
# SEcure Decentralised Intelligent Data MARKetplace

## D2.1 Use cases definition and initial requirement analysis

| Document Identification | |
|---|---|
| **Contractual delivery date:** | 30/06/2023 |
| **Actual delivery date:** | 30/06/2023 |
| **Responsible beneficiary:** | WINGS |
| **Contributing beneficiaries:** | ALL |
| **Dissemination level:** | PU |
| **Version:** | 1.0 |
| **Status:** | Final |

**Keywords:**

Use cases, Requirements, Security, privacy and trust, Data quality management and data processing, Artificial Intelligence, Energy efficiency, Interoperability, Data storage, Publication and discovery, User interface, Smart Contract and Tokenization, Business and technology landscape, Trust and security analysis.

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. and is also partly funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the European Commission or UKRI.

# Document Information

| Document Identification | | | |
|---|---|---|---|
| **Related WP** | WP2 | **Related Deliverables(s):** | |
| **Document reference:** | SEDIMARK_D2.1 | **Total number of pages:** | 118 |

| List of Contributors | |
|---|---|
| **Name** | **Partner** |
| Elias Tragos<br>Diarmuid O' Reilly Morgan<br>Erika Duriakova | NUID UCD |
| Luis Sánchez<br>Pablo Sotres<br>Juan Ramón Santana | UC |
| Tarek Elsaleh | SURREY |
| Maxime Costalonga<br>Jairo Rojas-Delgado | ATOS |
| Gabriel Mihail Danciu | SIE |
| Luc Gasser<br>Gilles Orazi | EGM |
| Maroua Bahri<br>Nikolaos Georgantas | INRIA |
| Alberto Carelli<br>Andrea Vesco | LINKS |
| Panagiotis Vlacheas<br>Panagiotis Demestichas<br>Loizos Koutsantonis<br>Konstantinos Almpanakis<br>Evangelia Lamprakopoulou | WINGS |

| Document History | | | |
|---|---|---|---|
| **Version** | **Date** | **Change editors** | **Change** |
| 0.1 | 06/02/2023 | MYT | Added Section 5.1 |
| 0.2 | 06/02/2023 | UCD | Added Section 2 and ToC for section 6 |
| 0.3 | 06/02/2023 | UC, SDR | Added Section 5.2 |
| 0.4 | 08/03/2023 | WINGS | Added Section 1 except the glossary |
| 0.5 | 08/03/2023 | MYT | Formatted Section 5.3 |
| 0.6 | 12/03/2023 | SIE | Added chapter 3 |
| 0.7 | 13/03/2023 | LINKS | Added draft of Section 6 and subsections |
| 0.9 | 05/04/2023 | MYT | Added subsections to Section 5.3 |
| 0.10 | 17/04/2023 | LINKS | Finalized Sect. “Initial trust and security analysis” |
| 0.11 | 12/05/2023 | UCD | Added content in section 6.1 |
| 0.12 | 16/05/2023 | UCD | Added content in section 6.2 |
| 0.13 | 17/05/2023 | LINKS | Updated Figures of Sect. “Initial trust and security analysis” and typos corrections |
| 0.15 | 22/05/2023 | INRIA | Added content in Section 6.6 |
| 0.16 | 26/05/2023 | LINKS | Added requirements tables in Sect. “Functional and non-functional requirements” (i.e., Sect. 7.2, Sect. 7.10 and Sect. 7.11) |
| 0.17 | 30/05/2023 | UCD, UC | Added more content in Sections 6.1, 6.2, 6.3, 6.8 |
| 0.18 | 30/05/2023 | UC | Added content in Section 4.2 |
| 0.19 | 30/05/2023 | ATOS | Updated |
| 0.20 | 31/05/2023 | SURREY | Added content in Section 6.4, 6.7 |
| 0.21 | 31/05/2023 | SIE | Updated requirements |
| | 01/06/2023 | MYT | Added Exceptions and Alt. Flows in Section 4.3 |
| 0.22 | 02/06/2023 | INRIA | Added more content in Section 6.6 |
| 0.23 | 05/06/2023 | LINKS | Review of requirements tables of responsibility |

| Document History | | | |
|---|---|---|---|
| **Version** | **Date** | **Change editors** | **Change** |
| 0.9 | 10/06/2023 | WINGS | Editing version for being ready for internal review |
| 0.9b | 27/06/2023 | WINGS | Internal review comments addressed and version ready for quality review |
| 0.95 | 28/06/2023 | ATOS | Quality Review Form |
| 1.0 | 30/06/2023 | ATOS | FINAL VERSION TO BE SUBMITTED |

| Quality Control | | |
|---|---|---|
| **Role** | **Who (Partner short name)** | **Approval date** |
| **Reviewer 2** | Pablo Sotres, Luis Sánchez (UC) | 19/06/2023 |
| **Reviewer 1** | Arturo Medela (ATOS) | 15/06/2023 |
| **Quality manager** | María Guadalupe Rodríguez (ATOS) | 27/06/2023 |
| **Project Coordinator** | Arturo Medela (ATOS) | 30/06/2023 |

# Table of Contents

# List of Figures

# List of Tables

# List of Acronyms

| Abbreviation / acronym | Description |
|---|---|
| AI | Artificial Intelligence |
| API | Application Programming Interface |
| BDVA | Big Data Value Association |
| CIA | Confidentiality, Integrity, Availability |
| CKAN | Comprehensive Knowledge Archive Network |
| CS | Customer Support |
| DICOM | Digital Imaging and Communications in Medicine |
| DID | Decentralized Identifiers |
| DLT | Distributed Ledger Technology |
| DVFS | Dynamic Voltage and Frequency Scaling |
| Dx.y | Deliverable number y belonging to WP x |
| EC | European Commission |
| EU | European Union |
| GDPR | General Data Protection Regulation |
| GPS | Global Positioning System |
| ICS | Industrial Control System |
| IDE | Integrated Development Environment |
| IEC | International Electrotechnical Commission |
| IEEE | Institute of Electrical and Electronics Engineers |
| IP | Intellectual Property |
| ISO | International Organization for Standardization |
| IdM | Identity Management |
| IDSA | Interactive Digital Software Association |
| IoT | Internet of Things |
| IOTA | Internet of Things Application |
| JSON-LD | JavaScript Object Notation for Linked Data |
| KPI | Key Performance Indicator |

| Abbreviation / acronym | Description |
|---|---|
| LoRaWAN | Lo(ng) Ra(nge) Wide Area Network |
| ML | Machine Learning |
| NFR | Non-Functional Requirement |
| NFT | Non-Fungible Token |
| OWL | Web Ontology Language |
| QoS | Quality of Service |
| RAM | Random-Access Memory |
| RDF | Resource Description Framework |
| SDI | Serial digital interface |
| SSI | Sensitive Security Information |
| SSL | Secure Sockets Layer |
| TLS | Transport Layer Security |
| UC | Use Case |
| UI | User Interface |
| URL | Uniform Resource Locator |
| VC | Version Control |
| WAsP | Wind Atlas Analysis and Application Program |
| WFS | Web Feature Services |
| WP | Work Package |
| ZKP | Zero-Knowledge Proof |

# Executive Summary

This deliverable serves as the initial comprehensive documentation of the project's use cases and requirements, playing a crucial role in guiding the design, development, and validation of the SEDIMARK marketplace. The key outcomes of this deliverable include detailed use case descriptions and an analysis of the derived requirements.

SEDIMARK encompasses four main use cases: Mobility Digital Twin, Urban Bike Mobility Planning, Valorisation of Energy Consumption and Customer Reactions/Complaints, and Valuation and Commercialisation of Water Data. The use case descriptions provide in-depth information about each use case's objectives, relevance to the project, required functionalities, and measurement of success. Various aspects are covered, such as brief descriptions, involved actors and stakeholders, current status, preconditions, postconditions, exceptions, services offered, data generation and sharing, associated data models, key performance indicators (KPIs), and expected flows.

The requirements engineering process in SEDIMARK follows a double iterative approach, inspired by the ISO/IEC/STANDARD IEEE 29148 [1] and aligned with the agile project methodology. Requirements are categorized as functional or non-functional, prioritized, and described in detail. They include identifiers for traceability, short names, descriptions, purpose goals, validation criteria, constraints, dependencies, conflicts, relevant use cases, and advantages and disadvantages of each requirement. The requirements are classified into ten categories, addressing different research aspects and implementation activities within the project.

Additionally, this deliverable provides an analysis of the business and technology landscape surrounding SEDIMARK, encompassing available and emerging technologies and their adaptation to meet market demands. Furthermore, an initial trust and security analysis is conducted, considering the use cases and envisioned solutions, given the critical importance of trust and security within SEDIMARK.

Overall, this deliverable presents the analysis of four niche use cases and over 70 core requirements, serving as a guiding framework for the design and development of the decentralized, trustworthy, interoperable, and open SEDIMARK marketplace. The requirements analysis influences the development activities and aims to create intelligent, energy-efficient data management tools that ensure high-quality data and services for consumers. The requirements are mapped to architectural patterns, supporting the instantiation of the platform in WP5 “Integration, testing and evaluation”. The defined use cases are utilized in WP5 to develop pilot demonstrators and evaluate and monitor the solution's performance using relevant assessment metrics.

# 1 Introduction

## 1.1 Purpose of the document

This document provides the initial detailed description of the project use cases (UCs) and the initial requirements. As far as it concerns the use cases, the goal is to define them with specific details about their implementation and how they will use the project tools, and especially aiming to combine data from different data sources and platforms to show the potential for secure combination and sharing of data across sites. With respect to the requirements, the aim is to gather requirements from various stakeholders, industrial applications, the UCs and the concept of EU Data Spaces, and analyse them, in order to extract functional and non-functional requirements for making the data marketplace decentralised, trustworthy, interoperable and open to new data (open data), with intelligent AI-based and energy efficient data management tools capable of providing high quality data and services to consumers.

## 1.2 Relation to another project work

D2.1 is a very important document, setting the context of all the next activities in the project. D2.1 will feed the project's architecture to be defined in D2.2 (SEDIMARK architecture and interfaces- First version) [September 2023], which will be built upon the requirement elicitation process of D2.1. This architecture will be then used for the design and development of all the activities in WP3 (Distributed data quality management and interoperability) and WP4 (Secure data sharing in a decentralized Marketplace). In addition, after the requirements of D2.1 are analysed and mapped to the architectural patterns, this will be further used to create the instantiation for the development of the platform towards WP5 (Integration, testing and evaluation). The use cases that have been defined will play a significant role in WP5 of the project. They will be utilized to create pilot demonstrators, which are practical implementations of the solution being developed. These demonstrators serve as real-world examples that showcase how the solution works in practice. During the pilot phase, the solution will be thoroughly evaluated and monitored using performance assessment metrics. These metrics are specifically designed to measure the solution's effectiveness and efficiency in addressing the activities and objectives outlined in the use cases. By assessing the solution against these metrics, the project team can gain valuable insights into its performance and make any necessary adjustments or improvements.

## 1.3 Structure of the document

This document is structured in 8 major chapters.

**Chapter 1** introduces the objective of the document and how it relates to the project's activities.

**Chapter 2** presents the methodology for defining use cases and requirements.

**Chapter 3** presents the business and technology landscape, comprising the available and/or emerging technologies on the market and business adapting to meet the demands. There are a few key trends which SEDIMARK considers when analysing the context of our project.

**Chapter 4** presents the SEDIMARK use cases, with focus on the Stakeholders involved, the Preconditions, the Postconditions, the Data sources and data models, and the Use Case flow.

**Chapter 5** focuses on a very special aspect for SEDIMARK, the initial trust and security analysis of the use cases and the solutions envisioned by the project.

**Chapter 6** presents the functional and non-functional requirements of the project, categorized in 10 areas: 1) Non-functional - system architecture requirements, 2) security, privacy, and trust, 3) data quality management and data processing, 4) AI, 5) energy efficiency, 6) interoperability, 7) data storage, 8) data discovery and data sharing, 9) openness and 10) user requirements. This categorization helps to map more easily the requirements to the subsequent development tasks of WP3 (Distributed data quality management and interoperability) and WP5 (Integration, testing and evaluation).

**Chapter 7** concludes the document, summarizing the main results and the following steps to take in the project in alignment with the project roadmap.

# 2 Methodology for defining use cases and requirements

SEDIMARK aims to adopt a use-case based approach for the design of the overall system and the system architecture. This means that the system design will be initially driven by the specific use cases that the project will support, with a subsequent generalisation, so that it can be extensible to support a plethora of use cases, services and scenarios and avoid conflicting requirements, as described in [2].

Requirements engineering is one of the most important activities in the preparation of a software system, because broadly speaking the success of a system can be measured as the degree to which it serves the purpose(s) it was built for [3]. A system that is functioning perfectly but serves a different purpose than the one that it was built for should not be considered as a successful system. Requirements engineering can be considered as the process that identifies what are the needs and the purposes of the stakeholders and translates them into specific goals so that they can be used in order to develop the overall system components and architecture. Thus, requirements engineering is the process of identifying the real-world goals, functions and constraints of a system, and then converting them to precise specifications for the software [4].

Considering the above, requirements engineering should first identify the high level “goals” of the system and the “purpose” it will serve and then it will have to go into details, breaking the goals/purpose into categories, providing lists of specific “targets” that have to be met in each category, including definitions of “what” needs to be built and how it can be verified [3].

