SEcure Decentralized Intelligent Data MARKetplace

D5.5 Demonstrators integration, testing and assessment of system performance. First version

Document Identification	
Contractual delivery date:	31/12/2024
Actual delivery date:	15/01/2025
Responsible beneficiary:	MYT
Contributing beneficiaries:	All partners
Dissemination level:	PU
Version:	V1.0
Status:	Final

Keywords:

Use cases, development process, datasets, supported scenarios, evaluation methodology, evaluation process, evaluation criteria, trial definitions, KPIs

This document is issued within the frame and for the purpose of the SEDIMARK project. This project has received funding from the European Union's Horizon Europe Framework Programme under Grant Agreement No. 101070074. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the European Commission.

Document Information

Document Identification			
Related WP	WP5	**Related Deliverables(s):**	-
Document reference:	SEDIMARK_D5.5	**Total number of pages:**	585 8

List of Contributors	
Name	**Partner**
Panagiotis Vlacheas Grigorios Koutantos	WINGS
John Tsogias Nikos Babis	MYT
Franck Le Gall Thomas Bousselin	EGM
Juan Echevarria Sonia Sotero	SDR
Eero Jalo	FV
Luis Sánchez Victor González Pablo Sotres	UC
Elias Tragos Diarmuid O'Reilly Morgan Erika Duriakova Honghui Du	NUID-UCD
Miguel Angel Esbri	ATOS/EVIDEN

Document History			
0.1	13/11/2024	WINGS	First draft of ToCS
0.2	20/12/2024	MYT	First Consolidated Version, included the first contributions of MYT and SDR/UC
0.3	23/12/2024	MYT	Added contributions of FV, SDR/UC, EGM, MYT.
0.4	3/01/2025	MYT	Formatting, additions to section 4.3.
0.5	7/01/2025	MYT	Formatting, table of figures, list of tables, bibliography added.
0.51	7/01/2025	NUID-UCD	Contribution to section 2.1.5
0.6	8/01/2025	UC	Contributions / Small review to Santander use case
0.7	8/01/2025	FV EGM	Contributions to section 4.1 Contributions to section 4.4
0.9	10/01/2025	MYT	Version ready for peer review
0.92	13/01/2025	MYT	Version ready for quality review
0.95	14/01/2025	ATOS/EVIDEN	Quality Format Review
1.0	16/01/2025	ATOS/EVIDEN	FINAL VERSION TO BE SUBMITTED

Quality Control		
Role	**Who (Partner short name)**	**Approval date**
Reviewer 1	Danciu Gabriel (SIE)	13/01/2025
Reviewer 2	Miguel Angel Esbri (ATOS/EVIDEN)	13/01/2025
Quality manager	María Guadalupe Rodríguez (ATOS/EVIDEN)	16/01/2025
Project Coordinator	Miguel Angel Esbri (ATOS/EVIDEN)	16/01/2025

Table of Contents

List of Tables

List of Figures

List of Acronyms

Abbreviation / Acronym	Description
AI	Artificial Intelligence
API	Application Programming Interface
AWS	Amazon Web Services
BI	Baseline Infrastructure
BIFs	Baseline Infrastructure Facilitators
CARF	French Riviera Agglomeration Community
CI/CD	Continuous Integration and Continuous Delivery/Continuous Deployment
CKAN	Comprehensive Knowledge Archive Network
CLI	Command Line Interfaces
CPU	Central Processing Unit
DAG	Directed Acyclic Graph
DCAT	Data Catalog Vocabulary
DID	Decentralized Identifier
DLT	Distributed Ledger Technology
DRACO	Data Reduction and Calibration Operation
DT	Digital Twin
Dx,y	Deliverable number y belonging to WP x
EC	European Commission
ECDSA	Elliptic Curve Digital Signature Algorithm
ERC	Environmental Regulation Commission
GIS	Geographic Information System
GUI	Graphical User Interface
GWs	Gateways
HRI	Helsinki Region Information
HSY	Helsinki Region Environmental Services
HTTPS	Hypertext Transfer Protocol Secure

Abbreviation / Acronym	Description
IdM	Identity Management
IEC	International Electrotechnical Commission
IM	Identity Management
IoT	Internet of Things
IOTA	Internet of Things Application
ISO	International Organization for Standardization
JRE	Java Runtime Environment
JSON	JavaScript Object Notation
KPIs	Key Performance Indicators
LoRaWAN	Lo(ng) Ra(nge) Wide Area Network
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLOps	Machine Learning Operations
MSE	Mean Square Error
MVP	Minimum Viable Product
NAP	Network Access Point
NFT	Non-fungible tokens
NGSI-LD	Next Generation Service Interfaces for Linked Data
ODRL	Open Digital Rights Language
OS	Operating System
P2P	Peer-to-peer
PCA	Principal component Analysis
PDP	Policy Decision Point
PEP	Policy Enforcement Point
RDF	Resource Description Framework
PoC	Proof of Concept
RERUM	REliable, Resilient and secUre IoT for sMart city applications
RESTful	REpresentational State Transfer
SDI	Serial Digital Interface

Abbreviation / Acronym	Description
SHA	Secure Hash Algorithm
SPARQL	SPARQL Protocol and RDF Query Language
SQuaRE	Software Quality Assessments and Recommendations
SSCP	Santander Smart City Platform
SSI	Self-Sovereign Identity
SSL	Secure Sockets Layer
SVN	Software Virtual Network
ToC	Table of Contents
UC	Use Case
UMAP	Uniform Manifold Approximation and Projection
URL	Uniform Resource Locator
UUID	Universally Unique IDentifier
VC	Verifiable Credentials
VMs	Virtual Machines
WMS/WFS	Web Mapping Service/ Web Feature Service
WP	Work Package
XML	eXtensible Markup Language
YALM	Yet Another Markup Language
ZIP	Zone Improvement Plan

Executive Summary

The document is the second deliverable of WP5: Integration, testing and evaluation of the SEDIMARK platform and reports the results of planning, execution, and evaluation of pilot trials designed to integrate the digital tools of SEDIMARK for urban and environmental management across Europe. Structured into six main sections, it provides insights into the project's objectives, methodologies, and outcomes.

The introduction establishes the purpose of the document, its relationship to other project tasks, and its overall structure. It sets the stage for understanding how the goals of the project align with broader work packages and outlines the document's roadmap for readers.

The core of the document focuses on the refinement of trial definitions for the four distinct pilot sites, each addressing region-specific challenges.

The subsequent section delves into the integration of demonstrator systems at each pilot site. These systems are designed to interact seamlessly with the SEDIMARK toolbox. The integration of each trial is detailed, covering the systems used and their interactions, ensuring alignment with project goals.

Testing and performance assessment follow as critical components of the evaluation phase. Each pilot undergoes rigorous testing to ensure the effectiveness of the systems and their ability to meet predefined KPIs. The connection between the local systems and the SEDIMARK toolbox is emphasised as a key aspect of achieving desired outcomes.

The document concludes with a summary of the key findings and insights derived from the trials. This section highlights the challenges encountered, lessons learned, and actionable recommendations for future implementations. The project's results serve as a foundation for advancing similar initiatives in other regions or sectors.

1 Introduction

1.1 Purpose of the document

The main purpose of the deliverable is to outline the design, implementation, and evaluation of pilot trials aimed at integrating innovative digital tools for urban and environmental management. It serves to define the trial objectives, datasets, and key performance indicators while detailing the integration of systems and their interaction with the SEDIMARK toolbox. Additionally, it provides a framework for assessing the performance of these solutions, offering insights and recommendations to guide future initiatives and drive sustainable development in diverse urban and environmental contexts.

1.2 Relation to other work packages and tasks

This deliverable is the result of the work done during the first year of the project and especially following the first integrated delivery of the platform. This deliverable can also be viewed as a follow-up to SEDIMARK_D5.1 [1], a very important deliverable because it establishes the context for the integration activities, as well as the evaluation process and methodology. The work presented in this document is linked to T2.1 (Use Case definition) and document SEDIMARK_D2.1 [2] as well as SEDIMARK_D2.2 [3]. The defined use cases are important in WP5 (SEDIMARK_D5.3 [4]) because they serve as practical examples of the solutions being developed. This document establishes a framework that thoroughly evaluates and monitors the performance of the solutions using a detailed evaluation framework and performance metrics.

1.3 Structure of the document

This document is structured into 7 major chapters:

- **Chapter 1** is the current chapter and introduces the objective of the document, its structure and how it relates to the activities of the project.
- **Chapter 2** presents a refinement of the trial definitions by the pilot leaders.
- **Chapter 3** analyses the integration process of the demonstrators (pilot leaders). Each pilot leader outlines the datasets per trial site and the associated KPIs.
- **Chapter 4** describes the integration of the systems for each pilot leader as well as the interaction with the SEDIMARK toolbox.
- **Chapter 5** includes a detailed description of the testing and assessment methods of the system performance by the pilot leaders.
- **Chapter 6** concludes the document, summarising the main outcomes and the future steps in alignment with the objectives and project roadmap.

2 Trial definition refinement

In this section, the document builds on the foundations laid in SEDIMARK_D5.1, focusing on the refinement of trial definitions for each pilot site with updated details on datasets and key performance indicators (KPIs) to align with evolving project objectives. The section outlines the datasets required for the four trials. Additionally, trial definitions and KPIs are revisited and refined for each site, incorporating insights gained since SEDIMARK_D5.1 to ensure these elements remain relevant and actionable, further enhancing the capacity of the trials to deliver meaningful results.

2.1 Datasets per trial site

The available trial sites include advanced digital systems to address challenges in specific sectors. Mobility and traffic infrastructure integration, planning of urban bike mobility for enhancement of municipal services, advanced AI techniques that predict energy distribution and / or churn prediction and valuation / commercialization of water data to optimize resource management and support sustainable development are all initiatives that exemplify how cutting-edge technology and data-driven strategies can address urban and environmental challenges across Europe.

2.1.1 Mobility Digital Twin in Helsinki

Helsinki use case wishes to utilize external data sources as part of its digital twin, to enhance the (local) data economy, and to diversify the options for data acquisition and management.

Helsinki's digital twin comprises a network of interoperable systems that exchange data using standardized APIs, enabling real-time urban management insights. Geospatial data forms the backbone of the digital twin, on which additional topic specific data assets can be overlayed. The digital twin approach has also been introduced in the field of mobility. Here the digital twin is a means to combine information from different data sources describing the traffic infrastructure and environment, the traffic itself, and related conditions and context. It thus comprises numerous data sources. So far, the digital twin of mobility has been developed on a conceptual level. However, potential data sources belonging to it already exist and are available from Helsinki.

As the Mobility Digital Twin in Helsinki is being developed as a 'system of systems' at a conceptual level, interoperability and machine readability are emphasized concerning the data marketplace. The next phase of development for the Mobility Digital Twin is currently in progress, to collect all traffic volume and sensor data to Helsinki city's own system on a database (Azure) instance. This initiative is underway as part of The City of Helsinki's Smart Transport Program 2030.

The types of data are:

- Infrastructure: The Register of Public Areas in the City of Helsinki contains data about the city's "street and green areas," namely street network as polygons, i.e., the area the street, road or a path occupies, with additional administrative information, such as classification and maintenance responsibilities. The registry is available in WFS format (https://kartta.hel.fi/ws/geoserver/avoindata/wfs). The data is also available at https://kartta.hel.fi/.

- Mobility / traffic: The city maintains a number of automated traffic counters (based on induction loops) that provide data over an open API. Induction loops are physical sensors embedded in the road surface that use electromagnetic fields to detect vehicles passing over them. These loops help in collecting data on traffic volume and patterns, allowing for effective traffic management and planning.
- Conditions: Helsinki Region Environmental Services HSY maintains a set of air quality measuring stations providing information on air quality in the city, available over an open interface).

The relationship between the data marketplace and the urban digital twin of mobility is envisioned to become two-directional. As the digital twin of mobility is formed as a "system of systems", the significance of interoperability and machine readability is highlighted concerning the data marketplace.

2.1.2 Urban bike mobility planning in Santander

Santander City Council, as it was explained in SEDIMARK_D5.1, has set up the Santander Smart City Platform (SSCP) which constitutes a centralized point to gather and distribute data from and to municipality departments.

The Urban bike mobility planning use case that is being implemented leverages the SSCP by accessing the data stored in it, having at its disposal those data that are specific to the use case and other datasets that are available and can complement the main pieces of information that will be employed within the use case. On the other hand, the SSCP will, eventually be able, once the Marketplace is developed, to play the opposite role and collect information that is shared through the Marketplace by other providers, and it is considered useful for the activities of the municipal departments.

Regarding the SSCP data sources that feed the Marketplace, the following is an update of what was already described in SEDIMARK_D5.1:

- The municipality electric bicycle rental service is already running and operating normally. Currently, the following sources of data are being used:
    - On-bicycle: there are 221 bicycles that provide periodic updates when the bicycle is parked, along with an event-based notification approach: status change (parked, on route, etc.), battery level changes and location changes.
    - Bicycle stations: there are 28 docking stations (see Figure 1, left) also providing information with an event-based notification approach. This information includes changes in status (working, on maintenance, etc.), as well as changes in available bicycles and empty slots.
    - The previous bicycle rental system has been cancelled when the new one came into operation.
    - Data on the use of additional public bicycle racks that are available for citizens to park there their bicycles and provide information about its usage.
    - The development of new devices under the project has been discarded. These devices were intended to be installed on volunteers' bicycles, but the municipality rental service is extensive and well distributed throughout the city and already provides very rich

information for the use case. Thus, it was considered that the effort required for this activity was not worth considering the reduced additional input it could provide.

- Data from the bicycle counting sensors mounted on totems (see Figure 1, right) located in some lanes of the city might be imported to SSCP and consequently ready for being shared through the Marketplace in the framework of the use case.

Figure 1: Santander e-bikes and totem

In all cases, as it has been described, data is already available through the SSCP API, which plays the role of the Provider's domain in the system architecture defined in SEDIMARK_D2.3. Specifically, the SSCP allows access to the data through an NGSIv2 API and data model. Thus, the necessary data mappers have been developed to integrate the available datasets and made them ready for consumption in the Marketplace Information model which is based on NGSI-LD.

2.1.3 Valorisation of energy consumption and customer reactions/complaints in Greece

In Greece, analysing energy consumption patterns and customer reactions is vital for optimizing energy distribution and enhancing the customer experience. Energy consumption data offers valuable insights into usage behaviours, seasonal trends, and regional variations. Integrating this data with customer feedback, such as complaints about billing issues, service disruptions, or dissatisfaction with energy providers, provides a comprehensive understanding of the factors influencing energy demand and customer loyalty. These insights enable the development of targeted interventions to reduce churn, improved customer service strategies, and dynamic pricing models aligned with consumption trends.

The trial will evaluate the accuracy of energy consumption predictions and assess how closely these align with actual usage patterns. It will also analyse clustering processes to determine whether users with similar consumption behaviours are grouped effectively. The accuracy of energy consumption predictions will be tested against historical data from a diverse set of users, ensuring the reliability of the outputs. Datasets for the trial include residential energy consumption records, customer feedback (e.g., complaints and satisfaction levels), geospatial information (e.g., zip codes), and weather data. These datasets will be used to explore correlations between consumption spikes, customer complaints, and external factors such as temperature or payment delays. Details are described in what follows:

- Contents: Numerical data (e.g., energy consumption, residential size, supply IDs), text fields, weather data (temperature, humidity), and wind direction data.
- Format: CSV files with columns for approximately 5,000 unique customers.
- Data Sources: Stored in MYT's corporate data warehouse.
- Anonymization: Public data but includes anonymized energy-oriented data (e.g., weather data, residential size).

