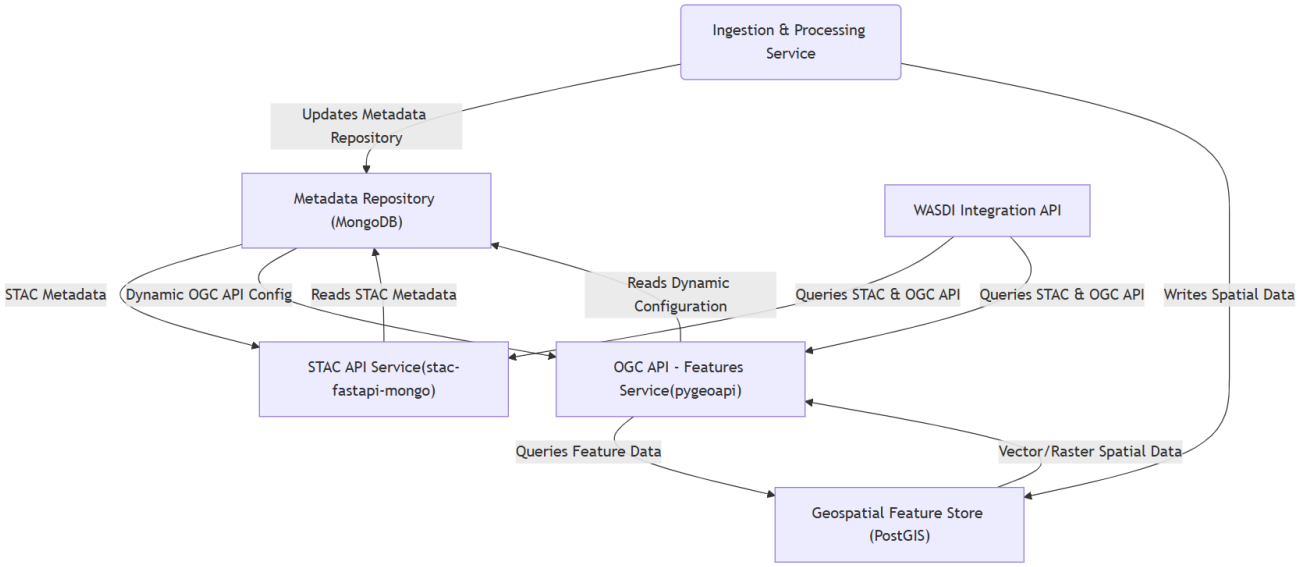

Figure 1 - Geospatial Data Service Architecture

4.2 Service features

The Data Ingestion and Processing layer is the first point of contact for all incoming geospatial assets. This robust engine is designed to handle new data dynamically, accommodating various formats including raster files and vectorial data such as zipped Shapefiles or GeoPackages. A key feature of this layer is its sophisticated handling of the accompanying metadata. For vector files containing multiple distinct features, the service supports supplementary tabular data in Excel or CSV format. This enables the precise association of rich attributes, identified by a unique ID, with individual features within the vector dataset. Upon successful ingestion, raster data is processed, and its spatial information is integrated, while vector data is meticulously parsed and loaded into unique, newly generated structures within the geospatial feature store. This ensures optimal performance and scalability for subsequent querying operations. The entire ingestion and processing workflow is fully automated, including the extraction of essential metadata and the dynamic generation of configuration entries required by subsequent API layers.

The STAC API is the central hub for asset discoverability. Powered by *stac-fastapi-mongo* library, this component exposes a compliant STAC API, enabling client applications to efficiently search and filter the entire catalogue of geospatial assets. All metadata, whether provided by the user or extracted and enriched automatically during ingestion, is stored in the metadata repository. Clients can perform highly granular queries based on spatial extents, temporal ranges and a rich set of properties, including any custom attributes derived from the accompanying tabular metadata. Each STAC item represents a single geospatial asset and provides direct links to its information within the geospatial feature store. For vector datasets, these STAC items are enriched with an alternative link that seamlessly guides clients to the corresponding OGC API – Features endpoint for precise, feature-level access.

The OGC API – Features complements the STAC discovery layer by providing standard-compliant access to individual vector features. This endpoint is dynamically served by an embedded *pygeoapi* instance whose resource definitions are managed in real time and retrieved from the metadata repository. This innovative approach ensures that, as new vector datasets are ingested and new structures are populated within the geospatial feature store, the corresponding OGC API – Features collections are automatically exposed and made available, eliminating the need for service restarts. Clients can query these collections to list all available features, retrieve specific features by their unique identifier or apply sophisticated filters based on spatial bounds, temporal references and detailed attribute values, making use of the robust filtering capabilities inherent in the OGC API – Features standard.