As mentioned above, SEDIMARK will employ a use-case driven requirement engineering process that can be very well combined with the agile development methodology for designing and developing the project system. Use-case driven requirement engineering has been favourable in the past as a method for designing software systems, because it effectively involves the stakeholders from the beginning, while capturing their requirements and keeping them involved throughout the whole project implementation period [5] [6] [7]. In SEDIMARK, we use the term “use-case driven” to note that the project use cases are playing an important role on the system design and on the development of the work, driving all the activities from design to coding.

As described in [3] , use cases are used to provide a description of the system functions, both under various conditions and for specific purposes, and. besides they can be used as a mean to interact with stakeholders, understanding their needs and analysing their requirements. Due to the important role of use cases in the system design, this requires that use cases need to be very carefully designed and described, with adequate details that will help towards extracting meaningful requirements for the system design.

The main process used in SEDIMARK for requirements engineering, as well as its relation to the main tasks for design, development and testing can be seen in Figure 1:

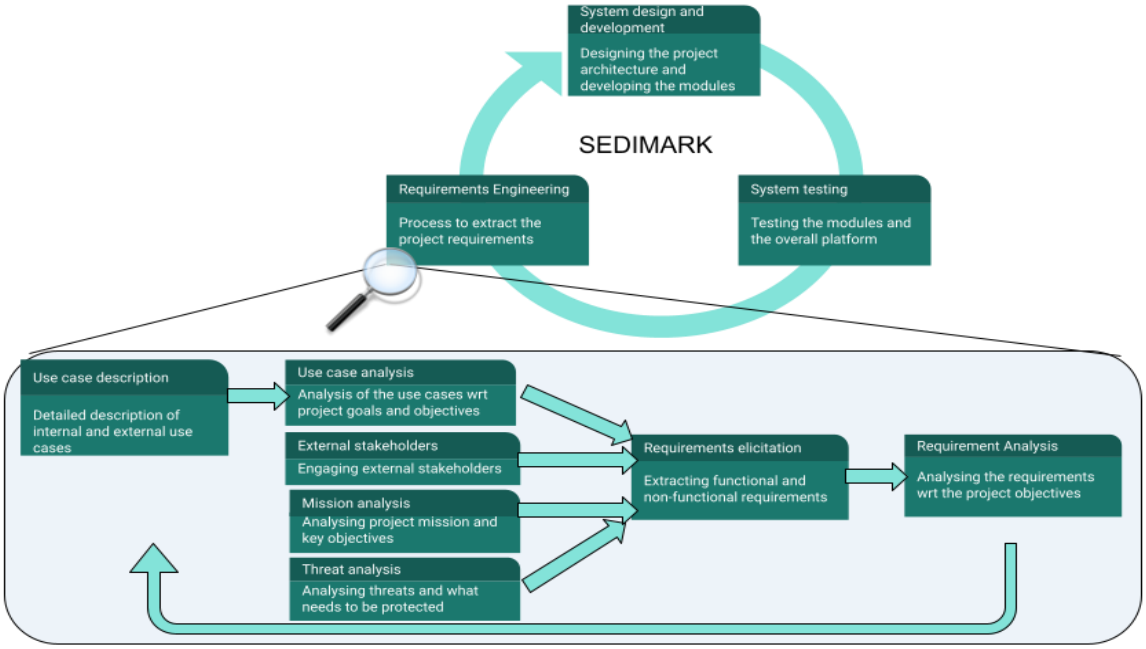


**Figure 1. Overall requirements of engineering process.**

The requirements engineering process of SEDIMARK is inspired by the ISO/IEC/STANDARD IEEE 29148 [ISO1] but follows a double iterative process since, as described above, it is aligned with the agile project methodology. On the higher level, the Requirements Engineering process is feeding the System Design and Development with (i) the various types of requirements (system, software, user, implementation, etc.), (ii) architectural requirements and (iii) key performance indicators for assessing if the project targets are met. Also, in an iterative process the Requirements Engineering process receives feedback from the system testing tasks in order to update the requirements to better reflect the project objectives and help redesign some modules or project features.

Internally, the Requirements Engineering process is also an iterative process which is done in 3 cycles as shown in Figure 2. Use case descriptions/analysis and requirements elicitation/analysis are done in parallel, exchanging feedback and information so that the next version of each task will be improved to better map the project design.

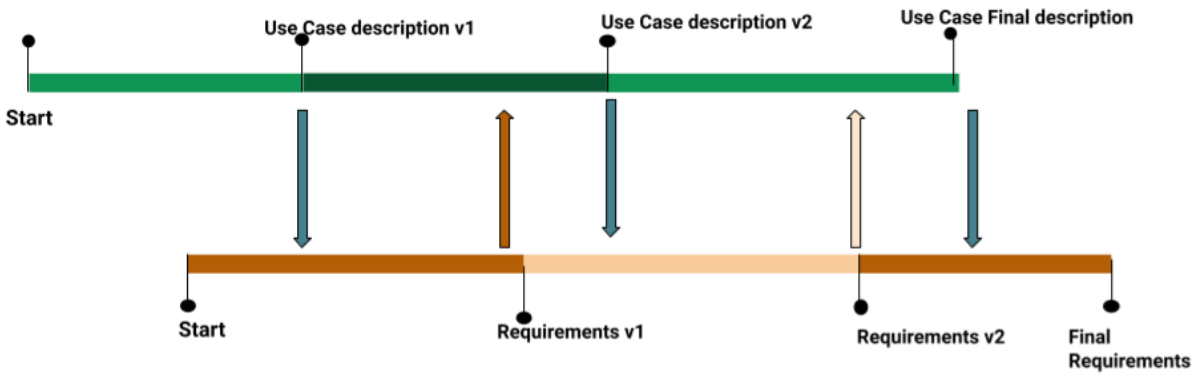


**Figure 2. Cycles of the iterative process for requirements engineering.**

The process for extracting the requirements of the project, as seen in Figure 2, consists of six main steps, which are analysed below in the following subsections.

### Use case description

This step involves all the processes required to provide an initial detailed description of the internal use cases of the project. As described in [8][2], use cases are a key factor into capturing system requirements and driving the design and the implementation of a software system. Use cases are basically expressing the views and the goals of the users and the customers in a textual format which is easier for non-technical users and provides a greater ability for understanding what the system has to provide.

Within SEDIMARK, we defined a common template that will be used for describing the use cases with detailed fields that can capture enough information so as to be able to understand what each use case will provide, why it is related with the SEDIMARK project, which modules it will require and how we can measure its success. Considering that the Requirements Engineering process is an iterative process as mentioned above, this template is a "live" template which might be updated in the future before the final system design is completed. The current version of the use case description template is shown in Table 1, together with a brief description of each field. It has to be noted here that for improved presentation and readability, in Section 5, the Use Cases are presented in plain text format and not within this table. Still, the structure of the sections follows closely the fields described in the template.

**Table 1. Template for the description of the use cases.**

| Name of Use Case: | Use case name example | | |
|---|---|---|---|
| Created By: | (Use case owner) | | (Last user updating the text) |
| Date Created: | DD/MM/YYYY | Last Revision Date: | DD/MM/YYYY |
| Description: | This section should provide a description of both the reason for using the use case and the expected outcome of the use case. | | |
| Actors/Stakeholders: | Persons or entities that are related with the use case or play a role in the execution of the use case, i.e., providers or users. | | |
| Current status of the UC | Description of how (and if) the use case is currently executed and what are the baselines. | | |
| Preconditions: | Description of any conditions that must be true or activities that must be completed prior to executing the use case. | | |
| Postconditions: | Description of the state of the system at the conclusion of the use case. | | |
| Exceptions: | Description of any errors/issues that may result during use case execution and how the system will react or respond to those errors. | | |
| Services to be offered | What services will be offered by the use case and to whom. | | |

| Name of Use Case: | Use case name example |
|---|---|
| Data to be generated/shared / data models | Brief description of the types of data to be generated/shared together with privacy characterisation. |
| KPIs | How the success of the use case will be assessed. |
| Flow: | Description of all actions of the user and the expected system responses for planned normal execution of the use case. |
| Alternative Flows: | Alternative flows are usually the result of options or exceptions built into the use case which may alter the primary flow. |
| Requirements: | This section should describe any non-functional or special requirements for the system as the use case is executed. |

### Use case analysis

This step will analyse the descriptions of the use cases against the project goals and objectives. In this step, each use case will be analysed separately, in order to find out the specific characteristics of each use case with respect to the system design. Next, the use cases will be jointly analysed in order to find commonalities and differences, as well as conflicting details that should be critical towards the system design. The overall goal of the use case analysis is to assist in the requirements elicitation process.

### External stakeholder analysis

This step will consider external stakeholders of the project and their objectives for the SEDIMARK project. In this respect, past projects in the same areas as SEDIMARK, EU initiatives, global associations and end users will be taken into account so as the project design and implementation take into account their objectives and requirements too. Additionally, legal aspects will also be analysed, especially with respect to privacy and data management laws.

### Mission analysis

In this step, the project main objectives, goals target and the overall technical and business problems or opportunities that exist, in order to identify potential solutions that could address such problems or that could take advantage of the opportunities. This process will mainly drive the elicitation of the non-functional requirements of the project that cover high-level aspects of the project design and implementation.

### Requirements elicitation

This is the main process that will identify the project requirements, transforming the analysis of the use cases, stakeholders' needs and project mission in well-defined requirements that can help drive the project design and implementation. The SEDIMARK requirements are split into non-functional and functional requirements. For the requirement specification, SEDIMARK was inspired by ISO 29148:2011 [9] that provides a detailed set of instructions regarding extracting and describing requirements. Mapping this to the project specifics resulted in the template that can be seen in

Table 2.

Table 2. Template for the description of requirements.

| Section | Description |
|---|---|
| ID | The requirement ID at a specific format |
| Short Name | Short title of the requirement |
| Type | Functional/Non-Functional |
| Priority | High/Medium/Low |
| Requirement Level | Required/Recommended/Optional |
| Description | Short description of the requirement |
| Additional information | Additional information that might be related with the requirement. |
| Purpose/Goal | Why is this requirement required or important in the project |
| Validation criteria | Metrics that can be used if/how this is met in the project |
| Constraints: | Any constraints that might hamper the implementation of the requirement |
| Dependencies: | If this requirement is depended on another requirement |
| Conflicts: | If it conflicts with other requirements |
| Relevant Use Cases | The Use cases that are related to this requirement |
| Pros/Cons | The advantages and disadvantages on having this requirement in the project |
| Category | Which requirement category includes it |

### Threat analysis

This step provides an initial analysis of the threats for the system that is currently under design and not for an operational deployment. The fact that SEDIMARK is not developed yet poses several limitations on the type of threat analysis that can be performed. Thus, here, the analysis will be limited to the basic security and privacy issues that might exist in data marketplaces, what are the assets that need to be protected and what type of attacks can be performed in such a system. This analysis will help identify the system requirements with respect to privacy, data management and overall trustworthiness.

### Requirements analysis

This step includes the analysis of the requirements in order to extract required information regarding the design of the project architecture. More information regarding this step will be given in the next project deliverable D2.2, to be delivered in September 2023.

# 3 Business and technology landscape

The digital transformation trend is driving businesses to adopt new technologies and approaches to improve their efficiency and customer experience. This requires careful consideration of system architecture requirements to ensure that the technologies and applications are well-integrated, scalable, and robust.

The growing volume of data, along with the development of AI algorithms, raise concerns around data security, privacy, and trust. Businesses need to prioritize these areas to protect their data and ensure that they comply with regulations and best practices.

The growing volume of data requires businesses to prioritize data quality and effective data processing to derive valuable insights. This necessitates the use of appropriate data quality management techniques, including data cleansing, transformation, and normalization.

Businesses are adopting AI to automate tasks and enhance decision making. This requires expertise in AI algorithms as well as efficient data labelling, model training and deployment.

As Cloud Computing and Internet of Things (IoT) services become more used, businesses need to prioritize energy efficiency in their computing infrastructure by using energy-efficient hardware, optimizing resource usage, by improving the current processing methods and algorithms. For example, Google's DeepMind [10] used machine learning algorithms to reduce energy consumption in their data centres by up to 40%. Additionally, businesses can use compression techniques to reduce the amount of data that needs to be processed, thereby reducing power consumption.

Interoperability is essential for businesses to ensure that their systems communicate efficiently, and this imposes adherence to standards and the use of interoperable protocols like JSON-LD [11], CSVW [12] , OData [13] and others.

As data volumes continue to grow, companies need to adopt scalable and cost-effective data storage solutions such as cloud storage, distributed storage or hybrid solutions.

Data discovery is just as essential as data interoperability: by including metadata managements, data catalogues or APIs for data searching the data sharing and exploration tasks can be properly covered. Openness refers to use of open standards, open-source technologies, and open data in order to promote transparency collaboration and innovation. This requires using open principles and practices, such as Apache libraries, Linux OS, open educational services or open datasets to validate various use-cases.

In the context of data spaces, defining user requirements is crucial for developing data-driven products and services that deliver value to users. There are some key aspects one should keep in mind when defining user requirements in the context of Data Spaces.