The above have been mentioned in previous deliverables (e.g., SEDIMARK_D5.1) as well. For churn prediction, the trial will focus on evaluating the ability of machine learning methods to identify customers at risk of leaving and to group them into meaningful segments. Using anonymized private data, it will assess the accuracy, reliability, and fairness of the predictions. Key performance indicators such as accuracy, recall, and precision will measure the effectiveness of churn identification while minimizing errors. Comparisons with baseline models, such as logistic regression, will validate the added value of advanced AI techniques. The models will generate churn probabilities for each customer, and clustering algorithms will group users based on shared behaviours or risks. These clusters will be rigorously evaluated for their coherence and relevance, with expert validation ensuring alignment with real-world patterns.

The trial emphasizes transparency and trustworthiness, ensuring that predictions are interpretable and actionable. By highlighting the most influential factors contributing to churn, the trial supports evidence-based decision-making. Expected outcomes include a comprehensive performance evaluation, a detailed description of customer segments, and recommendations for retention strategies. This effort also represents a significant step toward integrating advanced AI tools into MYT's operational workflows, providing scalable solutions for addressing customer churn. The data for this trial are customer oriented and private. Details are described in what follows:

- Contents: Customer TaxID, Supply Meter No., Customer Churn, Payment delays, Amount of payment, Postal code, Energy amount, Customer Complaints From emails in text form.
- Format: CSV files.
- Data Sources: Stored in MYT's corporate data warehouse.
- Anonymization: Private data includes customer-oriented segmentation and churn prediction datasets, also anonymized.

2.1.4 Valuation and commercialisation of water data in France

The trial site of the valuation and commercialization of water data use case has been adapted to build on an opportunity which emerged from the on-going EGM deployments. The new trial site is located on the territory of the French Riviera agglomeration community which is situated on the south-Alps region along the Mediterranean coast. Water management in the catchment area of the French Riviera Agglomeration Community (CARF, Figure 2) presents several major challenges. The scarcity of water resources is a central issue. Due to climate variability and population pressure, freshwater resources are becoming increasingly limited. This situation is exacerbated by prolonged periods of drought that reduce available reserves, thus affecting the supply of drinking water for residents and local economic activities.

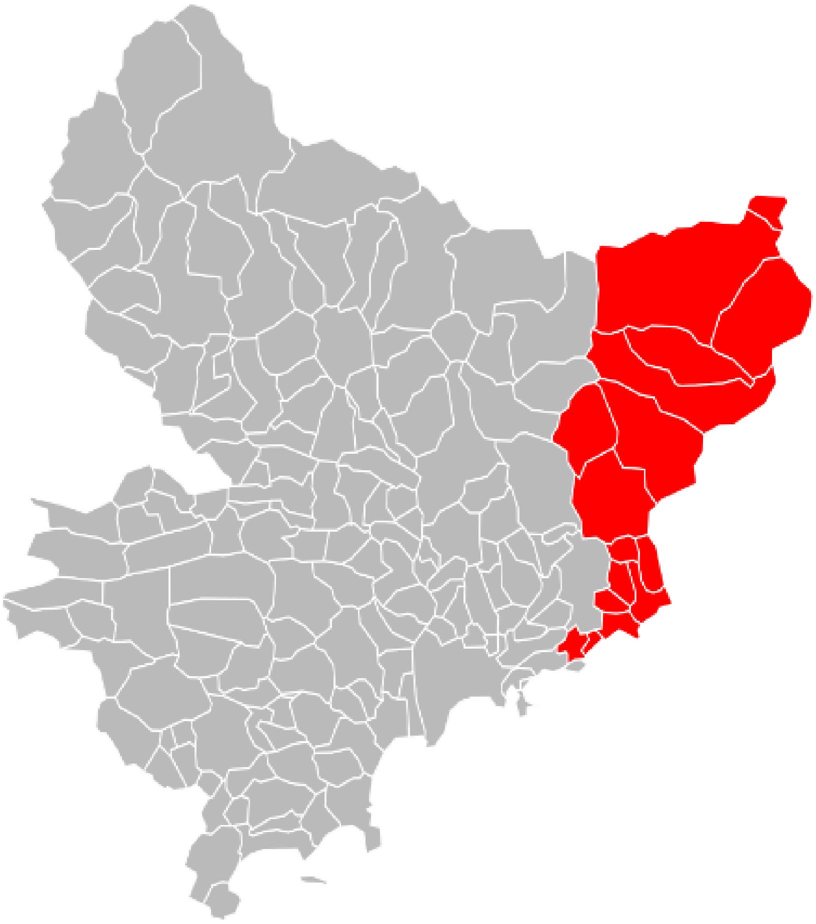


Figure 2: Municipalities of the French Riviera agglomeration community (source: [Wikipedia [5]](#))

Stormwater pollution is another significant challenge. Pollutant-laden runoff from urban and agricultural areas flows into rivers and groundwater, compromising water quality. Pesticides, fertilizers and industrial waste are all contaminants that threaten aquatic ecosystems and public health. In addition, coastal pollution, due to water discharges and maritime activities, degrades beaches and marine habitats, negatively impacting tourism, a key sector for the economy of Menton, a city in the vicinity of Monte Carlo.

The management of the various economic actors, the cross-border nature of water management, the proximity of a national park, add a layer of complexity to water management. Tourism, while vital to the local economy, puts considerable pressure on water resources. Hotels, restaurants and other tourist facilities consume large amounts of water, especially in high season. It is therefore crucial to find a balance between the needs of tourists and the preservation of water resources for permanent residents.

To meet these challenges, an integrated approach to water management is needed. This involves collaboration between local authorities, businesses, farmers and citizens. Measures such as promoting sustainable agricultural practices, improving wastewater treatment infrastructure, and raising public awareness of the importance of water conservation are essential. In addition, the adoption of innovative technologies for water management, such as rainwater harvesting systems and water-saving devices, can help alleviate the pressure on available resources.

Here is a description of the datasets available for water management in the Menton watershed:

- **Ultrasonic river level sensors**: Four sensors will be deployed in January 2025 to measure river levels in real time. These sensors will provide continuous data on changes in river levels upstream of the urban area, allowing accurate and rapid monitoring of hydrological conditions.
- **Connected rain gauge and rain detector**: These devices will be installed upstream to collect data on rainfall. Rain gauges will record the amount of rain that has fallen, while rain sensors will provide information on the intensity and duration of rainfall events.
- **Rainfall radar data**: This data will be collected to provide a detailed view of rainfall over the region. Rainfall radars can track the movements of rainfall systems and estimate the amount of precipitation over large areas.
- **Precision weather data**: This data includes detailed measurements of weather conditions such as temperature, humidity, wind speed, and wind direction. They are essential for understanding the interactions between climate and water resources.
- **Ship transit data**: Collected from AIShub [6] by AIS measurement, this data provides information on the movement of ships in the maritime area at the front. They are important for monitoring the impact of marine activities on coastal water quality. It is proposed to cover a rectangle from Nice to the Italian border (Figure 3) and covering a distance at sea of 6 miles (~10km) corresponding to the coastal navigation area.

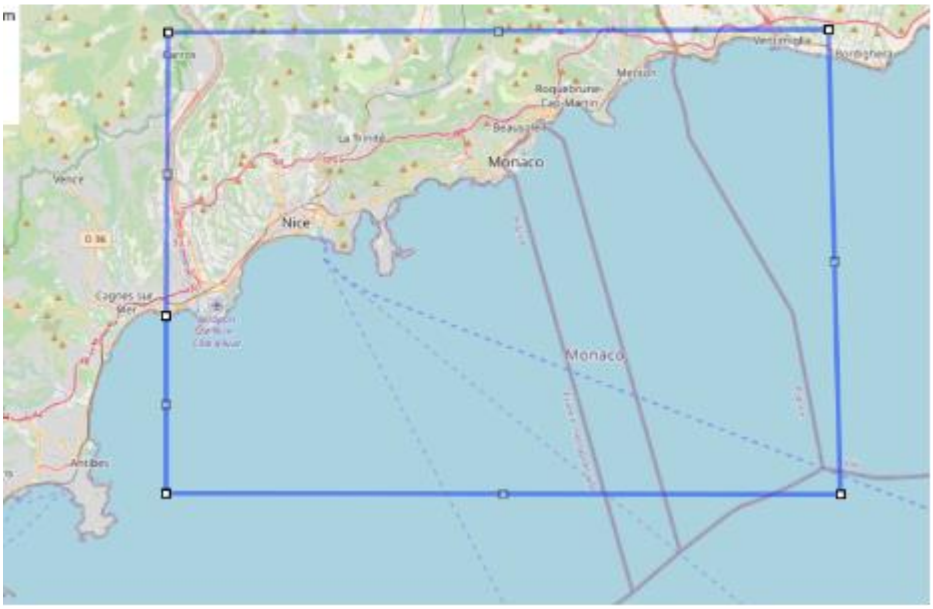


Figure 3: Maritime Traffic collection area (source openstreetmap)

- **Data available in open-data on the website eaufrance.fr [7];** in particular:
  - Sandre is the water data repository of the Water Information System (WIS). It consists of a set of specifications that describe the data as well as the rules and formats of exchanges. In particular, it contains the BD TOPAGE®, which is the new French hydrographic reference system, replacing the BD CARTHAGE® (Figure 4).

- Naiades provides access to quality data for continental surface waters (rivers and water bodies). Physical parameters, concentrations of chemical substances, species inventories and hydromorphology.
- Hydroportal provides water levels and flows at around 5,000 measuring stations.

Figure 4: example of geographical data available from the Sandre website (source Sandre [8]).

These datasets, collected at different frequencies and by various means, are essential for an integrated and efficient management of water resources in the Menton watershed. They will be aggregated at the level of the CARF territorial Hypervision platform so that they can be shared and used. This platform is based on an NGSI-LD (STELLIO) data core and is therefore aligned with the SEDIMARK architecture.

2.1.5 Exploiting datasets for optimising the components of the Data Processing Pipeline

The trial sites provided the consortium with early access to drafts of the datasets of each use case. The goal of this was to use the dataset samples in order to build and optimize components of the data processing pipeline, especially the data quality assessment and the data curation components. The data quality assessment module, as described in Deliverable SEDIMARK_D3.1 [9], aims to provide an evaluation of the quality of datasets based on various criteria. To build a generic module that can work with diverse datasets, the component exploited early versions of all the datasets and used them to test (i) that the component considers many possible input formats, especially with respect to dates and location information and (ii) that the assessment is realistic and according to the expectations of the provider. The first draft of the data quality assessment component was tested, and all datasets and feedback/errors were gathered in order to improve the module and derive the second version of it. Additionally, the datasets were also used to test and debug/improve the data cleaning components, with much emphasis on the outlier detector and the deduplication component. Here, the different nature of the datasets helped optimise the way the component transforms and uses the datasets as input, ensuring that the detection and deduplication algorithms can run smoothly on the datasets. Finally, a draft version of the Energy use case dataset was also used to test both the Local Training and the Shamrock Federated Learning components, helping run experiments and debug/improve the early version of the modules, guiding also the development of the more advanced versions.

2.2. Trial definition and KPIs

In this subsection, the scope of the trials is set and the pilot leaders explicitly define key performance indicators (KPIs) to guide the assessment of the outcomes.

2.2.1 Mobility Digital Twin in Helsinki

While the mobility digital twin is still in the conceptual phase, existing data sources from Helsinki are already available for integration. Exploring datasets such as infrastructure data, traffic data, and air quality data in Helsinki can provide detailed insights into various aspects of urban planning, traffic management, and environmental conditions. The digital twin of Helsinki is formed by a network of interoperable systems, exchanging data over standardised APIs. Below are examples of what these datasets will cover at the Helsinki use case.

- **Traffic data contains traffic counting information from multiple traffic modalities and locations in the Helsinki area**. An example below The City of Helsinki own domain, Lido-Tiku data enhances our project's ability to analyse traffic patterns more comprehensively, ultimately providing a more robust platform for traffic management and analysis in the Helsinki area. API for accessing traffic measurement data of the city of Helsinki. The measurement data is updated daily and hourly and consists of count and speed observations derived from sensor devices (counters) of different types (data sources), across different measurement intervals according to the type of the sensor; typically, 15 minutes. The data is provided on the Swagger API (Figure 5).

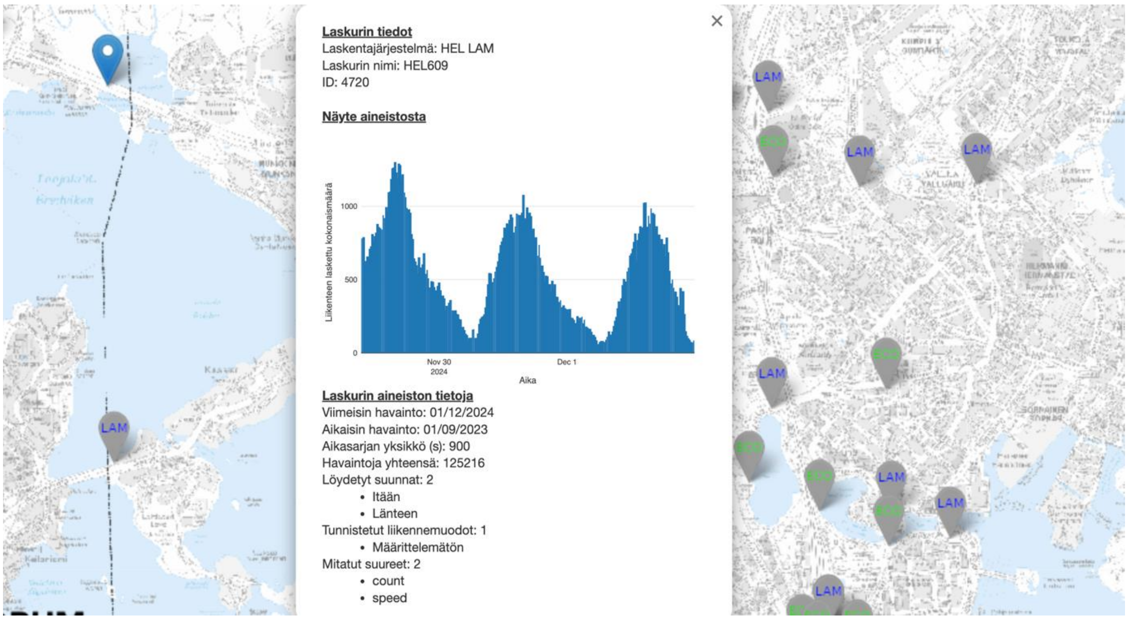


Figure 5: View of data from the Swagger API

- **Datasets from Register of Public Areas**. An example is provided in Figure 6: a visualization of the combined sidewalk, cycle path and separated cycle path. The data is provided as GeoJSON point features, accessible via a WFS API.