Finally, to ensure seamless integration with the WASDI platform, the service incorporates a dedicated WASDI Integration API. This specialised REST API layer has been designed to conform to WASDI's external data provider specifications. It allows WASDI to consume and leverage the extensive geospatial data hosted by our service directly. The integration layer translates WASDI's data request patterns into optimised queries for our internal STAC and OGC API – Features components. This allows WASDI users to effortlessly browse, search and access all ingested data types — including both rasters and vectors — directly within their existing WASDI workflows. This maximises the utility and analytical potential of the data.

5. Workflow Execution Engine

Executing workflow pipelines on the WASDI platform often requires the concatenation of operations that cannot be performed manually by end users and stakeholders in the digital ecosystem.

The proposed Workflow Execution Engine meets this need by orchestrating WASDI tasks programmatically. This system eliminates the inefficiencies and errors of manual processes, ensuring consistent results and enabling the scalability and reproducibility of geospatial analysis.

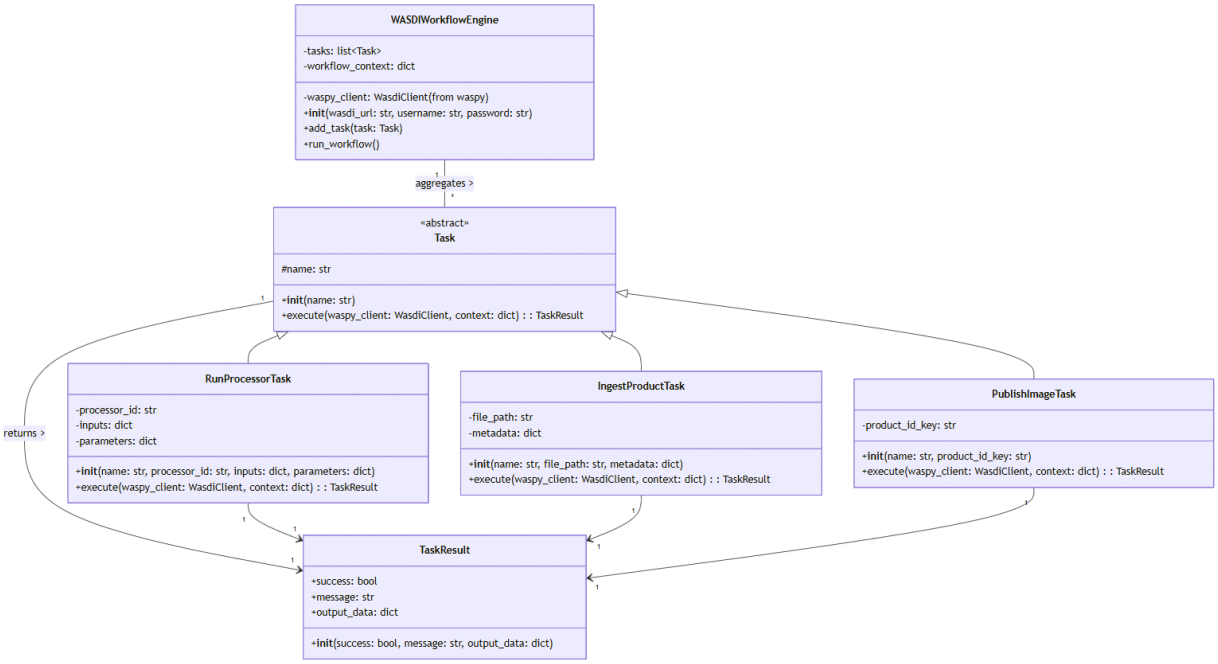

Figure 2 - Workflow Engine Architecture

The workflow engine facilitates the structured execution of a series of operations on the WASDI platform.

Each workflow is defined programmatically by adding specific tasks sequentially to the Workflow Engine, thereby establishing the sequence of operations.

Once all the tasks have been defined, the application or user triggers the execution of the workflow on the engine. Upon activation, the engine iterates through its registered tasks, executing each one in turn. For each task, the engine invokes its 'execute' method, passing along a shared workflow context. This context serves as a central repository for any data that needs to be exchanged between tasks.