To begin, it is necessary to identify the target audience. Subsequently, user feedback should be collected, followed by the analysis of the user data. Once the analysis is complete, and the user requirements are well defined, we need to prioritize these items, and develop data stories that describe how data will fit the needs of the users. In doing all from above, one should consider business aspects of data-driven service such as revenue models, costs and value propositions in order to create products that are commercially viable.

In the remaining of this chapter, we are reviewing/analysing the data spaces from a business perspective in section 3.1, as a relevant topic today that help business to perform better decisions or identify new opportunities. This is complemented in section 3.2 by the technology landscape organized by various domains and categories – the same categorization will be followed in our requirement analysis in section 6.

## 3.1 Data Spaces – the business perspective

Data Spaces is a concept that refers to a collection of interconnected data sources that are integrated via a single platform. The idea is to provide users with a unified view of multiple sources, thus making easier to analyse and extract information from various data.

Data Spaces are becoming a relevant topic since business generate an increasingly amount of data from various sources, like IoT devices. By integrating these data sources, data spaces ease the overview of the operations, customers, providing a means that could lead to better decisions or identification of new opportunities.

### 3.1.1 Data spaces initiatives

A list of European initiatives that focus on the developments in various domains and application of data spaces is listed below:

1. GAIA-X
    a. GAIA-X [14] is a European project that aims to create a federated, secure, and trustworthy data infrastructure for Europe. The project seeks to establish common standards, protocols, and governance frameworks for data sharing and data spaces across various industries.
2. IDSA
    a. The International Data Spaces Association (IDSA) [15] is a coalition of more than 130 member companies that share a vision of a world where all companies self-determine usage rules and realize the full value of their data in secure, trusted, equal partnerships. Their collective efforts are dedicated to transforming this vision into a tangible reality.
    b. Their goal is nothing less than a global standard for international data spaces (IDS) and interfaces, as well as fostering the related technologies and business models that will drive the data economy of the future across industries.
3. BDVA
    a. BDVA/DAIRO [16] focuses on enabling the digital transformation of the economy and society through Data and Artificial Intelligence by advancing in areas such as big data and AI technologies and services, data platforms and data spaces, industrial AI, data-driven value creation, standardisation, and skills.
4. OpenDEI
    a. OpenDEI [17] is an EU project that aims to support the development of digital platforms and data spaces in key industrial sectors. It focuses on creating an ecosystem of digital platforms, tools, and services that promote data sharing and collaboration between organizations.
5. FIWARE
    a. FIWARE [18] is an open-source platform that provides a set of APIs and components for the development of smart applications and services. It supports the creation of data spaces by providing tools and technologies that enable data sharing and interoperability between different systems.

### 3.1.2 Data Spaces related to SEDIMARK.

In the context of the SEDIMARK project, various data spaces are created to address the specific needs and requirements of each use-case. These data spaces facilitate the seamless

integration, sharing, and utilization of heterogeneous data to support efficient decision-making processes and enable innovative services. The data spaces related to SEDIMARK include:

#### Urban Mobility Data Spaces

The first use-case deals with urban mobility and aims to improve traffic management by predicting traffic congestion and recommending optimal routes. This dataspace gathers data from diverse sources such as traffic sensors, Global Positioning System (GPS) devices, and weather information. The urban mobility data spaces support stakeholders like city authorities, traffic management centres, and transportation companies in enhancing traffic flow, reducing travel times, and minimizing the environmental impact of traffic.

#### Energy Management Data Spaces

The second use-case addresses energy management by analysing consumers' energy behaviour and customer conduct in terms of complaints and churn. This data space combines data from residential energy consumption, weather data, customer complaints, and payment information. The energy management data spaces allow utility providers and residential customers to predict future energy consumption, calculate costs, and adopt more efficient pricing plans. These dataspaces also enable utility providers to reduce customer churn and improve overall customer satisfaction.

#### Smart Agriculture Data Spaces

This use-case focuses on smart agriculture, aiming to optimize irrigation and fertilization management by predicting crop water needs and nutrient requirements. This data space integrates heterogeneous data from sources like IoT devices, satellite imagery, and soil data. The smart agriculture data spaces enable stakeholders such as farmers, agronomists, and agricultural organizations to make better-informed decisions, reduce resource waste, and improve overall crop productivity.

#### Water Data Valuation and Commercialization Data Spaces

The fourth use-case focuses on the valuation and commercialization of water data, aiming to strengthen the Water Basin Authority's capacities in monitoring and analysing water quality and operating data. This data space integrates heterogeneous data from sources such as water quality measurements, weather data, flow rates, and energy consumption. The water data valuation and commercialization data spaces facilitate data sharing and collaboration among stakeholders like water unions, local authorities, citizens, and researchers. In conclusion, the SEDIMARK project establishes various dataspaces to address the unique challenges and requirements of each use-case. These data spaces enable the integration and sharing of diverse data sets, facilitating efficient decision-making processes and unlocking innovative services for stakeholders across multiple domains. By utilizing advanced AI tools, decentralized infrastructure, and secure data sharing mechanisms, the project data spaces contribute to a more sustainable and efficient future.

### 3.1.3 Business sectors in which the dataspaces can make a significant impact.

In various sectors like healthcare or transportation the integration and analysis of data from various sources can lead to improved decision-making, more efficient operations, and the development of innovative products and services.

#### Transportation [19]

The use of various data sources such as traffic sensors, GPS, and weather/pollution APIs can greatly improve transportation systems. Some potential benefits include:

- Optimizing transportation routes by analysing real-time traffic data, reducing fuel consumption and travel time.
- Enhancing public transit systems by using data to better predict demand and adjust schedules accordingly.
- Implementing smart traffic management systems that adapt to changing road conditions, reducing congestion, and improving road safety.
- Analysing the impact of weather and pollution on transportation networks and taking appropriate measures to minimize their effects.

### Energy [20]

Data Spaces can support the integration of data from various energy sources, including renewable energy generation, energy storage systems, and smart grid infrastructure. This can enable more efficient energy management, demand response programs, and the integration of distributed energy resources. Implementing Data Spaces in this sector can help address the challenges associated with energy churn by enabling better customer engagement, personalized offerings, and data-driven decision-making.

### Healthcare [21]

The integration of patient data, from various sources like EHR (Electronic Health Record), DICOM (Digital Imaging and Communications in Medicine), wearable devices, etc, can enhance the healthcare services. Using this type of data one can:

- Personalize patient healthcare plan and improve diagnostics by leveraging the patient specific data.
- Monitor patient health and provide near real-time feedback using wearable devices.
- Use AI to gain disease insights and identify patterns to predict outcomes.

Streamline data among healthcare providers to ensure the continuity of care.

### Finance [22]

In the finance sector integrating data from different sources like customer transactions, credit scores and market data, can lead to the development of more precise and customized financial product and services. This can include:

- Developing AI-driven trading algorithms that leverage market data and customer transactions to make smarter investments.
- Enhance the fraud detection and prevent using generative AI models previously trained on transaction data.
- Improve credit scoring models by incorporating alternative data sources to better assess customer trustworthiness.

### Manufacturing [23]

The Manufacturing process can be optimized by integrating data from various sources like sensors or hardware behaviour observation. This can lead to:

- Implementing predictive maintenance by analysing the sensors data in order to detect or predict equipment anomalies and schedule maintenance before the error happens.
- Optimizing the production process by analysing data to identify eventual bottleneck or inefficiency that can slow down or even shut down the production.
- Utilize data-driven insights to improve product design and the supply chain flow, making them more efficient, cost effective and environmentally friendly.

## 3.2 Technology landscape

The SEDIMARK project must be built on a robust and scalable system architecture, capable of handling diverse data sources, complex analytics, and advanced visualization tools. This architecture should support seamless integration of data from various domains, enable efficient data processing and management, and provide a high level of interoperability. In order to achieve all this, an assessment has been made to include all technologies organized by the various domains as follows.

### System Architecture Requirements

In the context of the SEDIMARK project, the system architecture landscape includes cloud-based platforms, microservices architecture, and containerization technologies like Docker [24] and Kubernetes. These technologies enable scalable and flexible infrastructure that supports diverse data sources and complex analytics while ensuring seamless integration and interoperability. Furthermore, this architecture ensures seamless integration and interoperability, keeping in line with the prevailing tendencies in the field.

### Security, Privacy, and Trust

The security, privacy, and trust landscape encompass various encryption algorithms, secure communication protocols (e.g., TLS, SSL), and identity and access management solutions (e.g., OAuth [25], OpenID Connect). Privacy-enhancing technologies, such as differential privacy and homomorphic encryption, can be employed to protect sensitive information. Additionally, blockchain technology can be utilized to establish transparency and trust in the system.

### Data Quality Management

The data quality management landscape includes tools and frameworks for data validation, cleansing, and enrichment, such as OpenRefine [26], Trifacta, and Talend. Metadata management solutions, like Apache Atlas, can be employed to provide context and facilitate data discovery and understanding.

### Machine Learning

The machine learning landscape comprises various frameworks and libraries that support advanced analytics, such as TensorFlow, PyTorch, Scikit-learn, and MLFlow. Additionally, distributed computing platforms like Apache Spark [27] and Dask can be used to scale machine learning workloads across multiple nodes for improved performance.

### Energy Efficiency

Energy efficiency in the technological landscape relates to optimizing hardware and software components to reduce energy consumption. Techniques like dynamic voltage and frequency scaling (DVFS), server consolidation, and energy-aware scheduling can be employed to minimize energy usage. Green data centres and the use of renewable energy sources can also contribute to improved energy efficiency.

### Semantic Web

The semantic web landscape involves technologies and standards that promote data sharing and reuse across different applications, such as RDF [28], OWL, and SPARQL. These technologies enable the creation of linked data and facilitate semantic interoperability among diverse data sources.

### Big Data Technologies

The big data technology landscape includes distributed data processing frameworks like Apache Hadoop [29] and Apache Spark, which can handle large-scale data storage and

processing. Data streaming platforms, such as Apache Kafka and Apache Flink, can be employed to manage real-time data flows and enable real-time analytics.

### Linked Data

The linked data landscape encompasses technologies and standards that support the creation, management, and querying of interlinked datasets, including RDF, SPARQL, and linked data platforms like Virtuoso and Apache Jena. These technologies facilitate the discovery, integration, and analysis of data from multiple sources.

### Open Source and Openness

The open-source landscape includes a wide range of software projects and tools that are publicly accessible and can be freely used, modified, and distributed. Promoting openness in the SEDIMARK project can be achieved by utilizing open-source technologies, adhering to open standards, and fostering collaboration and knowledge sharing among stakeholders.

### User Requirements

The user requirements landscape involves understanding the needs and preferences of various stakeholders, such as utility providers, customers, and researchers. This can be achieved through techniques like user interviews, surveys, and focus groups, as well as the analysis of usage data and feedback. Incorporating user-centred design principles and agile development methodologies can help ensure that the SEDIMARK platform meets the diverse needs of its users.

# 4 Use case description

## 4.1 Mobility Digital Twin in Helsinki

The digital twin of Helsinki is formed by a network of interoperable systems, exchanging data over standardized APIs (Application Programming Interfaces) [30] Geospatial data forms the backbone of the digital twin, on which additional topic specific data assets can be overlaid. The data sets involved in the digital twin also include data describing the traffic environment, in particular the road infrastructure. The open data offering of public administration in the Helsinki region is available from the HRI (Helsinki Region Info share) open data portal [31]

The relationship between the urban digital twin (and city SDI) and data marketplace is two-directional:

- City data is offered to the marketplace (both open and closed data cases exist). There are various reasons for this operating model, city may e.g., lack suitable own storage space or publishing tools, and utilising 3rd party publishing (e.g., marketplaces) provides better visibility to city's data.
- City may also utilize the data marketplace to gather information on the private data offering, and possibly to procure data to improve processes. The data city is interested in may be only available via marketplaces, or city may wish to avoid exclusive procurement and wish to procure service (i.e., access to the data) instead of investment (i.e., ownership)

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

There are two primary use cases:

1. Digital Twin uses data FROM data marketplace.
2. Digital Twin provides data TO the marketplace.

In both cases, the functional requirements include:

- The data can be free and public, or it can be e.g., restricted, exclusive or commercial. The metadata describing the availability of the data and licenses has to 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.

### 4.1.1 Current Status of the Use Case

- The digital twin of Helsinki is formed by a network of interoperable systems, exchanging data over standardized APIs. 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. These include, for example:
  - **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. [32].
  - **Conditions:** Helsinki Region Environmental Services HSY maintain a set of air quality measuring stations providing information on air quality in the city, available over an open interface [33] ).
- The relationship between the data marketplace and urban digital twin of mobility is envisioned to become two-directional, as it was introduced before in the beginning of the section 4.1.
- As the digital twin of mobility is formed as a "system of systems", the significance of interoperability and machine readability is highlighted in respect to the data marketplace.

### 4.1.2 Actors/Stakeholders

Actors are the people who will be initiating the system described in the use case. Initially, the following main actors have been identified in this use case:

- Forum Virium Helsinki, the innovation subsidiary of City of Helsinki, product owner of Helsinki's Digital Twin.
- Different departments of City of Helsinki. Namely KYMP (Division of Built Environment) and STARA (Construction and Maintenance Services).