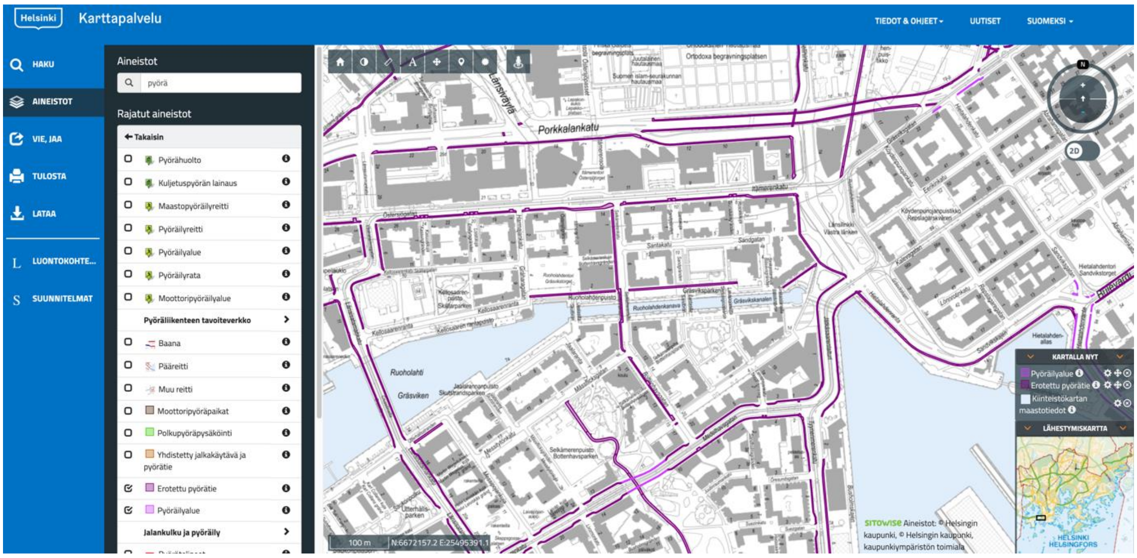


Figure 6: Datasets from Register of Public Areas

- **Real time air quality at HSY monitoring sites**. Data describes the air quality index on 13 measuring sites in Helsinki area, providing an example of geospatial data related to mobility conditions (as a part of mobility digital twin). The data is provided as GeoJSON point features, accessible via a WFS API, as seen in Figure 7:

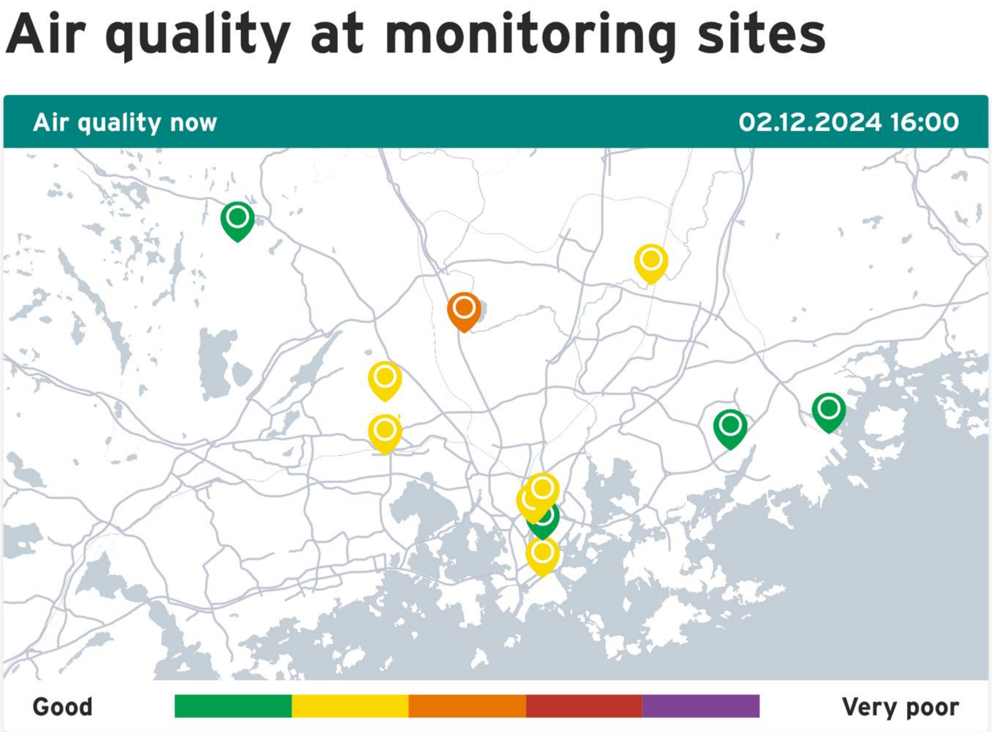


Figure 7: Air Quality at monitoring sites

The system offers data owners an alternative to traditional public procurement by facilitating data purchase and access while enabling effective data publishing and sharing.

Success will be measured By the KPIs described in Table 1.

Table 1: FV KPIs and evaluation

ID	MobDig.01	**Name**	Data from Mobility Digital Twin (DT) to SEDIMARK	**Category**	Data
Description	Data sharing from Mobility Digital Twin to Data Marketplace.				
Rationale	Number of datasets shared from the Mobility Digital Twin to the Data Marketplace.				
Evaluation responsible	FV with associated system representative.				
Evaluator	Expert				

Evaluation process	Expert: SEDIMARK expert validation with the FV/The city of Helsinki support. User: A system user associated with knowledge of the Digital Twin concept with data validation support on SEDIMARK.
Requirements	In both cases, the validation requirements include: • The data can be free and public, or it can be e.g., restricted, exclusive or commercial (traffic counter, air quality). • The metadata describing the availability of the data and licenses must be maintained and may be available from an external API. • The usage of/access to the data may be agreed outside the marketplace or within the marketplace. • The data may be hosted either in the marketplace or in an external service. • When hosted externally, the systems may use the marketplace as a publishing channel, preferably over an API.
Metrics and target	Target: 3 datasets
Rank	Mandatory
Type	Concept

ID	MobDig.02	**Name**	Data from Data Marketplace to Mobility DT	**Category**	Data

Description	Data utilization from the marketplace.
Rationale	Number of datasets obtained from the marketplace and applied in joint visualisation in the Mobility Digital Twin.
Evaluation responsible	FV with associated system representative.
Evaluator	Expert
Evaluation process	Expert: product owner of Digital Twin or a similar role at FV/The city of Helsinki. User: A system user associated with knowledge of the Digital Twin concept and data validation.

Requirements	In both cases, the validation requirements include: • The data can be free and public, or it can be e.g., restricted, exclusive or commercial (traffic counter, air quality). • The metadata describing the availability of the data and licenses must be maintained and may be available from an external API. • The usage of/access to the data may be agreed outside the marketplace or within the marketplace. • The data may be hosted either in the marketplace or in an external service. • When hosted externally, the systems may use the marketplace as a publishing channel, preferably over an API.
Metrics and target	Target: 1 dataset
Rank	Mandatory
Type	Concept

For the use case related to the digital twin of urban mobility, the following SEDIMARK services can be applied:

- Query of available data sets over an API, limited by attributes such as location, tag/classification and timestamp – to be used for retrieving an up-to-date list of datasets available for visualisation in the digital twin environment.
- Query of dataset metadata over an API to obtain license information, data source query URL etc. – to be used for retrieving an individual dataset for visualization in the DT environment.
- Discovery & query of available data sources and/or individual data source parameters from an existing data catalogue metadata API, such as CKAN – to be used for listing data sets to SEDIMARK from urban spatial data infrastructure utilized in the DT.
- KPIs: The system aims to provide data owners with a viable alternative to the traditional public procurement model by offering them a feasible option to purchase and access data. Additionally, the system aims to enable data owners to publish and share their data effectively. To measure the success of these objectives, key performance indicators (KPIs) have been established and defined in the above tables.

Experiment scenarios: This section should describe all actions of the user and the expected system responses for the planned normal execution of the use case. Uploading or linking ensures that the data will be made available for the marketplace, either by uploading it to dedicated hosting, providing a URL to an external hosting (such as the city's open data portal), or by providing a URL to the data stream for a dynamic data source over API. Metadata production involves inputting relevant metadata to the marketplace or providing a link to existing metadata, for example, in an external data-catalogue service. Metadata and/or data is provisioned over the marketplace, aggregating statistics of its use. Data is provided over the UI, as well as over standardised APIs.

The key functional components involved / tested in the use case are:

- Data Types and Accessibility: Data can be public or restricted, with metadata and licenses available through an external API.
- Data Usage and Access: Agreements can be external or within the marketplace.
- Data Hosting Options: Data can reside in the marketplace or on external services.
- Publishing via Marketplace: When hosted externally, systems can use the marketplace as a publishing channel via an API.

Currently, the mobility digital twin in Helsinki is at a conceptual stage, with potential data sources available from Helsinki, including infrastructure, mobility, traffic, and conditions datasets.

Experiment risks: This section should describe any errors that may result during use case execution and how the system will react or respond to those errors. If a dataset becomes expired or its maintenance is discontinued, the dataset should be flagged accordingly and then removed, with the data marked as removed. If the party responsible for the data is dissolved, metadata should be updated, and the data flagged accordingly. Changes to the publicity of the data, such as those arising from the data owner's decisions or changes in legislation must be addressed. For problems in real-time data streams, the system should define the level of real-time management it can handle, and an alternative way to connect to the data stream must be provided. In cases of problems in agreements or commercial arrangements between parties, the system must define its role and level of involvement in inter-party agreements or commercial arrangements. Additionally, there must be a manual way to provide access to data. Privacy breaches and issues with ownership and usage rights must be addressed within the system or through secure interaction with external management. The system should also consider including consent management or MyData features for end-users where appropriate.

Services to be offered: The Use Case may offer city data to external organizations through the SEDIMARK data marketplace and can utilize external data sources to enhance its digital twin of mobility. This highlights Helsinki's pivotal role as both a provider and consumer of valuable data, fostering collaboration and innovation within the marketplace while leveraging external resources to enrich its digital twin capabilities. By doing so, Helsinki not only strengthens the functionality of the marketplace but also demonstrates its commitment to advancing urban mobility solutions. These services benefit businesses, researchers, city planners, and other stakeholders by providing access to a wider range of data sources.

2.2.2 Urban bike mobility planning in Santander

In SEDIMARK_D5.1, it was described how essential it is for Santander City Council to obtain as much information as possible from the use and necessities of the recent setup, consisting of a network of bicycle lanes. Bicycle movement patterns and the use of bicycles and the associated support infrastructure are the main type of information that is needed.

A good number of datasets coming from the already mentioned Santander Smart City Platform (SSCP) are currently available for the project. However, it is the intention of the municipality to not only promote the use of this data but also attract more stakeholders from the municipal ecosystem to provide new datasets or enhance the current ones enriching the total set for the benefit of all citizens.

The KPIs defined in SEDIMARK_D5.1 are oriented to keep track of the success of this aim within the project. The degree of progress is summarized in Table 2.

Table 2: Urban bike mobility planning in Santander KPIs state report

ID	Name	Metric	Target	Achieved	Comments
BikeMob.01	Mobility data from SSCP to Marketplace	No. of elements	4	2	Bicycles + Docking stations
BikeMob.02	Mobility data from Pilot to SSCP	No. of elements	4	5	Provided to the city council – Bike turnover ratio, average waiting time, average trip time, daily/weekly distribution, route heatmap
BikeMob.03	Number of informed decisions	No. of elements	3	0	Pending further project development

It is important to mention that during the KPIs definition (as reported in SEDIMARK_D5.1) several of them were associated to the engagement of volunteer bicycle riders, as the use case was mainly involving this kind of users. However, as with the newly available municipal bicycle sharing service, the amount and diversity of data sources was much higher than the one that would be obtained with private volunteer bikers, that option was discarded and those KPIs are no longer applicable.

The intended outcome of the use case is mainly aimed to satisfy the KPI identified as BileMob.03, which is a measure of how the use case is able to provide valuable information for making informed decisions. After performing an interview with the main responsible for this area within the municipality, a set of interesting questions to be answered arose. The first questions to be considered are related to details about bike routes and stations performance. These questions can be summarised in the list below:

- Which are the most used bike routes? Do they fit with the available bike lanes?
- In which bike rental stations are there a higher turnover of bicycles?
- What are the emptiest and the fullest bike rental station? i.e., which is the degree of occupancy of each station along the day, the week, etc.

In Figure 8 a conceptual design of a report is depicted. This kind of reports should be automatically done (on several frequencies basis) for service's responsible to quickly have a good overview of potential flaws.

For example, the comparison between currently built bike lanes and actual bicycle rider routes will provide necessities of improvements in the form of bicycle lane changes and/or new additions. Further inquiries should be conducted to find out the cause of the differences, but the system will provide valuable hints. Something similar would apply when the bicycle station turnovers are provided in a specific report. Places with a need of a greater density of bicycle stations (or bigger ones) will be easily identifiable. Besides, the degree of occupancy of bicycle stations will complement this information by providing the pattern of use, allowing to set up not only infrastructure modifications but also logistic ones, transferring bicycles from one station to another.

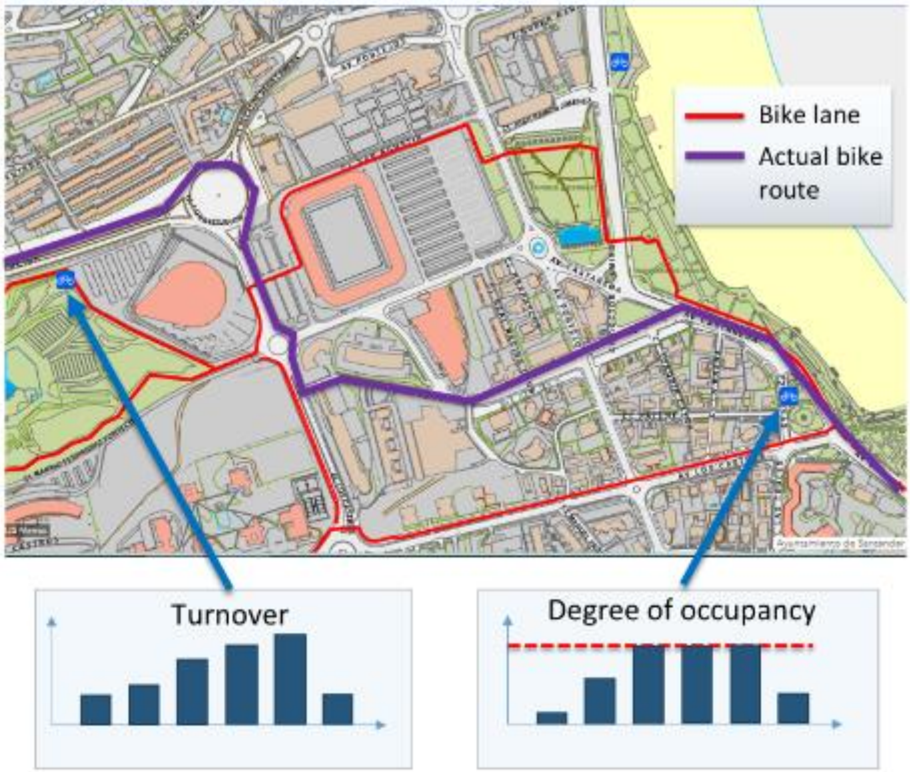


Figure 8: E-bike service bike routes and station turnover and degree of occupancy sample dashboard

Regarding the service to citizenship and visitor's performance, the set of questions to be answered are typically the following:

- What are the average times of use of a rented bicycle?
- What is the probability distribution of time that, at a station, a person must wait to rent a bicycle because there are no bicycles available?
- What are the geographic points where a greater number of bicycle stops are observed?

Figure 9 shows, for illustrative purpose only, a sample of a dashboard addressed to provide useful information for answering these questions. Average time of bicycle use and waiting time to rent a bike in each rental station will allow to adjust bicycle rental rates and other conditions of use. Additionally, a city heatmap counting statistics of bicycle stop and staying (taking advantage of the mounted GPS) will provide rich information about points of interest not only for the municipality mobility service but also for other areas like tourism, shopping, etc.