Three basic workflows are provided out of the box:

- Ingest Product: This initial step focuses on introducing new data into the WASDI environment. It involves retrieving an image or any other relevant product file, along with its associated metadata, from the platform's data catalog. This makes the product available for subsequent processing or analysis.
- Run Processor: Following ingestion, this step executes a specific WASDI processor on the previously ingested product. This could involve applying various algorithms for data transformation, feature extraction, or thematic mapping (e.g., calculating NDVI). The output of this processing step is typically a new, derived product.

- Publish Image: The final step involves making a processed image or product publicly accessible within WASDI. This operation ensures that the results of the workflow are discoverable and usable by a wider audience, facilitating data sharing and dissemination.

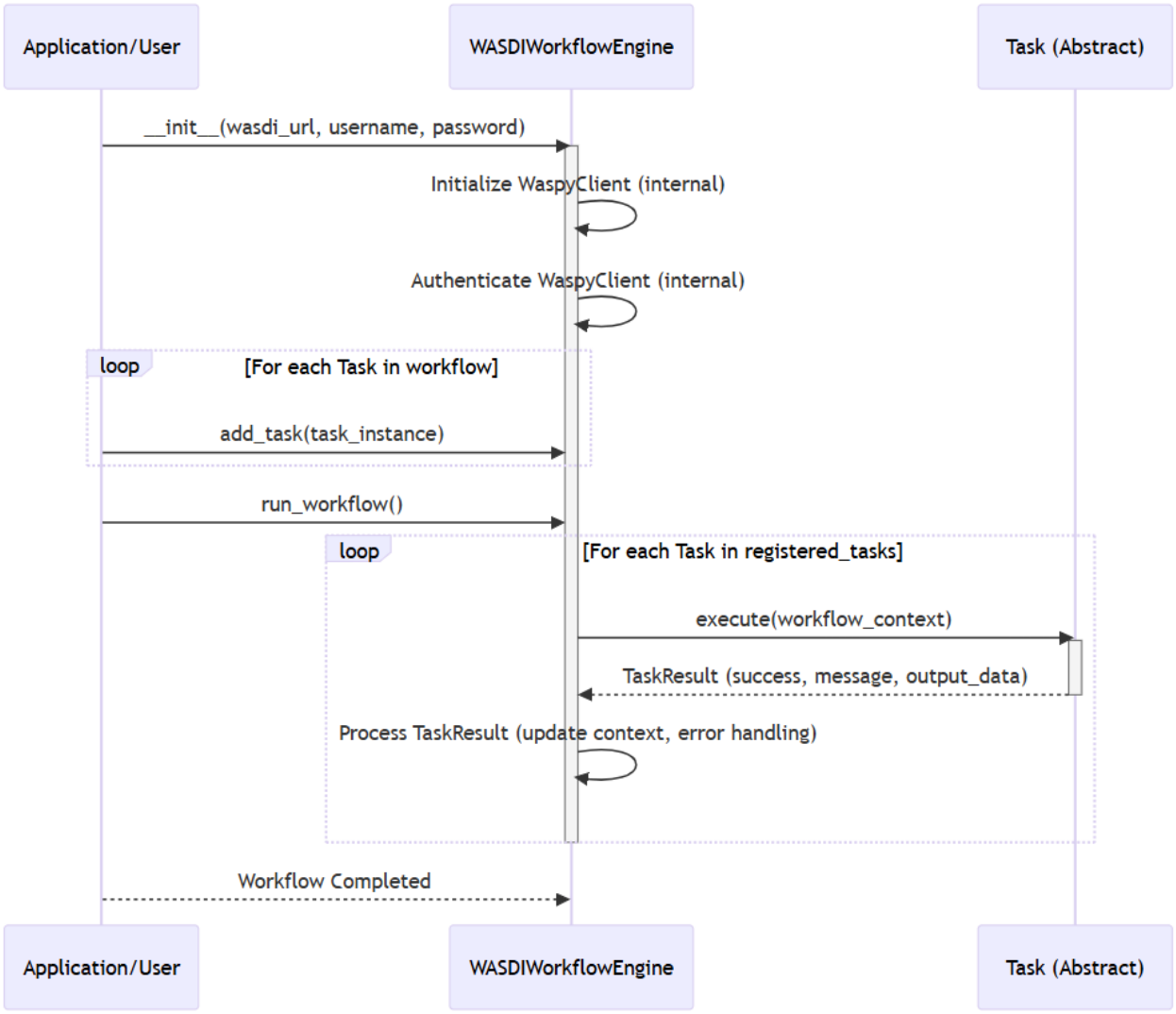

Figure 3 - Workflow initialization and execution

6. Geotools Service

A critical component within this ecosystem is a dedicated service providing a robust set of tools for in-depth analysis of data sourced from various inputs, whether it's directly extracted from available sources or generated by other models and services within the ecosystem. This analytical capability is structured around the following key features, each offering unique insights into spatial and temporal patterns.