Fintraffic, A subsidiary of Finnish Transport and Communications Agency, in charge of traffic control and traffic data services, e.g., operates the Finnish NAP.

### 4.1.3 Preconditions

This section should describe any conditions that must be true or activities that must be completed prior to executing the use case.

- Adequate number of technical interfaces for exchanging data (REST, JSON, …, …)
- Adequate features for metadata management (that support e.g., spatial data's specific requirements)
- Adequate features for access management (user rights, etc.)
- Adequate data timeline management
- Adequate data quality documentation

- Adequate data catalogue, searching and browsing features.
- (Cybersecurity requirements from digital twin perspective to be defined later)
- Optional:
  - Support relevant data catalogue standards, e.g., DCAT (Data Catalogue Vocabulary)
  - “Marketing features” for presenting the data.
  - Some elements from web shop systems, e.g., customers’ remarks about product (i.e., dataset)

### 4.1.4 Postconditions

This section should describe the state of the system at the conclusion of the use case.

The system shall be able to receive and host multitude of mobility-related data and be able to integrate with Helsinki’s Mobility Digital Twin in a level to be defined within the project.

### 4.1.5 Exceptions

This section should describe any errors that may result during use case execution and how the system will react or respond to those errors).

- Data set becomes expired, or maintenance is discontinued.
- Data set should be flagged accordingly.
- Data set is removed.
- Data should be marked as removed.
- Party responsible for the data is dissolved.
- Metadata should be updated, and data flagged accordingly.
- Changes to the publicity of the data due to data owner’s decisions, changes in legislation etc
- Metadata should be updated, and data flagged accordingly.
- Problems in real-time data streams
- Define the level of real-time system should be able to manage.
- There needs to be an alternative way to connect to the data stream.
- Problems in agreements or commercial arrangements between parties → Define the level and role of the system in relation to inter-party agreements or commercial arrangements.
- There must be a manual way to provide access to data.
- Privacy breach
- Ownership and usage rights must be dealt with within the system or build a way to interact securely with external management.
- Potentially include consent management / MyData -features for end user

Services to be offered Use Case may offer city data to external organizations through SEDIMARK data marketplace and can utilize external data sources to enhance its digital twin of mobility. These services benefit businesses, researchers, city planners, and other stakeholders by providing access to a wider range of data sources.

### SEDIMARK services:

For the use case related to 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 visualization 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.

### 4.1.6 Data to be generated / data models

The data pertaining to urban mobility can be divided into three categories:

1. Data describing the infrastructure.
2. Data depicting mobility events.
3. Data describing environmental and other conditions.

The datasets may be

- static, e.g., archived datasets, maps, statistics etc.
- dynamic, e.g., areas with occasional changes in e.g., size and other parameters.
- real-time or near-real-time*, e.g., traffic measurements such as volumes, speeds, routes, travel times.

The data will be in multitude of different kinds of formats. Helsinki is committed to using open standards, but some of the data (e.g., maintenance-related) may not yet have established standards for mobility domain and may be experimental.

Big share (> 50%) of the data is spatial data in some form, whether describing infrastructure or certain parameters of public space, or measurements or forecasts with spatial component.

### 4.1.7 KPI

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, the following key performance indicators (KPIs) have been established:

1. Data sharing from Mobility Digital Twin to Data Marketplace:
   - KPI: Number of datasets shared from the Mobility Digital Twin to the Data Marketplace
   - Target: 3 datasets
2. Data utilization from the marketplace:
   - KPI: Number of datasets obtained from the marketplace and applied in joint visualization in the Mobility Digital Twin
   - Target: 1 dataset

### 4.1.8 Flow

This section should describe all actions of the user and the expected system responses for planned normal execution of the use case.

1. Uploading / linking - the data will be made available for the marketplace, either by:
    a. uploading it to dedicated hosting,
    b. providing a URL to an external hosting (such as city's open data portal), or by
    c. providing a URL to the data stream (for a dynamic data source over API)
2. Metadata production, either by:
    a. inputting relevant metadata to the marketplace
    b. providing a link to existing metadata, e.g., in an external data-catalogue service
3. Metadata and/or data is provisioned over the marketplace, aggregating statistics of its use.
    a. Data is provided over UI.
4. Data is provided over standardized APIs.

### 4.1.9 Alternative Flows

The alternative flows in a use case typically arise from various options or exceptions that have been incorporated into the system, capable of modifying the primary flow. While not available at the moment, the following questions can be investigated to explore potential alternative flows or considerations:

- Real-time data: Is there a possibility to directly link to a data stream, enabling access to up-to-date information?
- Specific spatial data: Are there any unique spatial data requirements that necessitate a distinct flow or handling process?
- Stakeholder agreements: How should the management of necessary agreements between data owners/providers and data users/purchasers be handled? This may involve aspects such as usage rights, contracts, pricing, and payments, potentially leading to an unconventional process.
- Privacy-related concerns: How can privacy considerations, such as integrating MyData/Helsinki-profile, be addressed and incorporated into the flow?

## 4.2 Urban bike mobility planning in Santander

Santander City Council has made an important effort to evolve to "greener" commuting and mobility patterns. It has invested in building new bike lanes, but it has been identified the need to gather as much information of usage and bike mobility patterns to optimally define new bike lanes deployments and other ancillary infrastructures, such as parking spots. This information, adequately processed and analysed, will be useful for making informed decisions when taken as static information (black spots, degree of each section use, etc.). However, real-time information about streets incidents and other issues that may affect bikers are useful to be able to provide corrective or mitigation measures that reduce the impact.

The high-level outcome of the use case is quasi-real time information about bikers' positions that lead to a more dynamic bike mobility patterns and incidents at the bike lanes. In order to achieve this outcome, the use case will leverage different services provided by the SEDIMARK ecosystem.

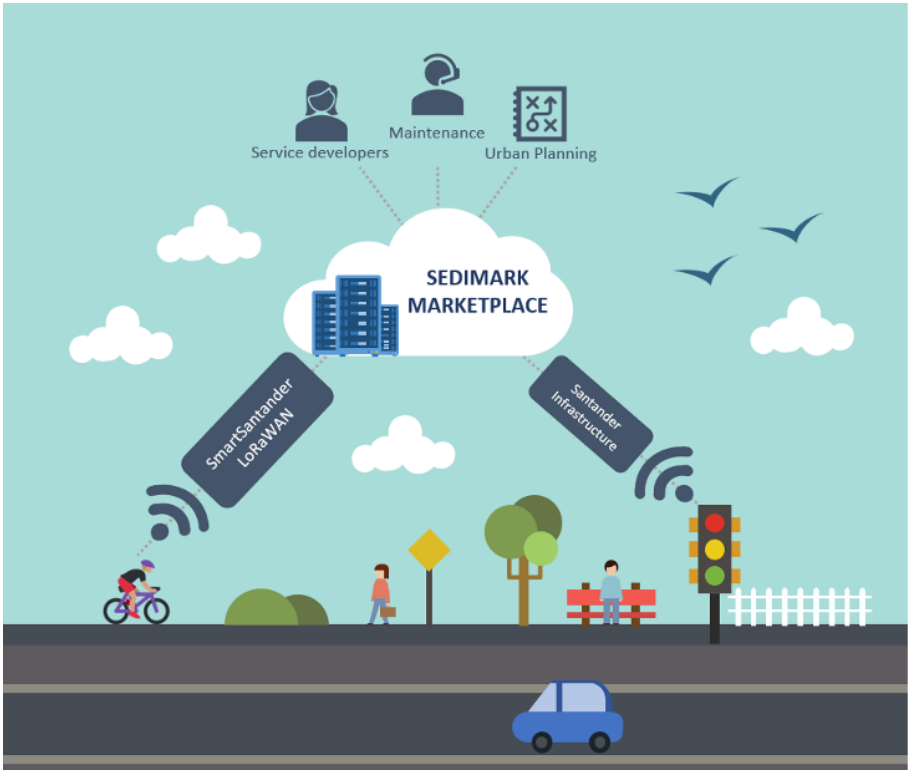


**Figure 3. Urban bike mobility planning high-level overview.**

### 4.2.1 Stakeholders involved

The popularity of bicycles in Santander has been growing in recent years, which means that they have to share space in the city with other means of transport. It is important to involve the main stakeholders to plan the actions. In the case of the Santander pilot at this stage of the project we have identified the following stakeholders:

- from the point of view of bike users: cycling associations and individuals in the city will be invited to be part of this pilot,
- from the point of view of the decision makers in the municipality: mobility councillor, director of mobility, municipal technicians and their supporting consultants will be involved,
- from the business & research perspective: entrepreneurs, bike-rental as well as bike-parking provider companies and research institutions will be engaged.

### 4.2.2 Current Status of the Use Case

The current status of the "Urban Bike Mobility Planning in Santander" use case involves several areas, as described below:

- Partial deployment of LoRaWAN infrastructure: A LoRaWAN (Low Range Wide Area Network) infrastructure has been deployed to provide LoRaWAN coverage in the targeted urban area of Santander. This infrastructure will support the communication between the intelligent tracking devices and the SEDIMARK platform.
- Additional LoRaWAN gateways deployment: To improve the coverage and ensure reliable communication between the intelligent tracking devices and the platform, we are planning to extend the LoRaWAN coverage in Santander. Thus, a thorough analysis of the existing bike lanes and their LoRaWAN coverage is being carried out, so as to identify suitable points to deploy additional LoRaWAN gateways.

- Design and development of the intelligent tracking devices: several LoRaWAN-capable devices are being tested so as to design and develop initial prototypes.

### 4.2.3 Preconditions

The following are the required conditions to carry out the Urban bike mobility planning use case:

- The intelligent tracking device will be deployed on both public and private bikes. Therefore, by using a shared bicycle system the use case will be able to identify a broader set of mobility patterns. When applicable, an Informed Consent will be requested to be signed by individuals to fulfil personal data protection regulation, in particular, GDPR. Proper anonymization and/or pseudonymization techniques will be used where applicable. GDPR stands for General Data Protection Regulation. It is a comprehensive data protection and privacy regulation implemented by the European Union (EU) to strengthen the rights of individuals and harmonize data protection laws across EU member states. The GDPR sets out guidelines and requirements for the collection, processing, and storage of personal data, aiming to ensure the protection and privacy of individuals' personal information. It imposes various obligations on organizations, such as obtaining consent for data processing, providing transparent privacy policies, implementing security measures, and enabling individuals to exercise their rights regarding their personal data. The GDPR applies to organizations that handle personal data of individuals residing in the EU, regardless of where the organization itself is based.
- The deployed intelligent tracking device which will serve as the basis for this urban mobility use case will transmit the generated data using LPWAN communication technologies. In particular, the use case aims to use LoRaWAN protocol for that purpose. Thus, an adequate network providing the appropriate LoRaWAN coverage level is needed.
- Data will need to be stored in an appropriate data repository.
- Have access to context information such as traffic condition, weather, environmental monitoring, and relevant GIS-related data.
- Available edge infrastructure to provide processing and storage capabilities.

### 4.2.4 Postconditions

The system will allow the publication of relevant data streams and events as part of the shared SEDIMARK marketplace as well as the automated analysis of such information to support the decision-making procedure within the mobility department at the city council.

### 4.2.5 Exceptions

- Intelligent Tracking Devices are damaged or malfunctioning.
  - If there is a halt in the data stream from the devices for a certain period of time, the maintainer will promptly contact the user to address any potential issues and replace the device if necessary.
- Failure related to the LoRaWAN network infrastructure.
  - In the event of a failure in the LoRaWAN network infrastructure (such as a power grid blackout or resource failure), appropriate measures will be taken to ensure the network functions properly. This may involve replacing or repairing any malfunctioning hardware components.

- Failure of services supporting the LoRaWAN network infrastructure and data gathering/processing from tracking devices.
  - Any issues related to the software or hardware of the supporting services will be promptly addressed. This may include tasks such as replacing server hardware or bug fixing in the software.
- Users want to quit from the pilot.
  - If users express their desire to withdraw from the pilot program, efforts will be made to engage new users and invite them to join the use case pilot.

### 4.2.6 Services to be offered

The following services are envisioned within the use case:

- Enabled edge data quality and analytics services to reduce the amount of information flowing through the network.
- Label or annotate multiple sources of heterogeneous data and combine them using AI.
- An interoperable set of the data, shared through a common SEDIMARK marketplace to learn and identify mobility patterns.
- Perform optimal route planning based on real usage information and emulating and inferring user acceptance of new routes in advance.
- Enable user/provider data control through the use of SEDIMARK decentralized marketplace while ensuring data anonymity.
- Generate information and adapters to enable the integration in a Digital Twin environment for green mobility in the city.

### 4.2.7 Data sources and data models

The use case will benefit from the data obtained from different sources, including IoT data streaming from deployed tracking devices in bikes and the existing IoT infrastructure in Santander. Besides, other data sources including static data will be also considered as part of the data harvesting for the use case. The following sections describe the data sources and are considered for the use case and the data models that will be used.