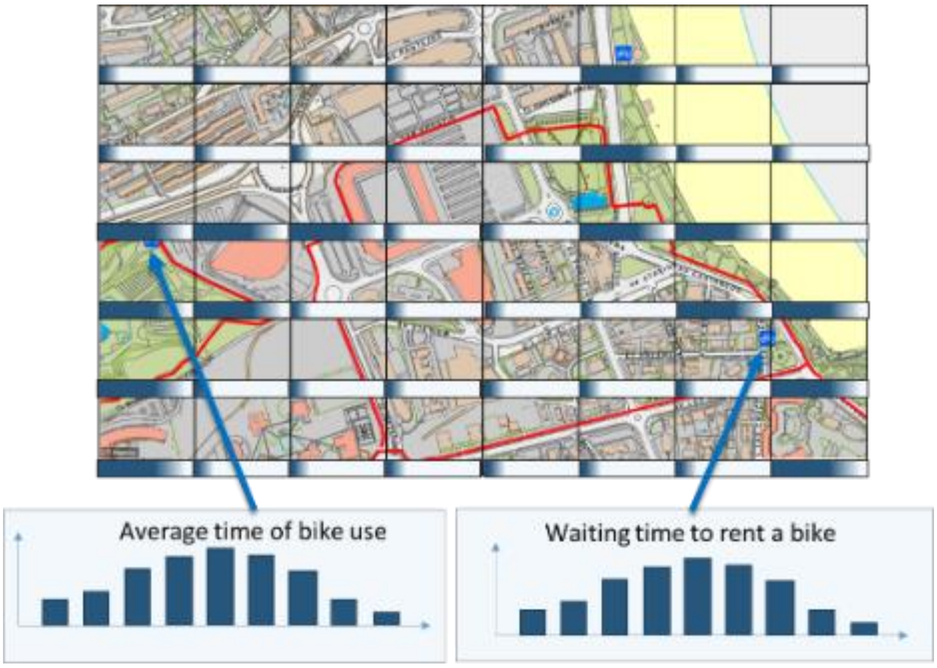


Figure 9: E-bike stops and staying heatmap, average bike use time and waiting time to rent a bike sample dashboard

Finally, the practical use of this information will give rise to informed decisions based on data which can take the form of planning and prioritization of new lanes and traffic calming retrofits, resizing of bicycle loan stations and/or creation of new stations and re-structuring of fees for bicycle use. Figure 10 shows a very simple example of this approach. A mismatch, as shown in Figure 8, between an existing bicycle lane and the actual route taken by bike riders will eventually prompt the creation of a new bicycle lane, potentially utilizing the space from the unused lane.

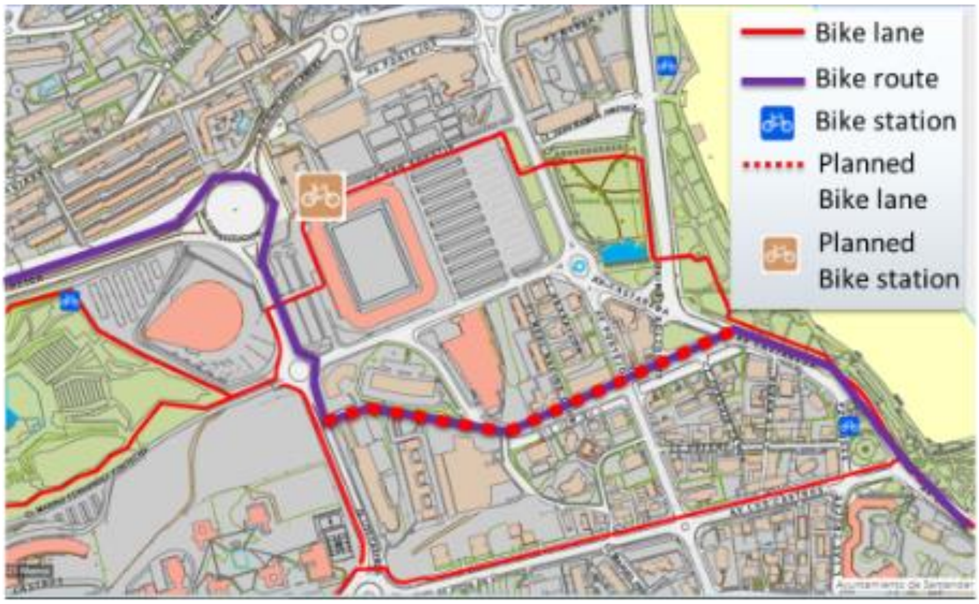


Figure 10: Example of remodelling plan of bicycle lanes infrastructure: new bicycle rental station and setup of a new bicycle lane

2.2.3 Valorisation of energy consumption and customer reactions/complaints in Greece

MYT's main performance indicators study how well AI models predict energy use, group user behaviours and examine customer leave patterns. The goal is to reduce mistakes in predictions, aiming for a Mean Absolute Percentage Error (MAPE) under 5%. Clustering success is checked with measures like the silhouette score and the Davies-Bouldin Index [10], which look at the quality and separateness of user groups found. High silhouette scores and low Davies-Bouldin Index numbers show useful and clearly divided clusters. The experimentation involves two primary data scenarios: energy-oriented clustering using public, anonymized datasets and customer-oriented churn prediction leveraging private, anonymised data. These models utilize decision trees for supervised learning and K-means for unsupervised learning, focusing on customer segmentation by region and churn probabilities at the postal code level. Risks include data privacy concerns, biases in data and algorithms, and challenges in model deployment.

The testing includes two key data situations: energy-based grouping using public, unnamed datasets and customer-focused churn forecasting using private, unnamed data. Supervised learning models, like decision trees, help with predictions, while unsupervised learning methods, such as K-means, aid in grouping customers. These models sort customers by area and predict the chances of churn at the postal code level. Possible risks include worries about data privacy, biases in datasets and algorithms and difficulties in using the models properly.

The combination of clustering quality metrics and predictive performance ensures robust evaluation and meaningful segmentation for operational use in MYT's customer management and energy optimization efforts. Additional KPIs at this point could include:

- **Churn Reduction Rate**:
  - Definition: The percentage of customers who are retained after receiving proactive intervention based on consumption or complaints.
  - Metric: Churn rate percentage before and after the implementation of predictive models and complaint resolution strategies.
  - Target: A 10-15% reduction in churn within the first 6 months post-implementation.
- **Customer Satisfaction Improvement:**
  - Definition: Change in customer satisfaction levels before and after the trial.
  - Metric: Customer satisfaction score (e.g., Net Promoter Score, NPS).
  - Target: A 5-point increase in customer satisfaction score within the pilot region.
- **Energy Consumption Clustering Accuracy:**
  - Definition: The accuracy of clustering energy consumption profiles based on customer usage patterns.
  - Metric: Percentage of correctly classified customers into energy usage clusters.
  - Target: 90% accuracy in clustering energy usage profiles.
- **Predictive Model Reliability:**
  - Definition: The consistency and reliability of AI models in predicting energy consumption and customer behaviour.
  - Metric: Precision, recall, and F1 score for models used in prediction and segmentation tasks.
  - Target: The KPIs identified in SEDIMARK_D5.1 are summarized in Table 3:

Table 3: MYT KPIs and evaluation

ID	Energy.01	**Name**	Data sharing and validation	**Category**	Valorisation of energy consumption and customer reactions/ complaints
Description	Data Sharing to the Marketplace & Model Validation.				
Rationale	Data accuracy and algorithm evaluation.				
Evaluation responsible	MYT				
Evaluator	Expert				

Evaluation process	Expert: • Prediction Accuracy KPI: Mean Absolute Percentage Error (MAPE) between the model's predictions and the actual values for electrical energy consumption (<5%). • Clustering Quality: Silhouette Score in combination with Davies-Bouldin Index - Evaluate the quality of clustering by calculating an average silhouette score, which measures how well-separated the clusters are and whether data points belong to the correct clusters. Then, measure the average similarity between each cluster and its most similar cluster respectively. • Target: Aim for high silhouette scores and low Davies-Bouldin Index values, indicating well-separated and distinct clusters. User: • Prediction Accuracy: Check the accuracy of the energy consumption predictions for the users' own energy usage. Are the predicted values reasonably close to their actual consumption? Higher prediction accuracy indicates a more successful model. • Clustering Relevance: Users apply their intuition to evaluate the clustering results to see if they make sense for their energy consumption behaviour. Do the clusters align with different usage patterns or user segments that they can identify with? Relevant and meaningful clusters indicate a successful clustering process. • Real-world Impact: Reflect on whether the energy-saving strategies suggested by the system are practical and effective in real-life scenarios. Has the system helped the user reduce energy consumption and lower utility bills? • Adaptability: Evaluate how well the system adapts to changes in the users' energy consumption behaviour over time. A successful process should continue to provide relevant predictions and recommendations as their habits evolve.
Requirements	Data cleaning and anonymisation.
Metrics and target	MAPE <5%, high silhouette score & Low Davies - Bouldin Index.
Rank	Mandatory

2.2.4 Valuation and commercialisation of water data in France

The KPIs identified in SEDIMARK_D5.1 are summarized in Table 4. Targets have been added based on the refinement of the pilot site definition:

Table 4: EGM Targets

ID	Name	Metric	Target	Achieved	Comments
Water.01	Number of data providers	Count	5		See section above
Water.02	Number of data consumers	Count	4		CARF CIO, Gemapi, Irrigation management, Veolia
Water.03	Number of AI Algorithms deployed	Count	2+		Rain meter, water flow prediction
Water.04	Number of datasets in the catalogue	Count	5+		Collected data information
Water.05	Number of far edge devices deployed	Count	6		4*water level 1*rainmeter 1* AIS Receiver
Water.06	Number of open datasets integrated	Count	3+		See section above

The business indicators for effective water management are being discussed with the operators in the field. They can be used to monitor and evaluate water management practices, identify areas for improvement and implement sustainable strategies for the preservation of water resources:

- **Water Quality Index (WQI):** Measures the overall quality of water by analysing various physical, chemical, and biological parameters. It helps to identify health risks and evaluate the effectiveness of water treatment processes.
- **Water recycling rate**: Assesses the percentage of water treated and reused for non-potable uses, such as irrigation or industrial processes. This KPI shows the effectiveness of infrastructure in promoting sustainable water management.
- **Water Leakage Rate**: Monitors the volume of water lost due to leaks or faulty infrastructure. Identifying and addressing these losses reduces water waste.
- **Water Consumption Per Unit of Production**: Quantifies the amount of water consumed in relation to production, helping to optimize water consumption efficiency.
- **Water Efficiency Index**: Measures the overall efficiency of water use by taking into account water intensity, recycling rate, and leakage rate.
- **Water Risk Assessment Score**: Assesses potential water-related risks, such as water scarcity, regulatory changes, or water quality issues. This helps mitigate these risks and ensure the sustainability of water resources.

- **Water Conservation Ratio**: Calculates the percentage of water saved through conservation activities, such as the use of water-saving technologies or behavioural changes.

3 Demonstrators integration

The scope of this section is to provide a detailed overview of the current landscape within the provider's domain, including an assessment of the existing systems slated for integration. It highlights the specific functionalities of these systems, their relevance to the pilot objectives, and the technical and operational adjustments required for seamless integration. Furthermore, it explains how these systems will interact with the SEDIMARK toolbox, outlining the mechanisms of interoperability, data exchange, and alignment with project goals, ensuring a cohesive and effective integration process.

3.1 Mobility Digital Twin in Helsinki

3.1.1 Systems to be integrated

The City's digital twin for mobility describes the traffic itself, traffic environment and related conditions and context. Digital twin includes various data sources. It is not one system, but instead a constantly evolving combination of diverse data sources and information, conceptualized as a system of systems. It can be applied for diverse use cases. Many parties contribute to producing and updating the data in the digital twin. The data can be studied, refined and combined with different digital tools, creating added value to its users. Mobility Lab facilitates the development and use of the digital twin for mobility and showcases its concrete applications in the field of mobility.

LIDO-TIKU platform: The LIDO-TIKU platform is an initiative by the City of Helsinki to develop a real-time traffic monitoring and data service. This system provides a comprehensive overview of traffic conditions, including real-time statistics and monitoring images, to support traffic planning, research, and management. Data is collected from various sources, such as traffic counters, and the platform makes this information accessible to users through machine-readable APIs, enabling better decision-making and improved urban mobility management.

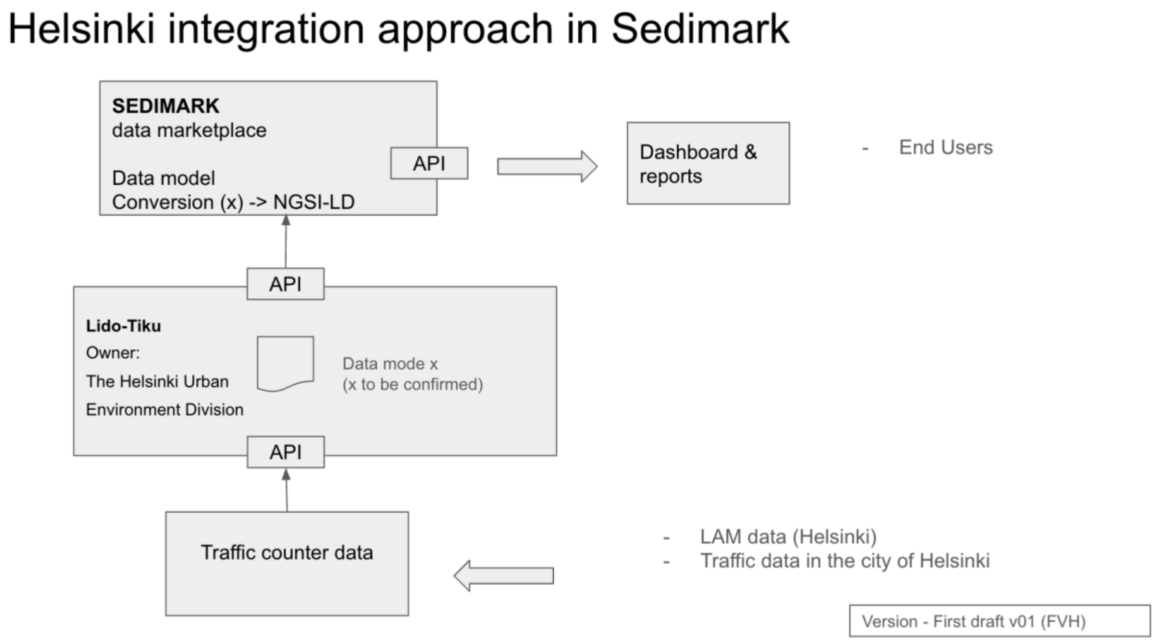


Figure 11: Helsinki integration approach in SEDIMARK

LIDO-TIKU is part of Helsinki's broader strategy for smart transportation, focusing on reducing emissions, enhancing safety, and increasing the efficiency of urban mobility systems. It aligns with the city's goals of leveraging modern technologies to ensure smoother and more sustainable urban operations.

The Helsinki SEDIMARK integration (Figure 11) emphasizes the use of traffic counters, such as induction loops. Measurement data is updated daily and hourly and includes vehicle counts and speed observations derived from various types of sensors (data sources). These measurements are recorded across different intervals, typically every 15 minutes, depending on the sensor type. An API for accessing the City of Helsinki's traffic measurement data is currently under development, and occasional data quality issues may arise during this phase. Swagger documentation can be found here [11].

Helsinki's use case for SEDIMARK focuses on utilizing data from traffic counters to gain detailed insights into traffic flow. These counters measure vehicle counts, speeds, and other parameters, enabling urban planners and researchers to analyse patterns, and plan infrastructure improvements to enhance the efficiency and safety of the city's mobility system.