- **Automatic hydrological basin identification**: this service provides a streamlined method for delineating hydrological basins, also known as watersheds or catchments, using the accumulation computed from flow direction extracted from a Digital Elevation Models (DEMs). Each identified basin represents the entire geographic area that contributes water to a specific outlet point, such as a river mouth or a lake. The purpose of this feature is to automate what has traditionally been a complex and time-consuming manual process, making it accessible and efficient for a wide range of applications.
- **Statistics calculations**: this service provides the capability for calculating aggregated statistics over user-defined areas on a spatial product. This feature allows for the rapid extraction of key summary metrics from geospatial datasets, enabling a deeper understanding of spatial distributions and characteristics. Given a specified area, the service can compute various statistical measures from the underlying spatial data. This includes determining the maximum value, useful for identifying peak intensities like the highest pollution concentration in a region or the warmest temperature. It also calculates the minimum value, which could pinpoint areas of lowest risk or minimal impact. The average value provides a representative measure of the central tendency within the selected area, offering insights into typical conditions, such as average rainfall over a specific agricultural zone. Furthermore, the standard deviation quantifies the dispersion or variability of values around the mean, indicating the spread of data and highlighting areas of greater heterogeneity. The purpose of this feature is to simplify the process of gaining quantitative insights from complex spatial data, supporting informed decision-making in fields ranging from environmental monitoring to urban planning and resource allocation.
- **Trend analysis**: this service is designed to extract temporal patterns from sequences of geospatial products. When provided with a series of spatial datasets, ordered chronologically, this feature calculates the sequence of **aggregated statistics** over a user-defined area or in a single point for each element within that sequence. This capability is paramount for understanding how phenomena evolve over time
- **Overlay analysis**: this feature enables users to combine two or more spatial datasets to create a new dataset that contains the attributes and spatial characteristics of both. While more complex overlays exist, a simple intersection or union can be immensely valuable. For instance, *intersection* allows users to identify areas where two features overlap, such as agricultural land within a flood plain, or residential zones within a specific administrative district. This reveals common areas where different spatial criteria apply. *Union* combines the geometries of two datasets, creating a new feature that encompasses the combined extent of both inputs, while preserving their respective attributes. The purpose of simple overlay analysis is to answer "where" questions by identifying areas that meet multiple spatial criteria
- **Proximity analysis**: this feature enables the calculation of distances and relationships between spatial features, providing insights into their closeness or separation. Users could define a point, line, or polygon, and the service would then determine the closest features of another type, or compute distances to a defined boundary. For instance, it could identify the nearest school to a residential area, calculate the distance of a new development to existing infrastructure like roads or utility lines, or determine how far a pollution source is from a sensitive ecological zone.

7. Conclusions

The proposed solutions aim to create a robust and efficient Digital Ecosystem that supports the goals of the RETURN project in addressing multi-risk science for resilient communities under a changing climate. By integrating these services, researchers and stakeholders will have access to the tools and data needed to conduct complex analyses and inform decision-making processes.

The Geospatial Data Service, leveraging STAC and OGC API standards, provides a dynamic and user-centric approach to managing diverse geospatial datasets. It ensures efficient data discovery, granular feature access, and seamless integration with the WASDI platform.

The Workflow Execution Engine automates complex operations within the WASDI platform, streamlining workflows and ensuring reproducibility. It handles tasks such as data ingestion, processor execution, and product publishing, thereby enhancing efficiency and reducing errors.

The Geotools Service provides a suite of analytical capabilities for in-depth data analysis. This includes automatic hydrological basin identification, statistics calculations, trend analysis, overlay analysis, and proximity analysis. These tools offer critical insights into spatial and temporal patterns within the data.

Further development and testing of these components will be crucial to ensure their effectiveness and scalability within the project's timeframe. Future work may also include expanding the range of Geotools services and enhancing the integration capabilities with other external data sources.

8. References

Bartos, Matt (2020). pysheds: simple and fast watershed delineation in python. https://doi.org/10.5281/zenodo.3822494