#### 4.2.7.1 Data sources

**Data to be generated from bike tracking devices:**

- Bike mobility patterns
  - Identifier
  - Geographical position
  - Battery
  - Timestamp
- Event detection (e.g., pothole, disturbance by traffic, disturbance by urban furniture, disturbance by pedestrian, etc.)
  - Identifier
  - Geographical position
  - Event (two predefined types)
  - AI-enabled events (e.g., fall)

**Data features for bike tracking devices:**

- **Data format**: tracking devices will send information through LoRa/LoRaWAN, which has a reduced data rate and impose limitations in the channel occupation with a 1% of duty cycle. Therefore, the overhead must be reduced as much as possible. To this end, it is foreseen the usage of Cayenne-LPP format for the payload sent by the tracking devices.
- Additional information that does not require near-real-time, will be stored in the device to be accessed afterwards, either manually or opportunistically, depending on the limitations of LoRaWAN.
- **Data report patterns:**
  - Periodical: geographical position to be sent periodically, either time-based (i.e., setting a minimum time between measurements) or location-based (i.e., after a significant variation in the position).
  - Event-based: if there is an event triggered by the user, data is sent even if the periodical conditions are not met. Besides, data delivered can be also triggered under certain opportunistic conditions (e.g., additional available network).

**Additional data (from existing infrastructure or available data sources):**

- Static
  - Bike lanes
- Streaming
  - Traffic volume
  - Environmental parameters (e.g., air quality)
  - Weather information
  - Bicycle parking occupancy
  - Bike usage from bike-rental services (e.g., bike-rental spots occupancy)

#### 4.2.7.2 Data models

Relevant Smart Data Models [34] for the data sources envisioned in this use case. Besides, new data sources might be needed (e.g., events information from tracking devices).

- Transportation [35]
  - Traffic Flow Observed
  - Fleet Vehicle Status
  - Vehicle
  - Bike Lane
  - Bike Hire Docking Station
- Environment [36]
  - Air Quality Observed
  - Environment Observed
  - Urban Mobility
- Weather [37]

### 4.2.8 KPIs

During the execution of the “Urban bike mobility planning in Santander” use case the following KPIs are envisioned:

- Increase the LoRaWAN coverage in the areas with existing bike lanes in Santander (>80% of the city coverage).
- Involvement of at least 10 users with 10 intelligent tracking devices to act as data sources. This number could be increased if it is decided to rotate the devices between different groups of users.
- Data provision of the following parameters per device:
  - Geolocation
  - Battery
  - Timestamp
  - Events
- Devices can support at least two predefined user-generated events.
- Number of incidents or events reported.

### 4.2.9 Use Case flow.

IoT devices will be installed on bicycles to track the route of the users that are part of the pilot with the aim of obtaining useful insights of the bike mobility patterns and infrastructure use in the city. In addition, the user will be able, through the IoT device, to report incidents or black spots during the cycling route.

With this information it will be possible e.g., to check if the most used routes coincide with the existing bike lanes, new needs for bike lanes can be identified, check if the black spots marked by the users of the pilot coincide with the black spots already identified, etc.

**User: rider**

1. Mounts the device in the bike with the help of the ‘maintainer’.
2. Check that the device has enough battery.
   a. (If needed) Plugs the device to the charging accessory.
3. Start a ride.
4. When the user finds an incidence on the streets, identifies its type and pushes the corresponding button in the device.
5. Access to personal space on a web platform to check if the information has been properly registered.

**User: maintainer**

1. Maintain the devices plugged to the charging accessory to assure it has the full charge before supplying it to a rider.
2. Mounts the device in the bike. If the bike belongs to a private user help with mounting and basic instructions are provided.
3. Monitor, by means of a control web platform and an alert system, that all the devices in use have their basic parameters (idle, level charge, etc.) in good condition and performance state.

4. Receives from private users broken or not properly working devices and substitutes them if required making the necessary adjustment in the monitor and control platform.

**User: urban planner**

1. Access to the exploitation platform with its role and view the general dashboard showing the basic information: number of devices on duty, number of routes (daily, global...) recorded.
2. Exports the information using filters for establishing the analysis period, zone, type of information (routes, incidents, etc.).
3. Imports the previous information to a GIS system, a digital twin or for publishing it in the open data portal and/or data marketplace. This will allow to join the information to other sources and be able to make decisions, i.e., improve bike lanes solving of incidents, change period of maintenance of bike infrastructures.

### 4.2.10 Alternative flow

**User: rider**

- If the user notices that the device does not work as expected, he/she will report a system failure through the established channel to receive the corresponding instructions.
- If the platform detects a failure in the registered information, the user reports through a web platform the issue or make the allowed changes (i.e., change the type of incident or remove it).
- The user notices that no data has been registered in the platform.
  - The user will manually upload the data from the device through the corresponding mechanisms (e.g., extracting the log file from the device and uploading it to a web platform.).

**User: maintainer**

- If the maintainer detects any of the exceptions defined in Section 4.2.5, it proceeds to solve it by following the defined procedure:
  - In case of LoRaWAN network infrastructure or IT infrastructure failure:
    - Replace or repair the malfunctioning equipment.
  - In case of data gathering/processing services error:
    - Restart the services and report any bug with a detailed description through the appropriate channel (e.g., GitHub Issues).
  - In case the Intelligent Tracking Device is not working properly
    - Contact the user through the appropriate communication channel and repair or substitute the device.
    - Extract the data from the internal storage device and manually upload it to the platform.

**User: urban planner**

- If data is unavailable, the urban planner will contact to the maintainer through the established communication channel.

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

### 4.3.1 Description

The accuracy of energy forecasting is critical for utility providers to run day-to-day operations efficiently. However, energy forecasting is challenging because of the impact of the dynamic demand and seasonal weather changes. Utility providers must balance supply and demand, purchasing energy to meet peak demand or selling excess capacity in the spot market. Demand forecasting has become challenging as well with renewable energy resources that do not always produce stable power. Furthermore, the rise of electric vehicle purchases and the unknown nature of when vehicle owners want to charge them at home contributes greatly to the instability. Improvements to forecasting enable the utility providers to foresee the structure of more cost-effective future contracts. Residential customers can also use these results to predict future energy consumption, calculate energy costs and move to more efficient pricing plans.

In many countries, power is provided in competitive retail markets. Consumers have a choice in buying electricity and can switch providers if they receive high energy bills or have bad customer experience. MYT can reduce customer churn by improving customer service and proactively reaching out with future bill spend alerts. These alerts can be based on accurately predicting electricity consumption and analysing the customer conduct in terms of complaints. MYT will provide the data and define the context of the use case. WINGS will support Protergia to develop the AI models and the integration with the SEDIMARK data and services marketplace. The stakeholders include utility providers, customers, research and university audience, internal & external sales network, energy advisors and, generally, the whole indirect channel of the Power and Gas BU of MYT, Protergia.

### 4.3.2 Current Status of the UC

MYT has developed an AI factory. This constitutes a platform with algorithmic structures and procedures for data analytics services for retail customers, which will be extended with new tools and connected to the marketplace. MYT has their own infrastructure for collecting and analysing energy related and customer data and has already conducted an internal investigation on data usage and collection. Furthermore, MYT currently explores models that have been created and tested for similar endeavours to SEDIMARK including but not being limited to other research projects. MYT may also exploit the myprotergia app, which enables customers to monitor their energy consumption. [38]

### 4.3.3 Actors / Stakeholders

MYT will provide the data and define the context of the use case. WINGS will support Protergia to develop the AI models and the integration with the SEDIMARK data and services marketplace.

The stakeholders include but are not limited to utility providers, customers, research and university audience, internal & external sales network, energy advisors and, generally, the whole indirect channel of BU Protergia.

### 4.3.4 Services to be Offered.

MYT will offer two main services based on the two subcases that will be examined.

1. Energy consumption prediction & clustering: The first service will offer predictions about energy consumption in residential customers though analysis of sparse data. As such, common energy consumption patterns will be discovered, and different regions will be clustered based on consumption. Prediction, profile extraction and clustering will use the SEDIMARK AI and data management tools. In case of energy consumption, the role of weather data (temperature, humidity, etc.) is particularly important and for this the SEDIMARK open data enabler will be utilized to exploit Open Data for weather. The quality and integrity of energy data is important for efficient predictions. Normally, the gathered data are “dirty” and need to be carefully cleaned. The processed data and the metadata will be shared in the SEDIMARK marketplace as raw data and as AI service models open to be used by any interested consumer. The data for this kind of analysis and, therefore, will be used are a) Residential energy consumption b) Zip codes, c) Weather data (external), d) Residential size.
2. Customer segmentation & churn prediction: The second service that will be offered is essentially comprised of two different AI prediction models which will analyse customer sales and behaviour on a geospatial level. The AI models will focus on the prediction of customer segmentation in different regions via postal code as well as customer churn in different regions via postal code. Customer segmentation & churn prediction will be used privately by MYT (BU Protergia) for the purpose of efficiently managing existing and potential business customers, analysing the complaints in customer support (CS), avoiding losses of existing customers (churn reduction) and boosting customers’ loyalty or increasing the market share in regions via informing local sales network. The role of these tasks is to pass the information to the sales network, energy advisors and, generally, the whole indirect channel of BU Protergia to gain a better view of the local market for which they are responsible. The following data will be anonymized and protected via the SEDIMARK privacy and security tools, as well as be processed through the data management and AI tool: a) Zip codes, b) Complaints of Customers, c) Churn of Customers, d) Delays of Payments, e) Amount of Energy Payment. The results of the AI analytics and tools will be fed in the SEDIMARK marketplace to be shared only with authorized clients, exploiting the SEDIMARK privacy and access control tools. In general, a combination of heterogeneous data will be processed in to produce the aforementioned results. Interoperability is of vital importance and the use case will utilize the SEDIMARK semantic models. For both subcases, specific AI models are in the process of being developed, whose results will be shared through the SEDIMARK marketplace together with the raw data either to an open network of stakeholders (subcase 1) or to the authorized internal & external stakeholders (subcase 2). SEDIMARK will provide the secure decentralized data and services marketplace, enabling Protergia to share raw data, AI services and insights derived by the AI models with trust, security, and privacy over the DLT. Distributed computing can assist in training the AI models with low latency and under preservation of security and privacy.

### 4.3.5 Preconditions and Postconditions

One crucial prerequisite to consider in the SEDIMARK project is the initial access to the data warehouse, which serves as a central repository for diverse data sources. Additionally, granting access to the repository becomes necessary at a later stage. Furthermore, it is important to engage in a discussion to define the Integrated Development Environment (IDE). The IDE enables software developers to perform tasks such as editing, building, testing, and packaging software, ensuring efficient software development processes within the SEDIMARK project. At the conclusion of the case, a successful platform to securely analyse both

consumers' energy behaviour and customer conduct in terms of complaints as well as churn will exist. The anonymized results will be forwarded to the marketplace.

### 4.3.6 Data to be Generated & Shared

As mentioned above, the energy-oriented data for the energy consumption prediction and clustering will be public and anonymized except for the ZIP Codes. This means that the only data that will have public status are the weather data, the residential size and any residential consumption related data. However, the customer-oriented data regarding segmentation and churn prediction will be private and anonymized. All data processed will concern a specific time range. The sources of the data include the internal corporate data warehouse as well as some external APIs (for weather data for instance). As far as specific data structures and formatting is concerned, any data will be provided in a column based .csv format.

### 4.3.7 Flows

As an overview of the whole system, the expected flow of any user is essentially to interact with the platform and have read-only access to the aforementioned metrics and results. Any user will be able to choose what data to see and interact with.

### 4.3.8 Alternative Flows

Regarding alternative flows, the user is expected to contact administrators of the platform primarily to verify absence of their data. Afterwards, if the administrators cannot draw conclusions regarding data availability, then they prompt the users to contact the respective data owners.

### 4.3.9 Exceptions

Since the whole use case of MYT is predicated on the existence of data, the first exception that must be raised is the case of the data owner not being willing to share the data anymore. In this case, the corresponding dataset is promptly deleted from the platform and the user is informed about the deletion. Secondly, the case the data has not been cleared properly is considered; in such a case, the user is informed about the cleaning process not being successful and they are promptly informed to contact an administrator to handle the issue. Finally, regarding the AI models that will be hosted on the platform, any conflicts regarding data governance will be handled by the data provider and the models will adjust accordingly when data is missing.

### 4.3.10 KPIs

- The vast majority (>95%) of the processing of the data to be done at the edge instead on centralized servers.
- Significant (>50%) reduction of communication load due to data being stored locally.
- Increased accuracy of AI-based techniques (e.g., outlier/error/noise detection) for data cleaning, data augmentation and data management should lead to a massive reduction of dirty data (>90%). A reduction in operational costs is expected as well.
- Improvements in data interoperability due to annotation and feature engineering.
- Improvements in data availability by >25%, improved data privacy, maintain near zero data tampering.

- Churn prediction metrics; improvements in areas such as customer churn rate, customer retention rate, customer health score, customer engagement rate, customer satisfaction score among others.

## 4.4 Valuation and commercialisation of water data in France

In the context of climate change water is a critical resource that must be managed very carefully. The ecosystem of water management involves many different actors, each having a different responsibility and their own datasets which may be of value for other stakeholders. Currently, these datasets are not or poorly shared. Allowing different actors to use data of others in an interoperable way may stimulate the innovation in water management by allowing new public services and better political decisions. The shared data will concern many aspects of the water for a territory like:

- Resource protection (ground water, rivers, flow and quality …)
- Drinking water production (quantities, infrastructure, …)
- Water distribution and usages (drinking, irrigation, leisure activities, …)

This use case will have a special focus on the specific aspects of data quality to ensure that they are trusted, to encourage their reuse, to provide value and new services. For this, the data quality services developed in SEDIMARK will be deployed and experimented in this use case, like validation, curation, and evaluation systems.