Currently, the LIDO-TIKU strategy does not support contract handling features; therefore, the data purchase function is managed manually at the city office within the economy division, considering data purchase laws. The SEDIMARK toolbox can take advantage of the separate development domain at FVH to meet the project's needs, such as processing data models.

HSY Air Quality Platform: The HSY Air Quality Platform, as seen in Figure 12, is an initiative by the Helsinki Region Environmental Services Authority (HSY) that monitors and provides real-time data on air quality in the Helsinki metropolitan area. Helsinki Region Environmental

Services (HSY) offers real-time air quality data in a machine-readable format (API). Additionally, open data in a geographic information format is accessible through the open data map service. SEDIMARK integration follows a similar approach to the LIDO-TIKU platform.

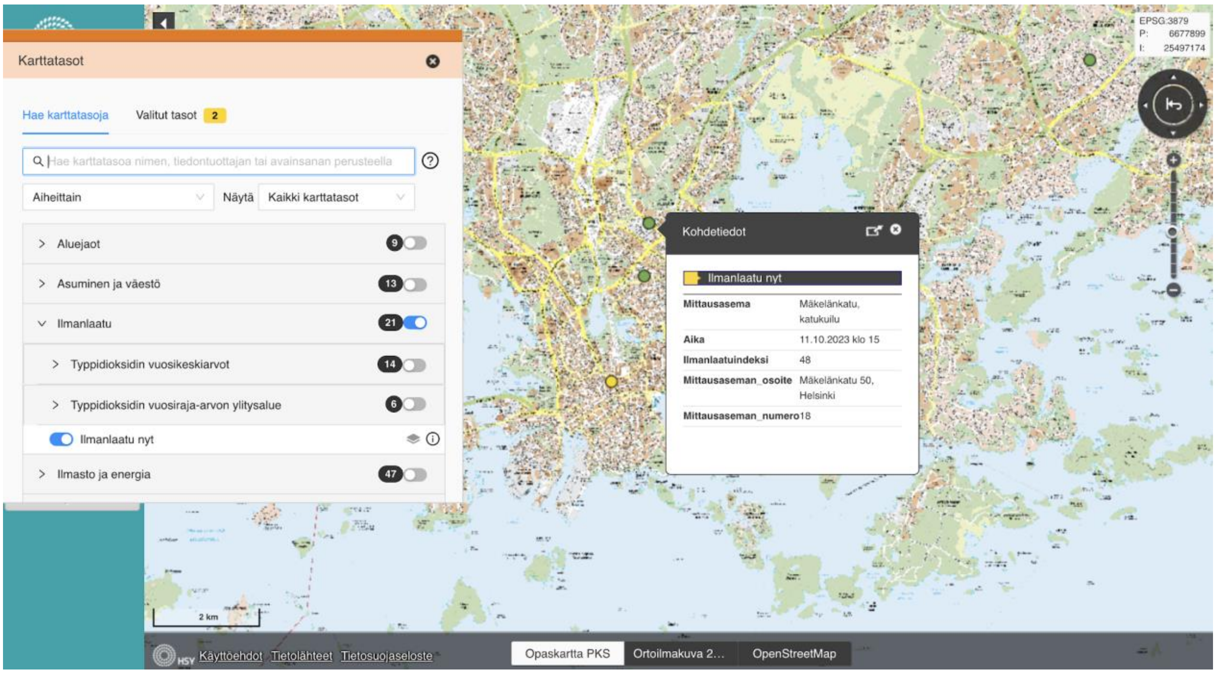


Figure 12: HSY air quality platform

The HSY Air Quality Platform, provides real-time air quality data for the Helsinki metropolitan area, serving a diverse range of users.

Air Quality datasets are GeoJSON point features available from a WFS API (Query URL:https://kartta.hsy.fi/geoserver/wfs?request=GetFeature&outputFormat=json&typeNames=ilmanlaatu:Ilmanlaatu_nyt&srsname=EPSG:4326

Register of Public Areas: Datasets from Register of Public Areas in the form of GeoJSON point features, accessible via a WFS API. GeoJSON features available from a WFS API (Query URL:
https://kartta.hel.fi/ws/geoserver/avoindata/wfs?request=GetFeature&typeName=avoindata:YLRE_Katu_ja_viherosat_ajorata_alue&outputFormat=json&srsname=EPSG:4326)
More information about the datasets is available in the tables provided in SEDIMARK_D1.3.

3.1.2 Interaction with the SEDIMARK toolbox

The toolbox deployment process integrates seamlessly with the necessary infrastructure to enable efficient context data management and exchange, employing a system-of-systems Digital Twin approach with the advantage of using the FVH system as the central repository for data model processing. This setup demonstrates standardized formats and supports updates and queries while smoothly interacting with domain-specific data, such as traffic, infrastructure, and condition data. By utilizing standardized APIs, the Digital Twin achieves interoperability across various technologies. Figure 13 presents the integration approach of the toolbox deployment concept and the owners of domains that will be leveraged in the Helsinki use case to support data sharing. The demonstration can also focus more on the selected domain in case of limited time for deployment.

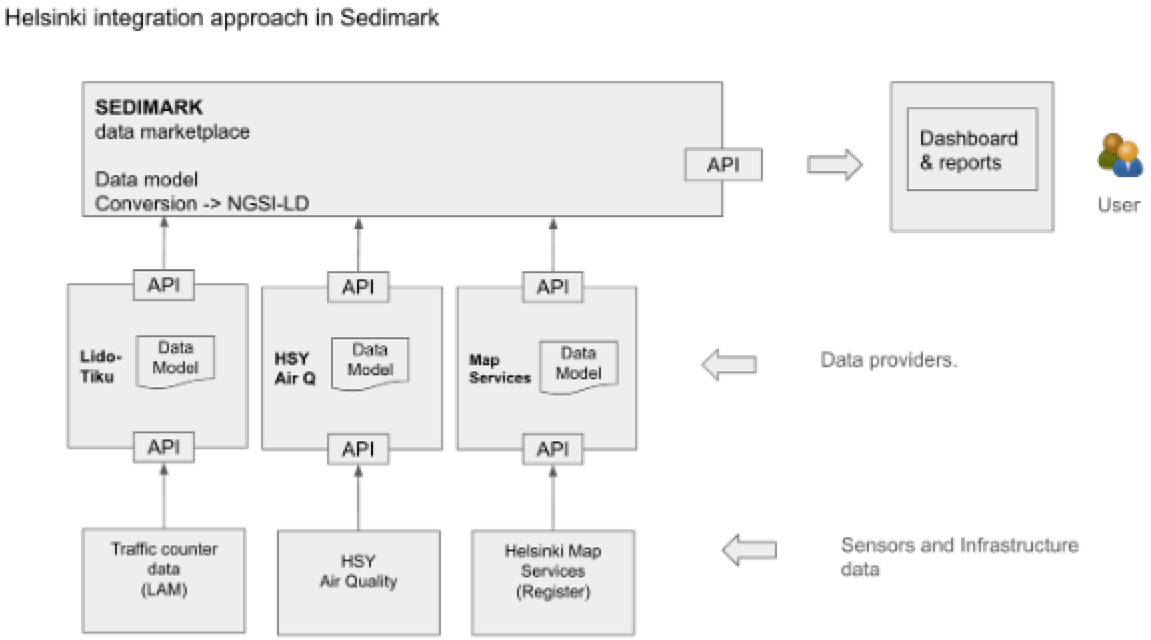


Figure 13: Helsinki interaction with the SEDIMARK toolbox

3.2 Urban bike mobility planning in Santander

3.2.1 Systems to be integrated

The integration of the Marketplace leverages the SSCP platform, which is basically a centralized data repository, hosted by Santander's Municipality, where all the information on the operation of municipal services is collected and from which this information is served to where it is deemed necessary.

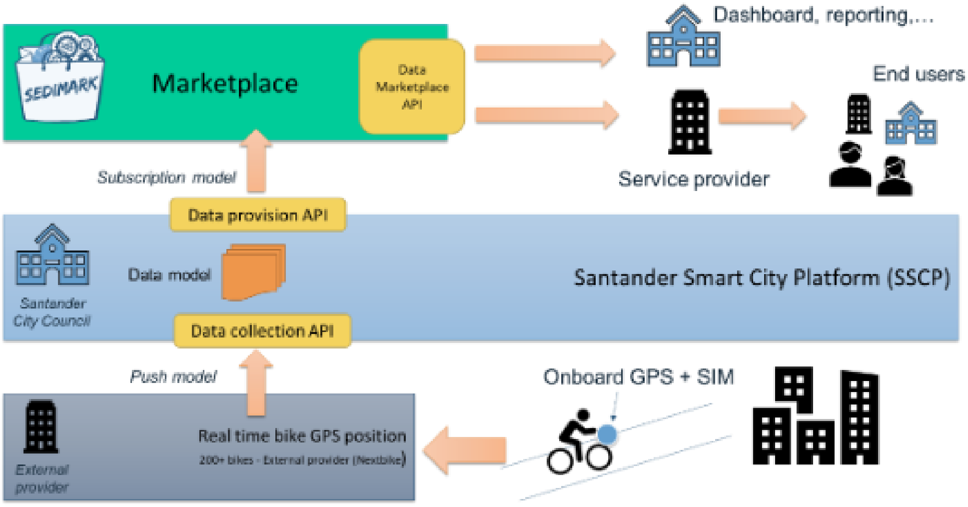


Figure 14: Santander use case integration approach

Figure 14 shows the high-level structure of the use case and how its elements are integrated. It is composed of three elements:

- Data source for the use case: located in the lower band of Figure 14, it consists of the system that provides information on bicycles, rental stations, bicycle counters on bicycle lanes, etc.
- SSCP: located in the central band of Figure 14, it is the platform that centralizes the data and has both the capacity to collect data and to serve it.
- SEDIMARK Marketplace: located in the upper band of Figure 14, it is the central element of the SEDIMARK project.

The information of the elements that make up the use case is managed by an external provider that uses the information for the bicycle rental service itself. Through an agreement established in the requirements of the Call for Procurement that gave rise to the contract, the external supplier, through a push approach, sends the information to the SSCP which, for this purpose, has an API for data collection.

This information is stored in the SSCP, which uses an NGSI v2 data model. Once this information has been collected, it is made available to whoever needs it (applying the protocols, permissions and conditions established by the municipality).

For this purpose, the SSCP has its corresponding data provision API based on a subscription model. The Toolbox that any participant of the SEDIMARK Marketplace must have, has access to the information needed and makes the corresponding data model conversion from NGSI v2 to NGSI-LD. For these two processes, information gathering and data mapping, the necessary modules have been developed and integrated within the Toolbox.

Once the information is in the Marketplace, the data is provided to its users through the Data Marketplace API. In the use case within this project, a set of dashboards and reports have been designed and are currently being implemented for the municipality services in charge of the urban mobility. Notwithstanding, the information is also available for the use of other stakeholders than can provide services to end users, including the municipality, and therefore closing a virtuous circle with the objective that, once a critical mass has been achieved, feedback will enrich the ecosystem and continue attracting more stakeholders.

3.2.2 Interaction with the SEDIMARK Toolbox.

The use case interacts seamlessly with the Toolbox to support various stages of data processing and exchange.

Firstly, as it has been already mentioned, to ensure interoperability and adherence to standardized formats, the use case already builds on a Scorpio NGSI-LD context broker which will be the provider's domain data storage. The interaction with the Connector at the Toolbox for the retrieval of the data upon successfully reached the necessary agreement with the consumers will be supported seamlessly as this will be one of the storage technologies that will be supported by the Connector's implementation that is being carried out in the project.

Furthermore, the data processing pipeline plays a central role, handling tasks such as data cleaning, normalization, and anomaly detection in the bike mobility datasets that will be shared through the Marketplace.

Finally, there is no plan, for the moment, to leverage the AI pipeline for the development of models from the datasets that will be employed in the use case.

3.3 Valorisation of energy consumption and customer reactions / complaints in Greece

3.3.1 Systems to be integrated

MYT's system architecture integrates several critical components to ensure the effective deployment and operation of AI-driven solutions. At the core of the system is MageAI (SEDIMARK_D5.3 [4]), an advanced AI orchestrator responsible for overseeing the training and testing of machine learning (ML) models. MageAI provides capabilities for distributed training, federated learning, and real-time model inference, enabling robust and scalable AI workflows. It integrates seamlessly with local energy datasets, ensuring localized predictions that reflect the unique characteristics of different regions. Alongside MageAI, MLFlow [12] is employed to automate the machine learning pipeline, facilitating efficient tracking of training metrics, model selection, and version management to maintain consistent performance and reliability.

A comprehensive data processing component supports these AI systems by curating and formatting energy datasets. This module is designed to preprocess and analyse data, incorporating tools for anomaly detection, feature engineering, and data cleaning. These processes ensure that the datasets are well-prepared for AI training and analysis, yielding accurate predictions and actionable insights. By leveraging robust preprocessing techniques, the system can effectively manage diverse data sources, including residential energy consumption records, customer feedback, geospatial information, and weather data. These datasets form the backbone of the AI models, supporting tasks such as energy consumption prediction, customer clustering, and churn analysis.

The platform also includes a marketplace interface where participants, acting as consumers or providers, interact with the system. This interface connects to backend components to offer a seamless user experience. It facilitates dataset selection, model sharing, and visualization of predictions, ensuring transparency and accessibility for end-users. Participants can access detailed insights into energy consumption patterns, customer behaviours, and model outputs, empowering them to make data-driven decisions.

The integration of these systems supports the broader objectives outlined in the valorisation efforts. MageAI orchestrates AI models that predict energy consumption with a MAPE target of less than 5%, identify well-separated user clusters using metrics such as the silhouette score and Davies-Bouldin Index, and predict customer churn with precision and recall rates exceeding 85%. The data processing pipeline ensures the reliability and accuracy of input data, while the marketplace interface enhances user engagement by providing intuitive tools for interacting with AI outputs.

The goal of this integrated approach is the delivery of scalable, reliable, and transparent solutions for energy optimization and customer management. The system not only addresses technical challenges such as data privacy and model biases but also aligns with strategic KPIs, including reducing churn by 10-15%, improving customer satisfaction scores by 5 points, and achieving 90% accuracy in clustering energy usage profiles.

3.3.2 Interaction with the SEDIMARK toolbox

The use case interacts seamlessly with the toolbox to support various stages of data processing, model development, and deployment within the energy sector. A general overview can be seen in Figure 15. The data processing pipeline plays a central role, handling tasks such as data cleaning, normalization, and anomaly detection in energy consumption datasets. To ensure interoperability and adherence to standardized formats, semantic enrichment tools within the toolbox standardize datasets using frameworks like NGSI-LD, enabling consistent representation and smooth integration with external systems.

The toolbox integrates advanced tools such as MageAI and MLFlow to automate the creation, tracking, and management of machine learning pipelines, particularly for energy consumption forecasting. These tools streamline the development of AI workflows, enabling efficient pipeline orchestration while ensuring reliability in predictive models. Federated learning workflows are also facilitated, leveraging the toolbox's utilities for weight aggregation, participant coordination, and the logging of training metadata, which are essential for maintaining performance and consistency across distributed systems.

The results of model predictions and workflow monitoring are exposed to end-users through the toolbox's intuitive user interface. Energy demand forecasting models are evaluated using predefined metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). These results are logged and visualized through the Marketplace GUI, providing end-users with actionable insights into energy consumption trends, forecast accuracy, and model performance.