### 4.4.1 Stakeholders involved

Regarding the valuation and commercialisation of water data use case, the stakeholders that were identified as this stage of the project are the following:

Data producers, which includes:

- Public authorities, to provide general basic data about the territory.
- Water management public and private institutions, to provide specialized water data.
- Algorithms / IA model and predictions providers

Data consumers, which includes:

- Public authorities to discover new insights for water management and risk mitigation.
- Start-ups or businesses to provide new services with provided data.
- Data producers to consult the quality of their own data.
- Administrators, such as Water data platform manager

### 4.4.2 Current status of the UC

Currently data providers have their data in their own information system. Data from different providers are not linked and are not interoperable. However, few data sharing based use cases already exists and have been identified:

- Prediction of mountain stream flows, link with water stocks in reservoirs
- Water quality in reservoirs
- Irrigation management
- Infrastructure management and maintenance

Indicators on the quality of these data are not assessed and some tools are needed to do so.

### 4.4.3 Preconditions

In order to successfully carry out the use case, local actors need to agree to share data and algorithms provider to implement their algorithm on the platform. Moreover, IoT devices deployment is required to get complementary datasets and perform far edge processing capabilities.

### 4.4.4 Postconditions

The SEDIMARK marketplace is expected to be fully deployed and running, along with the deployment of other use cases data, services and algorithms within it. Datasets are registered in the marketplace and accessible by the user. The quality assessment of the datasets is available for all datasets and is accessible by the user with the use of quality indicators.

Users of the SEDIMARK marketplace are expected to be able to perform multiple actions depending on their role.

The producers are able to add new datasets, grant access to dataset they own to other users, account for the usage of their data and monitor it.

The consumers are able to search or discover dataset, request access to a dataset, consult some quality indicators on a dataset and manage the quality of the dataset on its own (add tag, filter or notification).

### 4.4.5 Exceptions

- Data owner no longer wants to share the data on the platform.
  - Dataset should be removed accordingly.
- Quality assessment has failed for a dataset.
  - Flag concerned dataset and send alert to system administrator.
- Data from real-time data flow is no longer received.
  - Send alert to system administrator and concerned provider to fix the issue as quickly as possible.
- Contact with IoT devices is lost.
  - Send alert to system administrator and local contact point to fix the issue as quickly as possible.

### 4.4.6 Services to be offered

The marketplace will propose multiple services aiming at making the use of data as practical as possible. The main services are the followings:

- Tools to integrate heterogeneous data sources.
- Data validation, semantic enrichment, and transformation services
- Security and authorization policies
- Edge services for data quality and alerting systems.
- Maps, dashboards, and alerts
- Basic data marketplace services

### 4.4.7 Data sources and data models

The data sources to be registered in the SEDIMARK marketplace that were identified are the followings:

- Territory
  - GIS data
  - Location of major infrastructures (dam, pumping station, …)
  - Location of major measurement stations (flow, weather, …)
  - Directory of the territory (key persons and institutions)
- Water distribution
  - Water management and distribution infrastructure
  - Infrastructure monitoring data (pressure)
  - Water consumptions (anonymized)
- Water resources
  - Stocks in water reservoir
  - River, stream flows (discharge + water level)
  - Piezometry measurements
- Weather information
  - Weather observation from external provider
  - Weather observation from station
- Irrigation management
  - Irrigation programs
  - Soil moisture sensors measurements

Accordingly, relevant data models must be used for all mentioned data source and also for the modelisation and prediction results from deployed algorithm.

### 4.4.8 KPIs

Several KPIs have been identified, some related directly to the performance of this use case, and some related to some other part of the project, that will be impacted by this use case.

- Improve data quality by minimizing ratio of errors to data to <2%, maximise data completeness by minimizing the number of empty/unused values to zero, improve validity of the data to 99%.
- Improve usability of data by 60%, due to advanced feature engineering process and additional metadata.
- Reduce energy costs for storing and processing data by 20%.
- Improve AI model frugality score by 30%.
- Reduce floating point operations of AI mechanisms by >30% with <5% reduction in performance.
- Improve discoverability of the data by 30% due to the efficient semantic annotation.
- Improve the quality of data annotations by >20%.
- Improve conformity to certification and standards.

- Number of connected open data platforms to the marketplace >5.
- Increase in the subjective trust for the data marketplace by 50%.
- Minimization of possibility of data leakage due to unauthorized access to data.
- Transaction confirmation rate > 80%.

### 4.4.9 Use Case flow

#### Marketplace standard flow

- Administrator registers a new user as either a producer, a consumer, or both.
- User registers a new dataset.
- User registers a new data stream.
- User searches the marketplace for a dataset (title, tags, quality criteria, …) or a processing algorithm.
- Consumer requests access to a dataset or a processing algorithm.
- Producer and consumer users establish a contract for dataset or algorithm access.
- Producer gives permission to access a dataset.
- Accounting balance.
- User can use the dataset or algorithm.

#### Data exploitation flows

- User set up an alert based on a dataset.
- User accesses a dashboard or a map.
- User uses an algorithm on a dataset to generate a new dataset.

#### Data quality flows

- Indicators of data quality in the marketplace search results are generated.
- User requests an assessment of the quality of a dataset.
- User requests a data stream to be pre-processed to get quality indicators updated in real time. The pre-processing may be a chain of services provided by the platform.

### 4.4.10 Alternative flow

Sending alert in case of exception

- Exception concerning a dataset/ service is detected by the system.
- System consults the metadata of the dataset/service to retrieve a diffusion list to alert.
- The alert is sent.

# 5 Initial trust and security analysis

This section provides a preliminary security analysis focusing on the domain of the whole SEDIMARK Marketplace architecture. The goal is to analyse under the security point-of-view the preliminary design of the system for the only purpose of shedding light on the Trust and Security key points of attention to be considered in the next agile process of designing and developing the marketplace.

In this context, the properties of Confidentiality, Integrity, Availability, Authenticity and Non-repudiation of the assets are considered. Moreover, considerations related to privacy of actors as well as related authorizations policies to access the assets extend this analysis.

First, the section 5.1 defines and set the scope of such analysis and propose an overview of the related security properties to be provided.

Then, in section 5.2, each security dimension is analysed with respect to the Uses Cases, previously defined in Section 4.

Finally, in Section 5.3, further considerations detail the extension of Trust, in possible scenarios that could be considered during the design and future development of the system.

## 5.1 Scope and Context

Nowadays, data security is more relevant than ever. The obvious reasons are the protection against unauthorized access to the data and in general the protection against cyber-attacks.

Other reasons for data security involve the specific content of the data itself, such as the preservation of Intellectual Property (IP).

Finally, data security might be enforced from States and Governments for compliance with existing norms and regulations.

Considering the Uses Case descriptions for the Marketplace, provided in Section 4, it is obvious that data exchange is at the core of the Marketplace.

Each SEDIMARK scenario does require the proper high-level mechanisms of security, aiming at preserving the so-called CIA triad (Confidential, Integrity, Availability) (see Section 5.2) for the assets involved for the main purpose of building "Trust in Assets", a key feature for the success of the marketplace.

It is possible to define a domain in which the assets and the operations performed by the various actors belonging to the system are trusted. Such domain, hereby defined as **SEDIMARK Trust Domain**, is shown in Fig. 4.

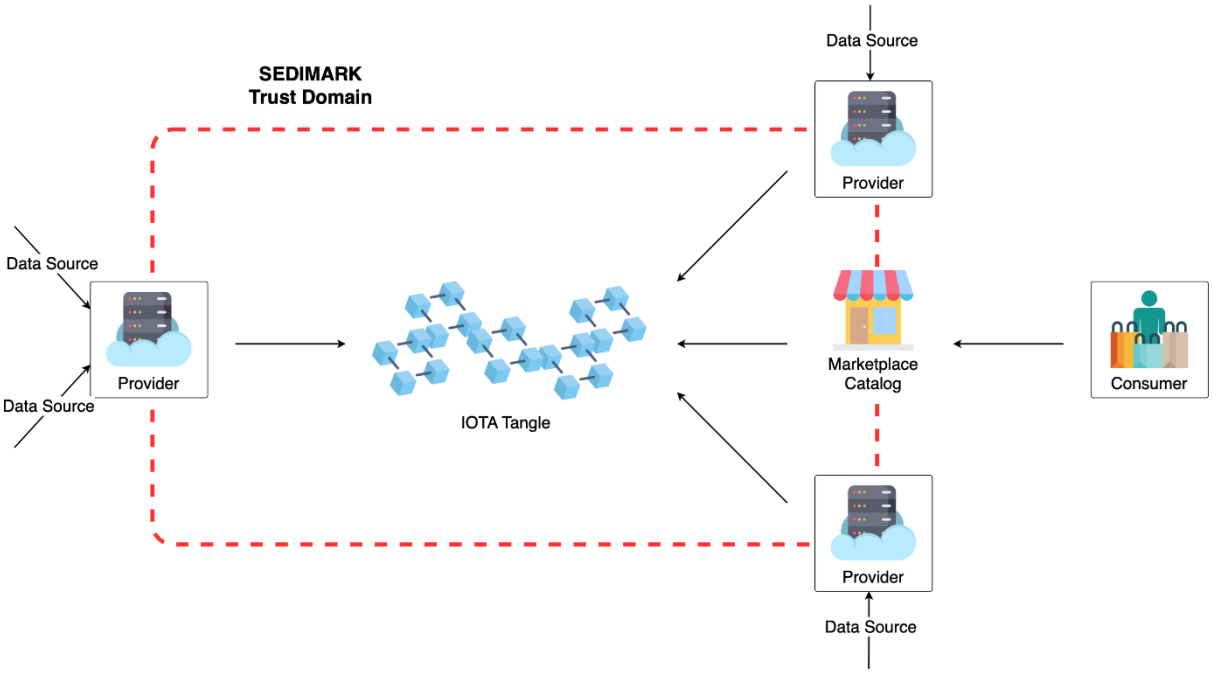


Figure 4. SEDIMARK actors and Trust Domain.

The main actors of the marketplace are the users, residing at the edge of the Trust Domain. The two types of users identified so far play the roles of the *Provider* and the *Consumer*. The Providers own the assets (*e.g.*, dataset, AI model, service, etc.) and can sell or offer such asset to the Consumers.

In the specific case of a data asset, the Provider can gather this content from a *Data Source*.

The Consumer can access the asset after browsing the *Marketplace Catalogue*, which shows the discovered assets offered by the Providers. Most of the described interaction happens through the *IOTA Tangle* [39], which is a particular type of Distributed Ledger Technology (DLT).

The SEDIMARK Trust Domain is outlined with red dashed square in Fig. 4.

It guarantees the so-called *"Trust"* (see Section 6.2) in the assets offered through the marketplace by means of a DLT-based approach.

The Trust is built with two components:

- **Trust in Data**: trust is derived from the verification of the Data Origin (*i.e.*, by authenticating the data source) and Data Integrity (*i.e.*, a proper cryptographic mechanism to assert the integrity of the data).
- **Trusted Interactions:** trust is derived with proper operations performed (e.g., buy/sell datasets, services offering, etc.) among the actors involved in the SEDIMARK Trust Domain and also the Consumers.

The previous components of trust are rooted in the management of proper metadata over the IOTA Tangle.

Such metadata have not to be confused with the one of a dataset. Indeed, it is related to the *secure interactions* taking place within the SEDIMARK Trust Domain.

In particular, it comprises:

- **The Trust metadata**, employed for ascertain data origin and data integrity, and

- **the Offering metadata**, needed to build the trusted catalogue and to provide secure pointers to Consumers.

Datasets are the main assets of the SEDIMARK Trust Domain and are given by Data Sources, which are considered external to the SEDIMARK Trust Domain. The exchange of datasets between the actors involved occurs Off-DLT, i.e., off the Tangle.

Also, such exchange must be protected to maintain the desired security properties. The protection of this interaction requires the establishment of secure communication channels according to best security practices.

The resulting interactions take place at two different layers, which are depicted in Fig. 5. On the one hand, the IOTA Tangle is used as a means of transportation for metadata (both, Trust and Offering). On the other hand, the widespread internet connectivity is employed for the exchange of the assets. In both cases, secure communication channels are required to protect the information.

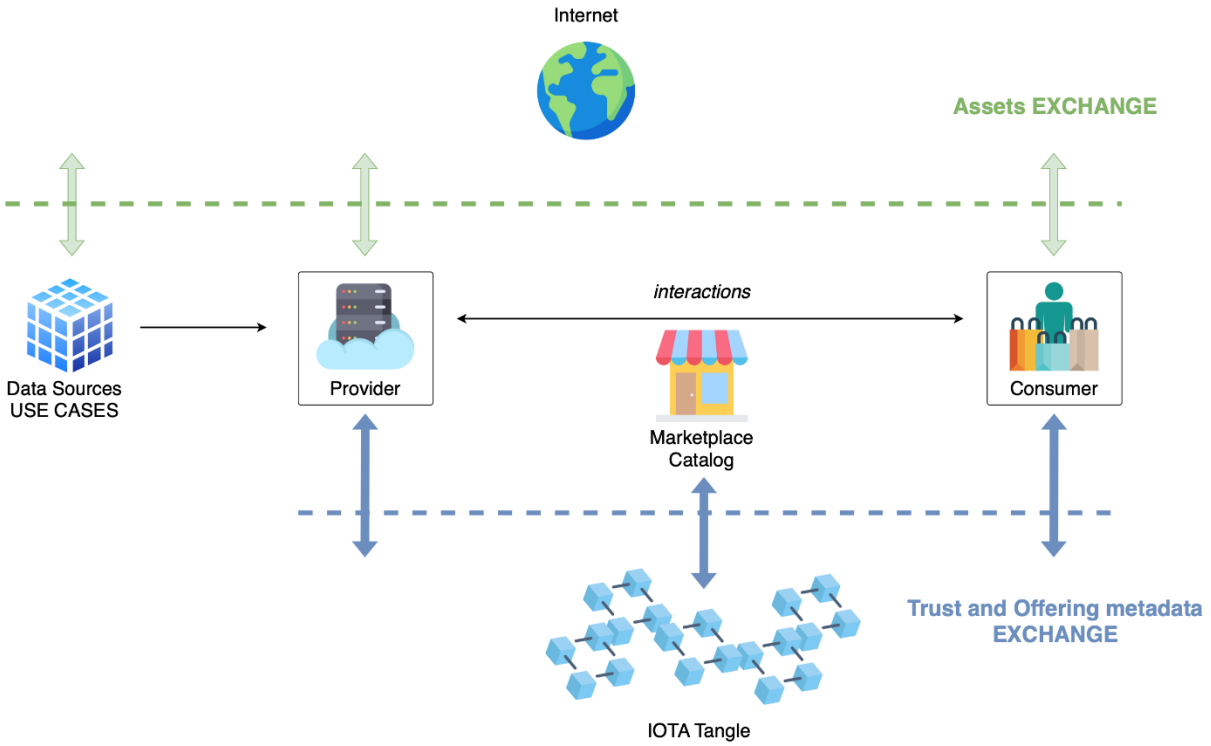


**Figure 5. Layers for assets exchange.**

## 5.2 Trust and security

The concept of Trust revolves around the assurance and confidence towards something.

For example, it can be considered as the confidence that an object or a process will function or behave in an expected way.

In the context of the SEDIMARK Marketplace, the Trust lies in its assets. For example, a Consumer achieves trust in the SEDIMARK Marketplace because the assets that he/she is buying are those he/she expects.

To achieve this level of assurance, it is necessary to fulfil the following properties:

- **Asset Origin:** the real origin of an asset is maintained, assured and verified from the source to the destination. Usually, the path could be from Producer to Consumer, however the origin must be safeguarded also among other actors within the Trust Domain.
- **Asset Integrity:** the integrity of the asset is maintained, assured and verified throughout the various interactions. By mean of integrity verification it is possible to trigger appropriate mechanisms to counteracts deliberate attacks, as well as accidental transmission errors.

Both Asset Origin and Integrity are maintained by means of proper Trust Metadata recorded onto the IOTA Tangle to build the “Trust in Assets”.

The Trust properties contribute to secure the SEDIMARK Marketplace, however they are not sufficient alone.

Indeed, usually, security involves the properties of Confidentiality, Integrity and Availability.

- Confidentiality aims at protecting an asset from access.
- Integrity copes with an unwanted alteration of an asset.
- Availability deals with the possibility to access an asset when it is needed.

Additionally, also the properties of Authenticity and non-reputability contribute to the security of the whole SEDIMARK Marketplace.

These security properties, mapped onto the UCs and the Marketplace assets, produce the following security needs:

- Secure communications by means of authenticated endpoints - implies the following properties:
  - Mutual authentication by means of trustworthy and unforgeable credentials: in order to set up a communication channel, the communicating parties must first identify each other.
  - Verification of integrity of the information exchanged: assets transferred together with the additional Trust metadata which can be used in combination with proper cryptographic mechanism to verify the integrity.
  - Confidentiality of communications: in general, every communication channel among different parties belonging to the marketplace shall be set up resorting to existing secure communication protocols.
- Confidentiality of critical data stored into the overall marketplace: the assets belonging to the Producers must be protected by the Producers themselves, according to their own security requirements and privacy policies.
- Availability of Trust Metadata: such information, residing onto the IOTA Tangle, is available according to the immutability features provided by DLTs.
- Availability of assets: such information, belonging to a Producer and residing on the Producer' premises, strictly depends on the availability of the internal storage system(s) of the Producer itself.
- Authorization of access to assets according to predetermined policies: such policies are defined the Producer for each of its assets. Proper authorization mechanisms shall be guaranteed.

Logging and Audit trails functionalities: realized by means of proper Trust, Offering and other metadata.

## 5.3 Privacy

Another important aspect related to the security domain is the concept of privacy, that is closely connected to the confidentiality property.

Privacy concerns the adherence to various legal and non-legal norms regarding the right to have something private or in a private way, *e.g.*, in the European context, privacy is often understood as compliance with European data protection regulations regarding the right to private life.

In the context of SEDIMARK privacy can be partitioned into two aspects.

The first, focuses on the users of the Marketplace, namely the Consumers and the Producers.

This distinction requires an Identity Management (IdM) system to be able to differentiate the actors based on the role they play. Hence, the users need to be provided with a digital identity to perform any operation within the Marketplace.

The Self-Sovereign Identity (SSI) paradigm [40] is adopted in SEDIMARK to let the actors to build their own identities without the need to rely on external identity providers. This option allows also to provide a decentralized identity to the actors of the Marketplace, by relying on the IOTA Tangle as a Root of Trust for identity data.

The identified requirements for the SSI framework with proper privacy management for the agents (i.e., the internal components of the SEDIMARK Trust Domain) are the following:

- Decentralized Identifier (DID) [41]: a globally unique identity designed to verify a subject.
- Verifiable Credentials (VCs) [42]: a VC represents the same information held in a physical credential. However, it embeds also cryptographic signatures that makes it more tamper-evident and more trustworthy than the physical counterpart.
- Proper revocation mechanism: such a mechanism shall be based not only on the expiration time, but also on specific events coming from the interested parties.
- Adoption of best security practices for data protection related to user into SEDIMARK agents (Issuer, Holder and Verifier).

The SSI model could also be enhanced by adopting Zero-Knowledge Proof (ZKP). ZKP could be used in combination with VCs to provide fine grain privacy management for the SEDIMARK users. However, ZKP adoption will be further investigated in the context of SEDIMARK, but it is not a requirement.

The second aspect of privacy is related to the usage of assets in the Marketplace. The access to assets needs to be regulated and based on specific policies of authorization. The SSI framework aids the implementation of such policies on a per-User base through VCs. Considering the SSI model, the identified and necessary mechanisms concerning the usage of datasets and services are the following:

- Shall be based on VC claims combined with other proofs, *e.g.,* proof of purchase.
- Shall allow the fruition of certain/specific assets only to certain/specific users, according to customizable authorization policies.

## 5.4 Further considerations

In principle, the Trust, described in Section 6.2, must be maintained *end-to-end*. The starting endpoint considered in the previous subsections is the Producer, which belongs to the trusted and secure SEDIMARK domain. The datasets, however, are gathered from the Data Source, which instead is outside the SEDIMARK Trust Domain.

From a security point-of-view, to achieve a complete form of Trust on the data assets, it is necessary to protect also the communications between Data Source and Data Provider. The protection of these communications also refers to the confidentiality, integrity, availability and authenticity properties previously described. However, securing the interactions between Data Source and Data Provider is outside of the scope of SEDIMARK Trust Domain and it is responsibility of the related UC's providers.

# 6 Functional and non-functional requirements (with use case flag)

## 6.1 Non-functional - system architecture requirements

There are many definitions of Non-Functional Requirements (NFRs) in the literature, but the majority of works consider as NFRs the requirements that represent quality attributes of a system and are differentiated from the Functional Requirements that are the designed operations of the system and are related with specific functionalities. Thus, NFRs are basically requirements that can be considered as the overall goals of a system, the constraints and the attributes and cannot be easily measured in terms of functionalities. As such, we present them in textual format in contrast with the Functional Requirements that are presented in more details in the Table format that was presented in Section 2. An overview of the non-functional requirements for improved readability is given in Table 3.

**Table 3. Non-functional requirements.**

| Identifier | Title | Priority | Req. Level |
|---|---|---|---|
| Req-NF-01 | Decentralisation | H | REQ |
| Req-NF-02 | Security, Privacy, Trust | H | REQ |
| Req-NF-03 | Interoperability | H | REQ |
| Req-NF-04 | Data availability and quality | H | REQ |
| Req-NF-05 | Intelligence | H | REQ |
| Req-NF-06 | Energy efficiency | H | REQ |
| Req-NF-07 | Resilience and Reliability | H | REQ |
| Req-NF-08 | Scalability | H | REQ |
| Req-NF-09 | Openness, Extensibility | H | REQ |
| Req-NF-10 | Usability | H | REQ |
| Req-NF-11 | Maintainability | H | REQ |
| Req-NF-12 | Adaptivity to data types and fast processing | M | REC |
| Req-NF-13 | Reusability | H | REQ |
| Req-NF-14 | Flexibility | H | REQ |

SEDIMARK aims to provide a secure and intelligent decentralised marketplace that will facilitate the exchange of high-quality data, services and models in an interoperable way. This overall goal of the system already reveals the five main NFRs for SEDIMARK. Decentralisation **(NFR#1)** is a main requirement for SEDIMARK, since the project aims to avoid centralised and cloud servers that are playing the roles of data lakes and can be single point of failures or attacks and promotes that providers and users should keep their data locally at their own

servers to have full control over who has access to them. In this respect, SEDIMARK is built on top of the IOTA distributed ledger technology that supports not only decentralisation, but also Security, Privacy and Trust **(NFR#2).** SEDIMARK aims to provide trusted transactions between consumers and providers, decentralised identities and guarantees for data integrity. Also, keeping data locally at your own servers without having to upload and store your datasets on central or cloud servers supports data privacy. SEDIMARK will also enable trust in the datasets and the offerings overall through additional Trust-related Metadata added to the offerings.

Interoperability **(NFR#3)** is also one of the key concepts that support the SEDIMARK system, since the goal is to annotate and enrich the data semantically using a common ontology so that data can be easily indexed, searched, used and reused within the system in a homogeneous way or across system to other data spaces. Ensuring data availability and data quality **(NFR#4)** will also help data to be used and reused not only for building services and applications upon them, but also for building machine learning and artificial intelligence models in an easy way, without the users or researchers having to spend huge amounts of time to clean the data, translate them from various formats and annotate the data in a way that they can be used for training models. In this respect, Intelligence **(NFR#5)** is also one of the key pillars of SEDIMARK, since the system will provide easy tools for building and training energy efficient machine learning (ML) models on top of high-quality interoperable data and mainly also in a decentralised way, employing techniques such as decentralised Federated Learning or Gossip Learning. SEDIMARK also focuses on model Interoperability, which supports both NFR#3 and NFR#5. As mentioned above, SEDIMARK also supports energy efficiency **(NFR#6),** through both lightweight data management and machine learning mechanisms and through optimisations that reduce the energy consumption of training and running ML models.

The SEDIMARK architecture should also ensure system reliability and resilience **(NFR#7)** with respect to failure and for fault, failure and attack recovery. The SEDIMARK system should ensure minimum number of failures so that it provides a reliable service and connectivity to the users. Having the IOTA DLT as a backbone for the interconnectivity of the nodes and the transactions supports these requirements, as well as the requirement for scalability **(NFR#8),** since there's no central node to become a bottleneck for communications or decision making, SEDIMARK will be scalable and be able to support large numbers of users and providers communicating and exchanging datasets, models and services. The implemented communication mechanisms should ensure low communication overhead, ensuring fast decision making and high performance.

The SEDIMARK architecture should also be Open and Extensible **(NFR#9)** so that new providers, open data platforms or other data spaces can be easily integrated with SEDIMARK and make their offerings available through the marketplace. In this respect, SEDIMARK will provide an open data enabler to provide open APIs so that third parties can easily be connected to the system and exchange offerings, without having to use the full functionality of the system. Additionally, the architecture will provide standardised APIs so that it can be easily extensible with new features, new data management modules and new services to improve all supported functionalities.

The SEDIMARK system should also be user friendly, meeting the requirements for usability **(NFR#10)** and maintainability **(NFR#11)** with its configuration being optimised as much as possible, requiring minimum user intervention. This will ensure that it will become widely usable not only by technology experts, but also by simple users that can exploit the offerings, without having to go through long lists of installation and configuration options. Additionally, the user

interfaces that will be included in the system, as well as the marketplace interface will have to be intuitive, easy to use and provide full functionalities in as simple way as possible.

Considering that SEDIMARK aims to support both offline (static) and streaming (dynamic) datasets, the functionalities for data processing, cleaning, annotation and enrichment will have to be adaptive to the data type and will have to be very fast **(NFR#12)** so that any processing doesn't cause delays into the data streaming. This must be taken into account especially when data streaming is related with critical applications that may have very strict requirements with respect to timeliness.