To ensure secure and trustworthy interactions within the ecosystem, the toolbox manages secure data exchange workflows. It incorporates the Trust Enabler for participant authentication and the DLT Enabler for transaction traceability, safeguarding data integrity and fostering trust among stakeholders. When AI models or datasets are shared within the SEDIMARK Marketplace, tools for tokenization and metadata annotation are employed, ensuring that transactions are both secure and transparent.

By integrating preprocessing, Federated Learning, and secure data exchange capabilities, the Toolbox empowers stakeholders to efficiently manage energy data and leverage advanced AI models. This interaction aligns with the broader objectives of improving energy forecasting accuracy, enhancing model transparency, and ensuring interoperability within the energy management ecosystem. Through its comprehensive set of tools and secure workflows, the toolbox supports the seamless adoption of AI-driven solutions for energy optimization and stakeholder engagement.

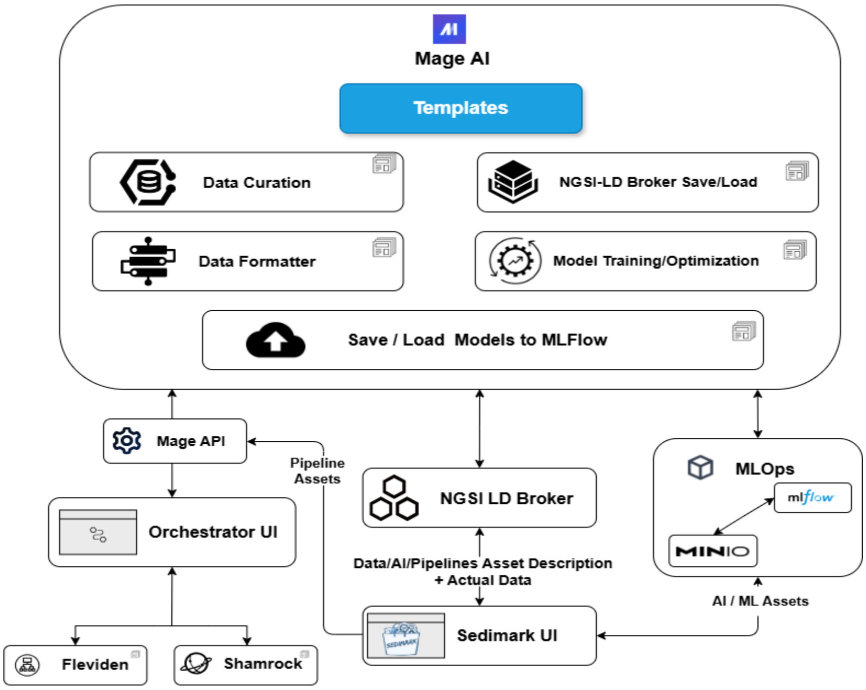


Figure 15: Orchestrator architecture

3.4 Valuation and commercialisation of water data in France

3.4.1 Systems to be integrated

The integration of the SEDIMARK ecosystem within the Twin·Picks platform from EGM is given in the Figure 16. The different data streams are retrieved by a NiFi connector which will ensure the transformation into NGSI-LD before storage in a Stellio data core. Two interface components will allow datasets to be exposed as open-data while the connector will allow transactions to be set up with consumers wishing to exploit the datasets.

The main part of the integration will focus on the implementation of a Data/AI processing chain connected to the core data.

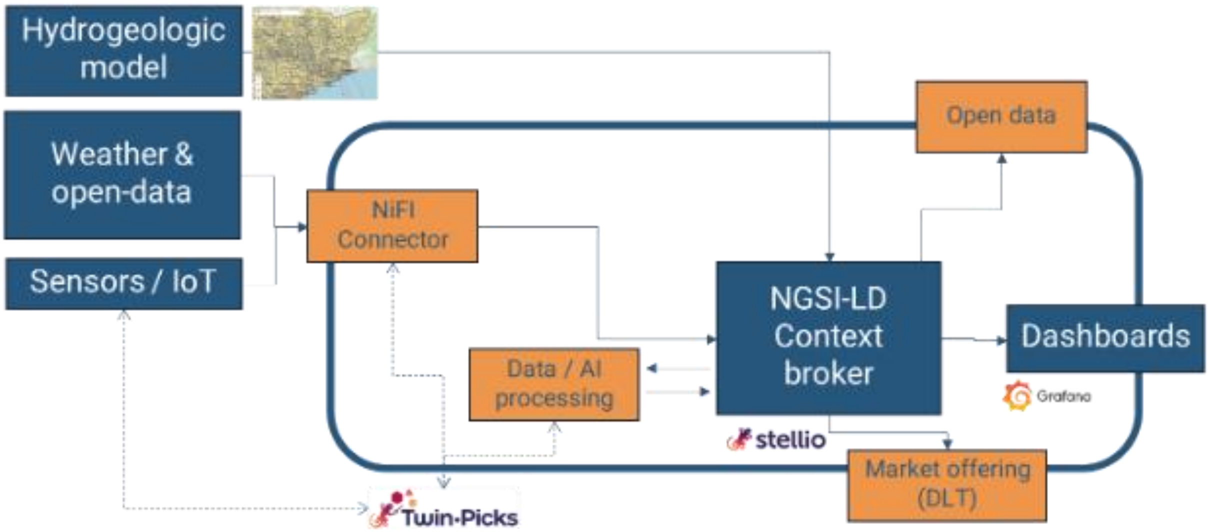


Figure 16: Integration of SEDIMARK modules within the EGM Twin·Picks platform.

A Twin·Picks is now available in production in the country. It already allows the identification of datasets by DCAT annotation of datasets (Figure 17). An experimentation instance with duplication of relevant datasets will be deployed for the project.

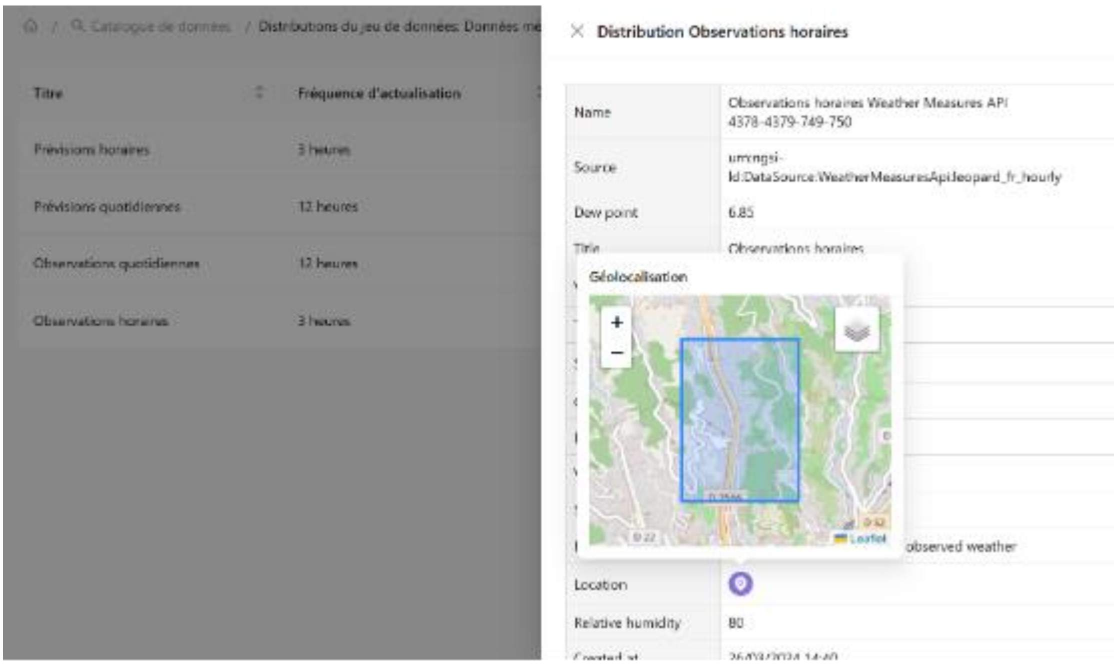


Figure 17: Export of NGSI-LD datasets by DCAT annotation

3.4.2 Interaction with the SEDIMARK toolbox

Twin·Picks already builds on a STELLIO NGSI-LD context broker which will be the natural point of data storage. In respect with the dataflow, the platform builds on NiFi which in its recent release better embed python processors. It is unsure yet if the AI/data processing pipeline orchestrator will be adapted to fit on NiFi or if the Twin·Picks platform should be evolved to integrate the Mage.AI based orchestration. The open-data publication is going to be handled by the addition of a CKAN module. Finally, the integration of the dataspace connector will be analysed with the intent to ease publication directly from the Twin·Picks data catalogue.

4 Testing and assessment of system performance

This section outlines the key performance indicators (KPIs) achieved during the initial trials, providing a clear assessment of how the SEDIMARK system aligns with project objectives. It discusses the early results, highlighting the system's performance, identifying strengths, and pinpointing areas for improvement. Lessons learned from these trials are analysed to refine methodologies and address challenges, forming a foundation for future assessments. Additionally, the section sets the stage for upcoming steps, including further testing, optimization, and scaling efforts to enhance the system's overall impact and effectiveness.

4.1 Mobility Digital Twin in Helsinki

4.1.1 Connection with the SEDIMARK toolbox

The SEDIMARK toolbox ensures interoperability among systems and facilitates the integration of use cases, including the different systems underlying Helsinki's Mobility Digital Twin.

The development of Helsinki's Mobility DT is connected to the ongoing LIDO-TIKU initiative, which in turn aligns with Helsinki's broader smart transportation strategy focusing on reducing emissions, enhancing road safety, and optimizing urban mobility systems for increased efficiency and sustainability.

The in-development LIDO-TIKU platform aims to provide real-time traffic monitoring and data service. This system provides a comprehensive overview of traffic conditions, including real-time data and statistics, and monitoring images, to support traffic planning, research, and management. Data is collected from various sources, such as traffic counters, and the platform makes this information accessible to users through machine-readable APIs, enabling better decision-making and improved urban mobility management.

In addition to LIDO-TIKU, Helsinki's Mobility DT includes information on traffic infrastructure and urban air quality. Currently, these are retrieved from separate systems, over a standard geospatial API. From an application point of view, this implies that interacting with the current data sources of Mobility DT requires connecting with multiple APIs and coping with their data formats.

By leveraging cutting-edge technologies, the toolbox supports seamless data exchange across platforms, ensuring smoother and smarter urban operations, potentially providing a unified access to the Mobility DT. With NGSI-LD standards, it bridges the gap between diverse data sources and applications, enabling urban planners, researchers, and developers to harness data for actionable insights and improved decision-making.

The integration scope for SEDIMARK Toolbox with Helsinki's Mobility Digital Twin is:

- **Data Collection and Integration:**
  - Collect data from available APIs from various systems, such as LIDO-TIKU traffic counters, Air quality monitors (HSY), and the Register of Public Areas in the city of Helsinki.
  - Integrate the data into the NGSI-LD format to ensure compatibility with the defined APIs in Mobility Digital Twin concept.

    o Ensure real-time data ingestion for dynamic traffic data management and urban planning.

- **Data Retrieval:**
    - Use the LIDO-TIKU API or other relevant endpoints to collect real-time traffic data.
    - Additional data that the use case may utilize includes air quality monitors (HSY) and the Register of Public Areas in the city of Helsinki.
- **NGSI-LD Entity Mapping**
    - Convert LIDO-TIKU traffic data into NGSI-LD entities (Figure 18) based on the NGSI-LD data model for compatibility with the Context Broker.

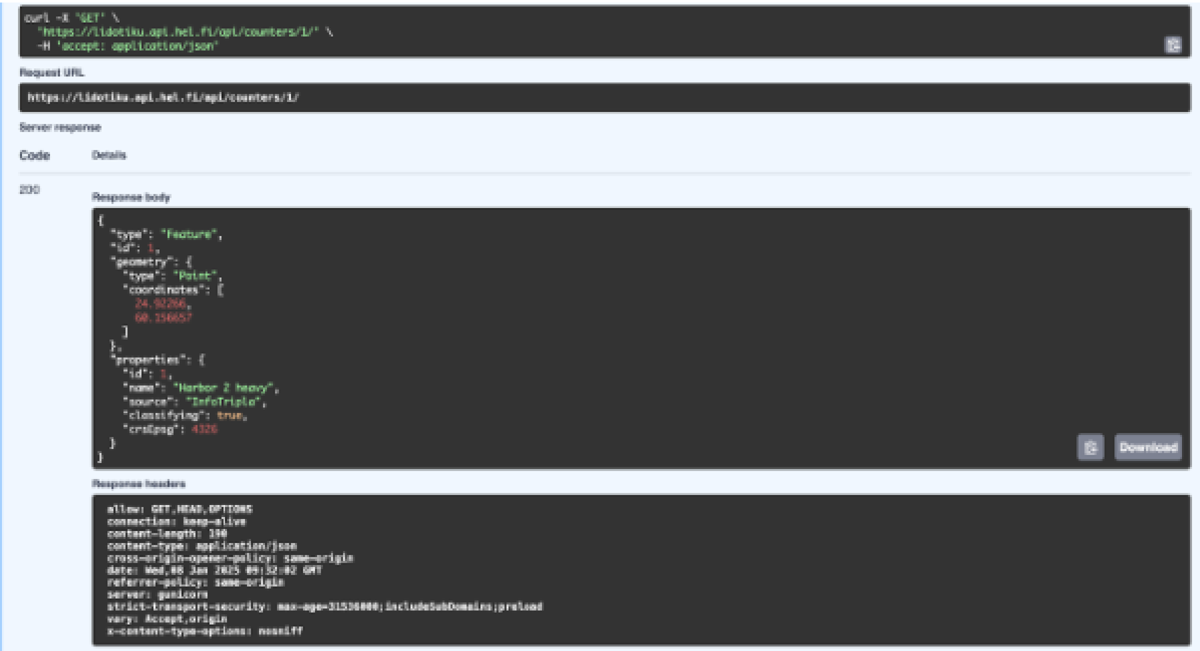


Figure 18: Example NGSI-LD entity for traffic counters

- **Stakeholder Validation and KPI Monitoring:**
    - Collaborate with stakeholders to validate the integrated solutions against the city's mobility and environmental goals.
    - Track specific KPIs to measure the success of the integration, including traffic flow improvements, emission reductions, and resource optimization.
    - Track success of integration by verifying the access to different underlying data sources of the mobility DT via a single NGSI-LD API.
- **Urban Planning and Environmental Benefits:** By providing a unified API, this use case opens opportunities for developing new decision-making tools and dashboards that offer a comprehensive view of urban mobility. The collected data can be used to guide urban planning decisions, focusing on congestion hotspots, pollution levels, and transportation system efficiency. Additionally, data sharing can introduce scenarios that promote sustainable mobility, such as encouraging the adoption of eco-friendly transport options.

This approach contributes to the development of next-generation traffic management services and urban transportation solutions, aligning with both environmental goals and future needs.

- **Ongoing Monitoring & Improvement:** Continuously monitor the effectiveness of the integration through real-time data analysis and feedback from stakeholders with KPIs.

Additional data that the use case may utilize includes air quality monitors (HSY) and the Register of Public Areas in the city of Helsinki.

- **Air quality monitors (HSY)**
  - The HSY Air Quality Platform provides real-time air quality data for the Helsinki metropolitan area, serving a diverse range of users. Air Quality datasets are GeoJSON point features available from a WFS API (Query URL: https://kartta.hsy.fi/geoserver/wfs?request=GetFeature&outputFormat=json&typeNames=ilmanlaatu:Ilmanlaatu_nyt&srsname=EPSG:4326
- **Register of Public Areas**
  - GeoJSON features available from a WFS API (Query URL: https://kartta.hel.fi/ws/geoserver/avoindata/wfs?request=GetFeature&typeName=avoindata:YLRE_Katu_ja_viherosat_ajorata_alue&outputFormat=json&srsname=EPSG:4326)

4.2 Urban bike mobility planning in Santander

4.2.1 Connection with the SEDIMARK toolbox

In order to test and demonstrate the functionality of the implemented systems, a mobility dashboard has been deployed. The dashboard is fed with post-processed NGSI-LD data derived from the mobility datasets (i.e., bicycles and stations) collected at the SSCP. Figure 19 shows an example of a pre-processed bicycle entity.

```json
{
    "id": "urn:ngsild:Vehicle:vehicle:MobilityManagement:196636",
    "type": "Vehicle",
    "category": {
        "type": "Property",
        "value": "tracked"
    },
    "vehicleType": {
        "type": "Property",
        "value": "e-bike"
    },
    "battery": {
        "type": "Property",
        "value": 0.87,
        "observedAt": "2024-09-25T16:23:06Z",
        "unitCode": "P1"
    },
    "currentTripCount": {
        "type": "Property",
        "value": 0,
        "observedAt": "2024-09-26T07:13:10Z"
    },
    "serviceStatus": {
        "type": "Property",
        "value": "parked",
        "observedAt": "2024-09-25T16:29:23Z"
    },
    "location": {
        "type": "GeoProperty",
        "value": {
            "type": "Point",
            "coordinates": [
                43.475682,
                -3.803682
            ]
        },
        "observedAt": "2024-09-25T16:29:23Z"
    },
    "@context": [
        "https://raw.githubusercontent.com/smart-data-models/dataModel.ERA/master/context.jsonld",
        "https://uri.etsi.org/ngsi-ld/v1/ngsi-ld-core-context-v1.7.jsonld"
    ]
}
```


Figure 19: Sample bicycle entity (only an excerpt of the complete NGSI-LD entity is shown)

It is an NGSI-LD entity, transformed from the NGSI-v2 version provided by the SSCP. As can be seen, several useful properties can be leveraged to obtain more meaningful statistics. For instance, the location property allows to build a heatmap when all the bicycles over a reasonable period of time are used. On a similar note, Figure 20 shows a pre-processed bicycle station NGSI-LD entity.

```json
{
    "id": "urn:ngsild:BikeHireDockingStation:bikestation:MobilityManagement:336927189",
    "type": "BikeHireDockingStation",
    "availableBikeNumber": {
        "type": "Property",
        "value": 4,
        "observedAt": "2024-09-26T07:41:46Z"
    },
    "freeSlotNumber": {
        "type": "Property",
        "value": 16,
        "observedAt": "2024-09-26T07:41:46Z"
    },
    "outOfServiceSlotNumber": {
        "type": "Property",
        "value": 0,
        "observedAt": "2024-09-18T08:34:35Z"
    },
    "stationName": {
        "type": "Property",
        "value": "Plaza de Italia"
    },
    "totalSlotNumber": {
        "type": "Property",
        "value": 20
    },
    "location": {
        "type": "GeoProperty",
        "value": {
            "type": "Point",
            "coordinates": [
                43.47177,
                -3.78128
            ]
        }
    },
    "status": {
        "type": "Property",
        "value": "working"
    },
    "@context": [
        "https://raw.githubusercontent.com/smart-data-models/dataModel.Transportation/master/context.jsonld",
        "https://uri.etsi.org/ngsi-ld/v1/ngsi-ld-core-context-v1.7.jsonld"
    ]
}
```


Figure 20: Sample bicycle station entity (only an excerpt of the complete NGSI-LD entity is shown)

In this case, properties such as the free slot number are critical to later calculate the average occupancy of the station.

The dashboard has access to data from 221 bicycles and 28 bike stations, which provide not only periodic updates every 20 minutes, but also additional observations when a meaningful event happens. In the case of a bicycle, this could be a new rental by a user or a change of location. For bicycle stations, every time a bicycle is rented or returned is considered an event. This means that the system receives approximately 30000 updates per day.

After enough data has been collected, the datasets are processed to extract value-added knowledge that may help the Santander City Council policymakers to execute an informed mobility plan. Currently, a total of five metrics are calculated and exposed through the dashboard:

1. Bike turnover ratio at a station is calculated by processing the changes in the available bike number property throughout the desired period of time. In this case, the ratio is averaged per day. This is repeated for every station to observe which ones have more activity.
2. Average waiting time at a station can be calculated by measuring the time when the station has 0 available bikes. This requires combined usage of the available bike number and the “observedAt” sub-property.
3. The average trip time is the total time that bicycles are in use divided by the total number of trips. It is a global metric, which aggregates all bicycles into a single calculation.
4. Daily and weekly distribution depicts the usage of bicycles throughout the day (by hour) or throughout the week (by weekday). It is calculated as a sum of all trips during each time slot, the same way a histogram would be.
5. The route heatmap uses the location of bicycles during every trip, including the starting point and the destination. This aggregated route dataset is used to build the heatmap to illustrate the parts of the city that are used the most by bicycle users.

Figure 21 shows the implemented station dashboard. The first two metrics mentioned above are summarized in a table with the visual aid of an interactive map with the location of every bicycle station.

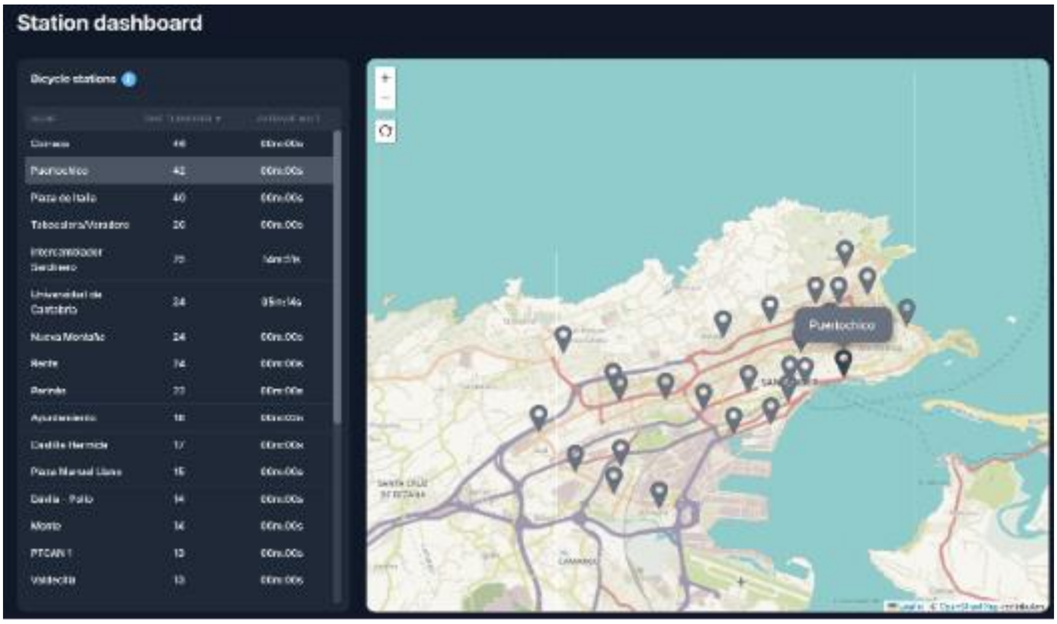


Figure 21: Station dashboard

Figure 22 depicts the bicycle dashboard. Metrics 3 and 4 can be seen, the former as a single value and the latter as a histogram.

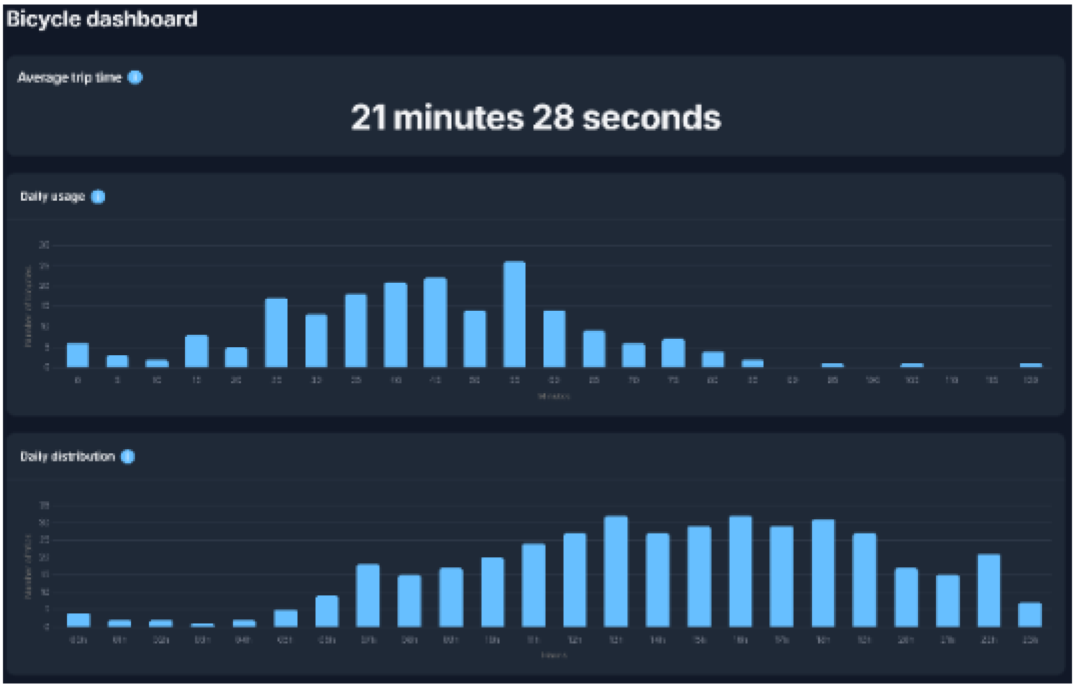


Figure 22: Bicycle dashboard

Finally, Figure 23 shows the heatmap, corresponding to the aforementioned fifth metric.

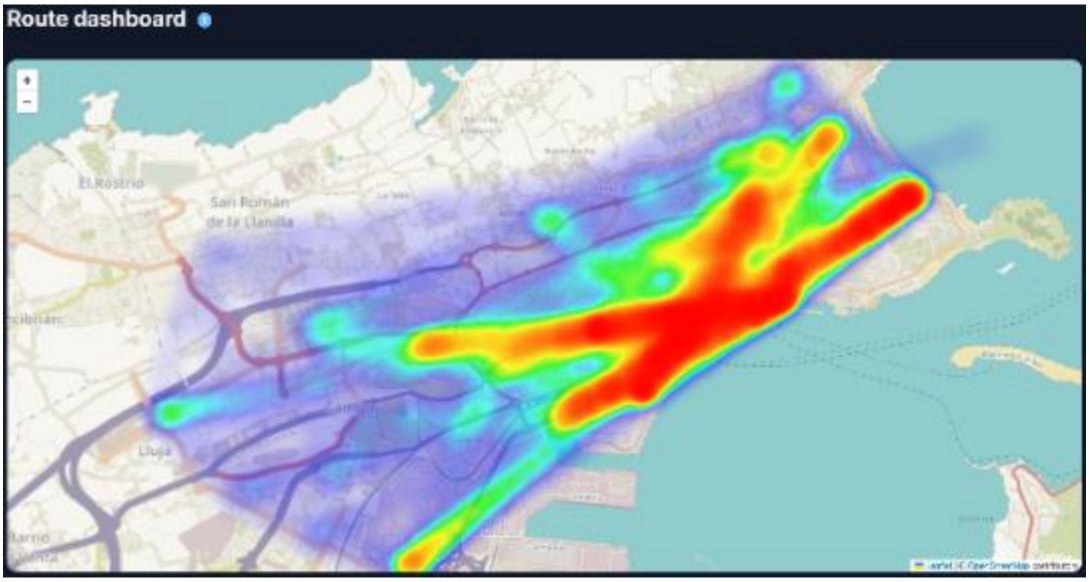


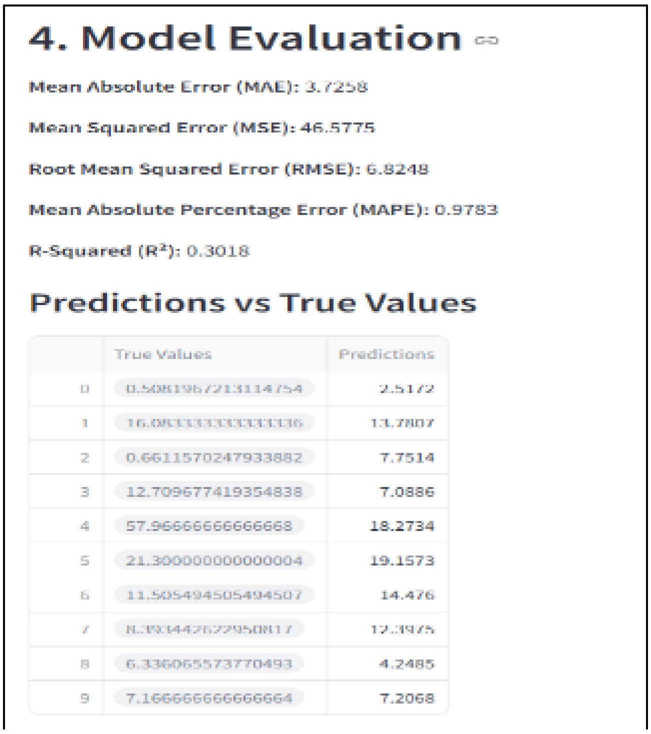
Figure 23: Heatmap corresponding to the fifth metric

These three dashboards, along with a main page to navigate through them, comprise the current state of the deployment, testing, and assessment phase of the mobility use case.

4.3 Valorisation of energy consumption and customer reactions / complaints in Greece

4.3.1 Connection with the SEDIMARK toolbox

After training, the UI displays model performance metrics and evaluation results. Users can view graphical comparisons, such as predictions vs. true values, to analyse the accuracy of the model and mean daily consumption predictions. The following dashboards (Figure 24, Figure 25 and Figure 26) illustrate how the chain of the results is displayed on the interface.


4. Model Evaluation

Mean Absolute Error (MAE): 3.7258

Mean Squared Error (MSE): 46.5775

Root Mean Squared Error (RMSE): 6.8248

Mean Absolute Percentage Error (MAPE): 0.9783

R-Squared (R^2): 0.3018

Predictions vs True Values

	True Values	Predictions
0	0.5081967213114754	2.5172
1	16.083333333333336	13.7807
2	0.6611570247933882	7.7514
3	12.709677419354838	7.0886
4	57.96666666666668	18.2734
5	21.300000000000004	19.1573
6	11.505494505494507	14.476
7	8.393442622950817	12.3975
8	6.336065573770493	4.2485
9	7.166666666666664	7.2068

Figure 24: Table displaying a comparison of the predicted values and the true values for energy consumption.

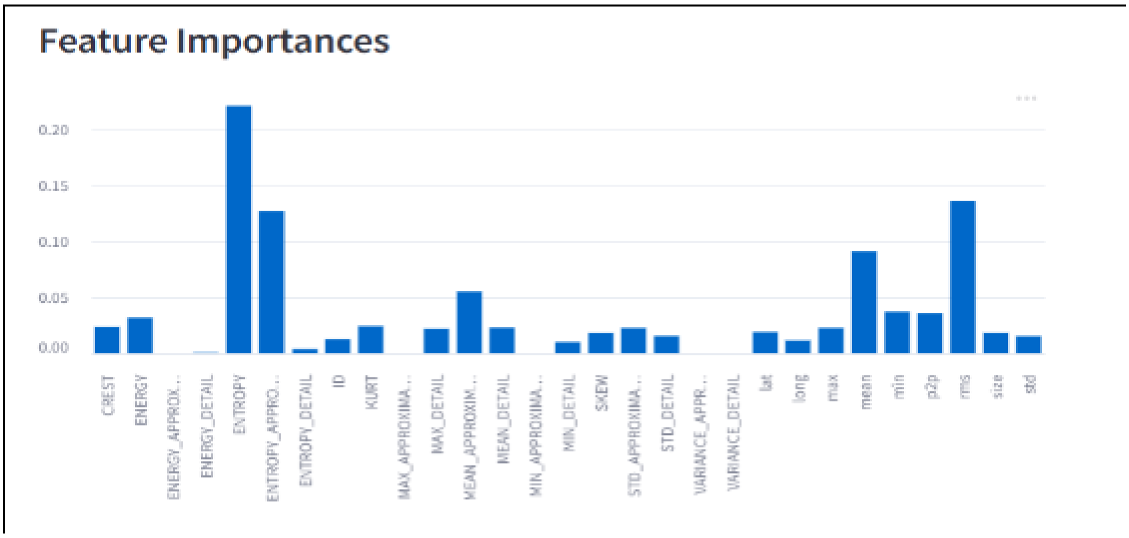


Figure 25: Visualisation of feature importances from the dataset

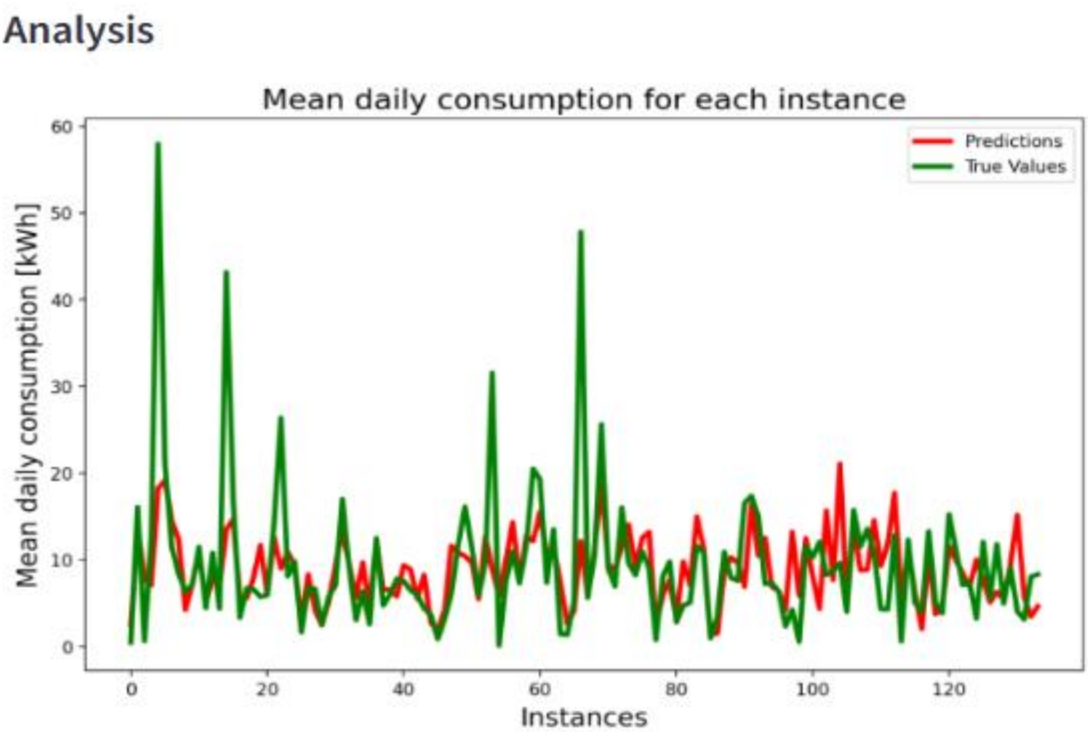


Figure 26: Line plot comparing the predicted vs. true values of mean daily energy consumption for each instance in the test set.

The graph shows that the model captures the general trend of energy consumption, particularly for low-to-moderate values. While the model slightly underpredicts peaks, it performs well in aligning with the overall behaviour of the data.

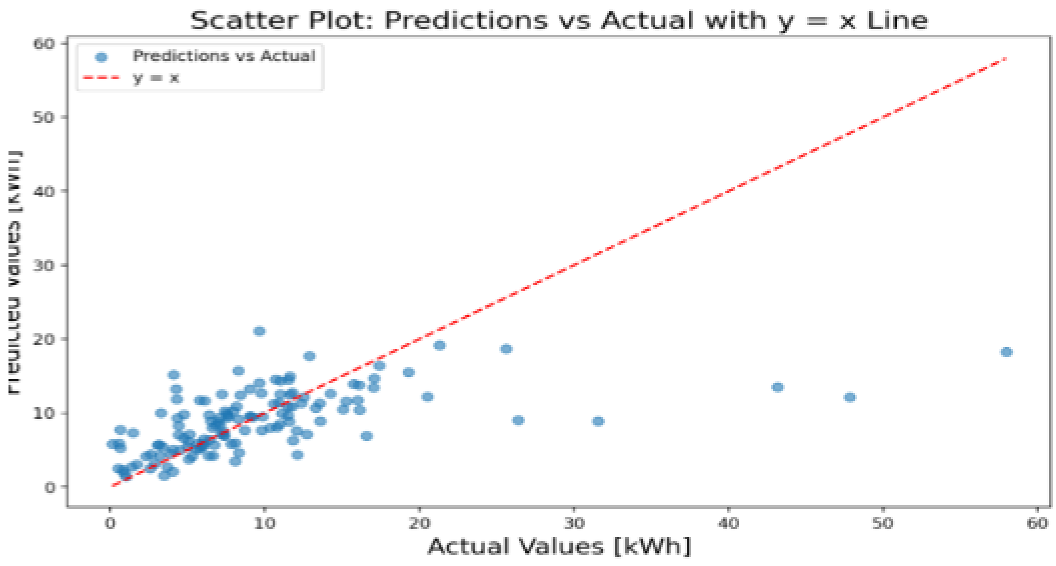


Figure 27: Scatter plot of predictions vs. true values

The above scatter plot (Figure 27) highlights that the model performs reliably for lower energy values where predictions align closely with the actual values. Although deviations increase for higher values, the overall pattern follows the expected trend, indicating the model's potential with further refinement.

The purpose of the whole analysis is to help the user assess how closely the predictions of the model align with the actual values, identify areas where the model might overpredict or underpredict energy consumption, and provide a visual overview of the model's accuracy and performance across the entire test set.

4.4 Valuation and commercialisation of water data in France

4.4.1 Connection with the SEDIMARK toolbox

Target KPIs of the water data used cases are summarised in Table 4. Initial focus has been on KPIs Water.01, Water.04 and Water.05, a Twin·Picks interface has been deployed on top of a Stellio NGSI-LD context broker. The ingestion flow to capture data from LoRaWAN devices has been made generic within NiFi so it adapts to any device deployed within the network. The flow, illustrated in Figure 28 register to payload sent by the LoRaWAN network through a MQTT broker. When an uplink is sent by a device, the NiFI processor intercepts and decode it using a decoder made available on-line, in a public git repository [13]. This repository is also used by the Twin·Picks interface to provide the user with the list of available device decoder for provisioning the device into the network server.

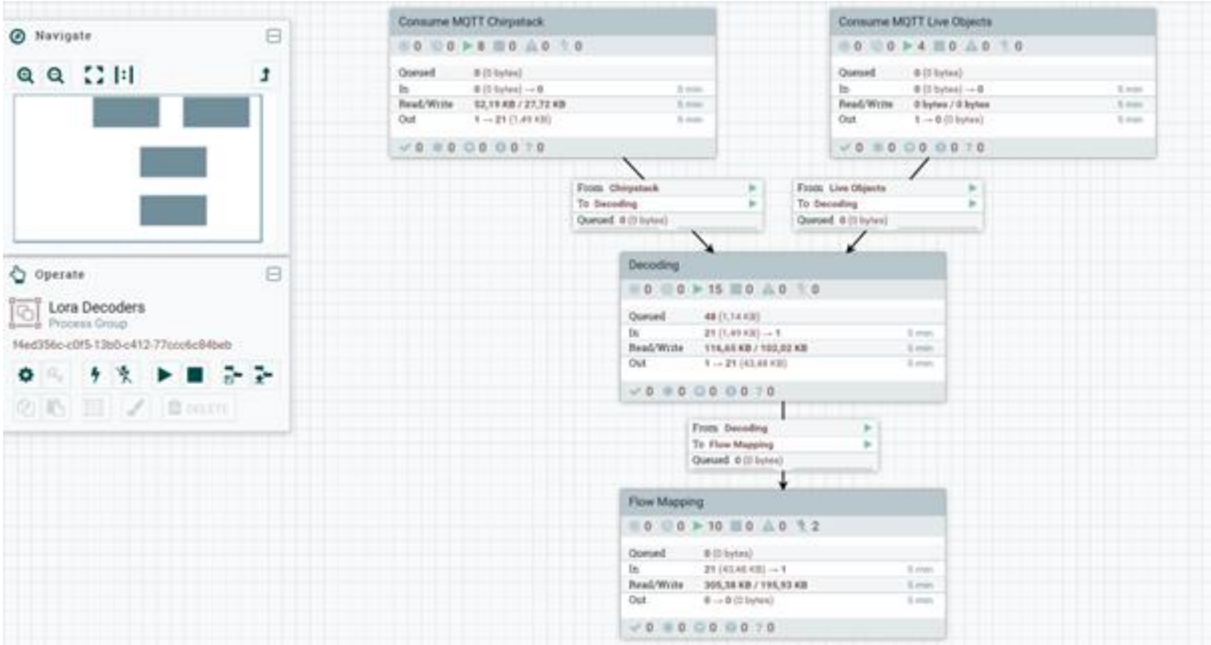


Figure 28: NiFi flow set for device payload decoding, mapping to NGSI-LD and saving.

As an illustration, a payload “83672201018482330701” sent by an ultrasonic distance sensor to be deployed for River height monitoring is decoded as seen in Figure 29:


```
[
 {
  "id": "urn:ngsi-ld:Device:24e124329e095877",
  "type": "Device",
  "temperature": {
   "type": "Property",
   "value": 29,
   "observedAt": "2025-01-07T14:55:20Z",
   "unitCode": "CEL",
   "datasetId": "urn:ngsi-ld:Dataset:Raw"
  },
  "alarm": [
   {
    "type": "Property",
    "value": 1,
    "observedAt": "2025-01-07T14:55:20Z",
    "datasetId": "urn:ngsi-ld:Dataset:Temperature:Raw"
   },
   {
```

```
      "type": "Property",
      "value": 1,
      "observedAt": "2025-01-07T14:55:20Z",
      "datasetId": "urn:ngsi-ld:Dataset:Threshold:Raw"
    }
  ],
  "distance": {
    "type": "Property",
    "value": 1.843,
    "observedAt": "2025-01-07T14:55:20Z",
    "unitCode": "MTR",
    "datasetId": "urn:ngsi-ld:Dataset:Raw"
  }
 }
]
```


Figure 29: Example of a NGSI-LD encoding of a packet sent by a river height sensor

Weather information has already been connected, and corresponding dataset are made available using DCAT based annotations, as per Figure 30. These will require to be aligned the DCAT mapping under finalisation within the SEDIMARK Ontology.

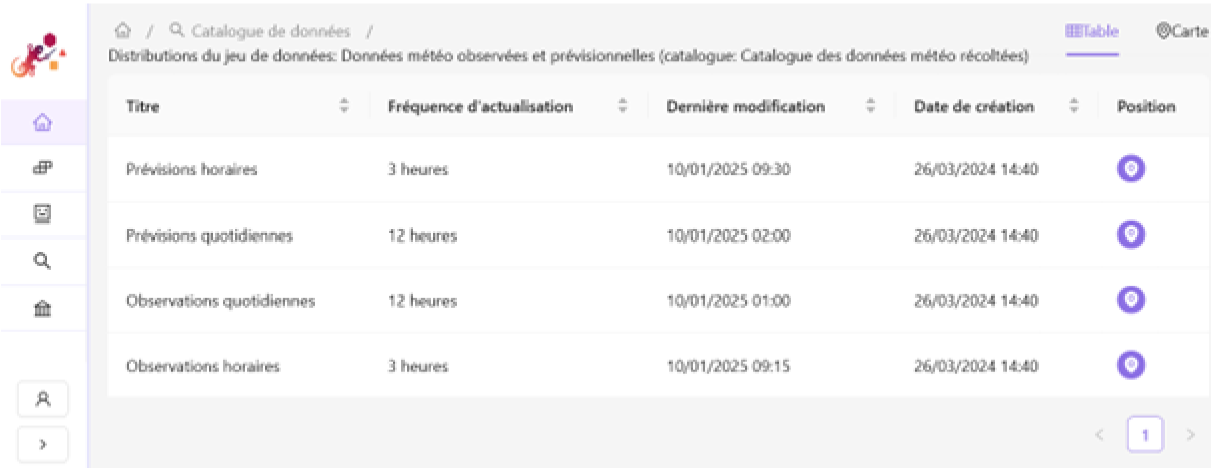


Figure 30: Datasets listing based on DCAT annotation.

5 Conclusions

This deliverable, in conclusion, is an update to how the SEDIMARK system plans to address diverse urban and environmental challenges. In particular, it focuses on the refinement of integration strategies, optimize system performance, and overcome identified challenges.

These findings set a strong foundation for future assessments and development, ensuring the project continues to advance toward its goals of delivering innovative, impactful solutions across pilot sites. The next steps involve enhancing the overall impact and applicability of SEDIMARK. Implementing the refined integration strategies and optimized system components in real-world pilot settings to evaluate their effectiveness and gather user feedback should be the first step. This phase will help identify additional areas for improvement and validate the proposed solutions. Additionally, efforts should focus on assessing the scalability and adaptability of the system to other cities or regions, ensuring that the solutions can be tailored to diverse urban and environmental contexts.

6 Bibliography

[1] SEDIMARK, Deliverable 5.1: Evaluation methodology, metrics and integration plan. Second version, SEDIMARK, September 2024.

[2] SEDIMARK, Deliverable 2.1: Use cases definition and initial requirement analysis. First version, SEDIMARK, June 2023.

[3] SEDIMARK, Deliverable 2.2: SEDIMARK Architecture and Interfaces. First version, SEDIMARK, September 2023.

[4] SEDIMARK, Deliverable 5.3: Integrated releases of the SEDIMARK platform. Second version, SEDIMARK, December 2024.

[5] Wikipedia: Communauté d'agglomération de la Riviera française — Wikipédia

[6] AISHub: Free AIS vessel tracking | AIS data exchange | JSON/XML ship positions

[7] Eaufrance | Le service public d'information sur l'eau

[8] Accueil | Sandre

[9] SEDIMARK, Deliverable 3.1: Energy efficient AI-based toolset for improving data quality. First version, SEDIMARK, December 2023.

[10] Davies-Bouldin Index : davies_bouldin_score — scikit-learn 1.6.1 documentation, Davies-Bouldin Index - GeeksforGeeks, Davies–Bouldin index - Wikipedia

[11] Swagger documentation : https://lidotiku.api.hel.fi/swagger

[12] ML Flow : MLflow: A Tool for Managing the Machine Learning Lifecycle

[13] STELLIO: https://github.com/stellio-hub/device-catalog