Finally, the SEDIMARK architecture and its modules should be designed in such a way so that they're easily reusable **(NFR#13)** to other projects, systems or platforms and also flexible **(NFR#14)** enough so that if requirements change, the modules can be adapted without significant extra effort or delays.

## 6.2 Security, privacy and trust

The following security requirements aim at ensuring secure interactions within the SEDIMARK Trust Domain among the involved parties and in protecting the assets exchanged in the SEDIMARK Marketplace.

| Section | Description |
|---|---|
| Id | REQ-SEC-01 |
| Short Name | Authentication of users |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Users (Providers and Consumers) must be able to authenticate in the SEDIMARK Trust domain |
| Additional Information | This requirement focuses on the authentication of the interacting users within the SEDIMARK domain. Employing the functionalities of the marketplace is dependent on the process of authentication. |
| Purpose/Goal | Only users belonging to the SEDIMARK Trust Domain are able to interact with other users and employ the services offered by the Marketplace. |
| Validation Criteria | - |
| Constraints: | No |
| Dependencies: | Req-UI-01, Req-UI-03 |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |

| Section | Description |
|---|---|
| Pros/Cons | Pros: only authenticated user can access services<br>Cons: overhead due to setup and verification for authentication process |
| Category | Security |

| Section | Description |
|---|---|
| Id | REQ-SEC-02 |
| Short Name | Authorization policies of assets |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Each asset must be characterized by an authorization policy decided by the respective Provider. Use of the Asset is dependent on such policies |
| Additional Information | This requirement focuses on the authorization of the usage of the assets within the SEDIMARK Marketplace. Each asset must be characterized with the desired authorization policies decided by the respective provider. The fruition of the asset is subordinated to the defined authorization policies. |
| Purpose/Goal | The target of this requirement is to enforce a form of control over the usage of the asset according to the will of the respective provider. |
| Validation Criteria | - |
| Constraints: | No |
| Dependencies: | Req-UI-02, Req-UI-04, Req-UI-05 |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: Provider has full control on its assets<br>Cons: Need to define a predefined set of authorization policies in the Trust Metadata |
| Category | Security |

| Section | Description |
|---|---|
| Id | REQ-SEC-03 |
| Short Name | Origin of assets |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Assets origin and integrity must be maintained in the SEDIMARK Trust Domain |
| Additional Information | This requirement focuses on maintaining the origin of the assets, when exchanged among the various interacting parties within the SEDIMARK Trust Domain. |
| Purpose/Goal | The target of this requirements is to guarantee that the assets exchanged are not modified during any exchange. Moreover, it guarantees that the real provider (origin) of an asset is maintained, assured and verified from the source to the destination. This guarantee however refers only within the SEDIMARK Trust Domain. |
| Validation Criteria | - |
| Constraints: | No |
| Dependencies: | REQ-SEC-04 |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: final users have the guarantee that the assets they are acquiring is exactly the one they want.<br>Cons: needs to define the respective information in the Trust Metadata. |
| Category | Security |

| Section | Description |
|---|---|
| Id | REQ-SEC-04 |
| Short Name | Trust Metadata on Distributed Ledger |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Trust Metadata must be written onto the Distributed Ledger |

| Section | Description |
|---|---|
| Additional Information | All the information required to guarantee the security properties (e.g., digest, signatures, policies) must be written on the Distributed Ledger. |
| Purpose/Goal | Writing information on the Distributed Ledger enables the trustworthiness of Trust Metadata needed for securing the interactions taking place within the SEDIMARK Trust Domain. |
| Validation Criteria | - |
| Constraints: | No |
| Dependencies: | Req-SEC-02, Req-SEC-03 |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: Data is verifiable by anyone<br>Cons: overhead of libraries required to interface with Distributed Ledger |
| Category | Security, Decentralization |

| Section | Description |
|---|---|
| Id | REQ-SEC-05 |
| Short Name | Decentralized provisioning |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Assets provisioning must employ a decentralised approach |
| Additional Information | N/A |
| Purpose/Goal | The purpose is to reduce complexity on the asset exchange stage, avoiding the involvement of additional parties in the procedure which are not controlled by either the consumer or the provider. |
| Validation Criteria | N/A |
| Constraints: | Too many possible scenarios to support, as the set of all existing assets is unbounded by nature. |
| Dependencies: | N/A |
| Conflicts: | No |

| Section | Description |
|---|---|
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: interactions to exchange assets are carried out between consumer and providers following a P2P approach, so no additional parties/domains involved on the communication<br>Cons: assessment of asset exchange cannot be easily done |
| Category | Decentralization |

| Section | Description |
|---|---|
| Id | REQ-SEC-06 |
| Short Name | Secure channel of the assets |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Assets must be transferred over secure communication channels (e.g., through TLS) |
| Additional Information | N/A |
| Purpose/Goal | The goal of this requirement is to enable an additional security layer on the asset exchange stage. By using secure communication channels, only participants are able to have access to the exchanged information. |
| Validation Criteria | N/A |
| Constraints: | There might be different requirements to enable secure communication based on the specific scenario. |
| Dependencies: | N/A |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: better security<br>Cons: higher complexity |
| Category | Security |

## 6.3 Data quality management and data processing

Data quality is one of the main pillars of SEDIMARK, which aims to provide the necessary tools to data providers and consumers so that they can curate their data according to their

preferences and improve the quality, so that they can both use it internally but also share high quality data through the SEDIMARK marketplace. As such, one key requirement for SEDIMARK is that it must provide tools for data cleaning and curation (REQ-DP-01).

| Section | Description |
|---|---|
| Id | REQ-DP-01 |
| Short Name | Data cleaning tools |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | SEDIMARK must provide a complete toolset for data cleaning, curation and quality improvement. The system must be able to flag records which are found to be problematic by the data processing pipeline. In this way the user will have the ability to define how they should be handled (removed or just flagged). Similarly, the system should provide tools to identify outliers or anomalies in the data. Problematic records may include: (i) Duplicates, (ii) Noise/Errors, (iii) Anomalies/outliers. |
| Additional Information | This requirement includes the following sub-requirements:<br>• must provide ability to identify and remove or flag noise or errors in the data before shared.<br>• must provide ability to remove duplicate records before sharing.<br>• bad records/entries must be clearly flagged as such.<br>• target for how to handle duplicates or errors has to be clearly defined |
| Purpose/Goal | SEDIMARK aims to promote the sharing of clean data, so providing tools for curating and cleaning data is of utmost importance. The flagging of problematic records allows the user to inspect the data quality and can also be used for calculating metrics regarding the quality of a shared dataset. |
| Validation Criteria | Data quality metrics compared against ground truth |
| Constraints: | Requires specific input format for the data. |
| Dependencies: | REQ-INT-01 ; REQ-INT-02 ; REQ-INT-03 ; REQ-INT-05 ; |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: transparency regarding data quality; improved data quality; reducing dataset number of records<br>Cons: Flagging increases the size of data in Random-Access Memory (RAM)/on disk |

| Section | Description |
|---|---|
| Category | Data processing |

SEDIMARK aims to provide support for sharing not only static offline datasets, but also data streams, i.e., originating from Internet of Things devices. In this respect, users will be able to register to a stream and SEDIMARK will need to handle the transmission of the streaming data. Considering the fact that SEDIMARK promotes data quality, the SEDIMARK data curation pipeline should be able to handle streaming data too (REQ-DP-02). This means that the data curation modules should be flexible to handle both static and streaming datasets, paying significant attention on the delay requirements of those streaming datasets.

| Section | Description |
|---|---|
| Id | REQ-DP-02 |
| Short Name | Flexibility to handling both static and streaming data. |
| Type | Functional |
| Priority | Medium |
| Requirement Level | Recommended |
| Description | Data curation modules should be flexible to handle different types of data and both static and streaming data. The tools within the data pipeline need to be adaptive to handle the strict requirements of data streaming and lightweight to be executed fast, without causing delays. |
| Additional Information | Define what data are static.<br>Define what data are streaming data.<br>Define the difference in how the different types of data must be processed. |
| Purpose/Goal | SEDIMARK aims to function for data providers who have both static and streaming assets. SEDIMARK should provide the functionality to pre-process, transform and clean both kind of assets, both in bulk and on the fly. |
| Validation Criteria | Cleaning accuracy; running time; curation delay; |
| Constraints: | Streaming application requirements. |
| Dependencies: | Req-DP-04; Req-DP-05; Req-DP-06; |
| Conflicts: | ML model accuracy and model size. |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: Extends SEDIMARK functionality to more kinds of data assets.<br>Cons: Some data processing modules might require a dual implementation for some of their functionality. |

| Section | Description |
|---|---|
| Category | Data processing |

The data curation pipeline in SEDIMARK offers users and providers the essential tools to clean and process their data. However, it does not impose any specific requirements regarding which modules should be executed or how they should be executed. The flexibility is given to users and providers to choose and customize the modules based on their specific needs and preferences SEDIMARK will allow the users of the curation pipeline to use it in the way they want, by selecting which modules they want to execute and allowing them to configure them to their preference. Thus, the data curation pipeline should be modular and configurable (REQ-DP-03). However, the pipeline should also allow for minimum user intervention in cases where users don't want to do any configuration themselves, so that even non-technical people can use it.

| Section | Description |
|---|---|
| Id | REQ-DP-03 |
| Short Name | Data processing pipeline configuration flexibility |
| Type | Functional |
| Priority | Medium |
| Requirement Level | Recommended |
| Description | The data curation pipeline should be configurable as per what modules will be run and how each module will run. The tools within the data pipeline may need to be figured to work optimally given e.g., the provider's domain and the provider's preferences. As such, the provider should be able to configure the tools within the pipeline. |
| Additional Information | SEDIMARK should provide a simple workflow e.g., UI or scripting protocol for defining how the modules will be run. |
| Purpose/Goal | For increased user satisfaction and considering that there's no single solution/configuration that fits all preferences and all domains, SEDIMARK aims to allow data providers to change the data curation configuration to achieve improved performance on curating their data. Additionally, for non-expert a very efficient default configuration per domain should exist. |
| Validation Criteria | User satisfaction |
| Constraints: | Data curation modules should become highly configurable but have some default configuration that allows them to run efficiently per domain. |
| Dependencies: | Req-DP-05; Req-DP-06; application QoS requirements |
| Conflicts: | No. |
| Relevant Use Cases | All use cases. |

| Section | Description |
|---|---|
| Pros/Cons | Pros: Gives data providers the ability to find an optimal transformation of their data; improves user satisfaction; improves accuracy/efficiency per domain<br>Cons: Some data quality metrics might become opaque when data providers can potentially configure how they are calculated; might require technical knowledge or domain knowledge for the configuration |
| Category | Data processing |

Sharing high quality data through the SEDIMARK marketplace requires to understand which data have been processed (if) and how, in a common, standardized and unified way, so that the users will be able to easily select the data they want for different quality criteria. As such, data shared through SEDIMARK should include data quality metrics or indicators (REQ-DP-04) that will provide statistics regarding the dataset, as well as additional data quality metrics that will be defined within the project.

| Section | Description |
|---|---|
| Id | REQ-DP-04 |
| Short Name | Data quality indicators |
| Type | Functional |
| Priority | Medium |
| Requirement Level | Recommended |
| Description | Data should include quality indicators or metrics. The data quality pipeline should produce a number of summary statistics about the data, including for instance basic profiling, and summaries of e.g., the number of outliers, errors, duplicates and missing values identified. Moder advanced ML-based quality metrics may also be attached to the datasets to help ML researchers to identify the best datasets to be used for training ML models. |
| Additional Information | SEDIMARK should define in detail the data quality statistics and metrics that will be used for assessing the quality of the datasets. |
| Purpose/Goal | SEDIMARK aims to improve the quality of data within its ecosystem. Data quality statistics and metrics should help both for external evaluation and allow consumers to better appraise the quality of data that they are purchasing. |
| Validation Criteria | Data quality metrics should be evaluated in the use cases against a ground truth, or domain expertise. |
| Constraints: | Data quality metrics ground truths might be different per domain; |
| Dependencies: | REQ-INT-02; REQ-INT-03; |
| Conflicts: | No. |

| Section | Description |
|---|---|
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: Allows consumers to better appraise data before purchase.<br>Cons: Some data quality metrics might be hard to define and might be open to manipulation by the data providers; increases dataset size due to additional annotation required. |
| Category | Data processing |

Data cleaning mechanisms may require significant computational resources to run, especially when the data to be cleaned are of large size (big data). Additionally, depending on the mechanism that will be exploited, in cases that it is a machine learning based mechanism, it might require a minimum set of hardware requirements to run efficiently. However, for making data cleaning to be run in a distributed manner, providers may choose to employ low-end devices at the sources of the data generation to clean the data to also reduce communication costs. Thus, the SEDIMARK data cleaning mechanisms should be adaptive to the technical capabilities of the devices that run them, and they should also be able to be run on low-end devices.

| Section | Description |
|---|---|
| Id | REQ-DP-05 |
| Short Name | Adaptability of data cleaning mechanisms |
| Type | Functional |
| Priority | Medium |
| Requirement Level | Recommended |
| Description | Data cleaning mechanisms should be adaptable to the technical capabilities of the device that runs them |
| Additional Information | N/A |
| Purpose/Goal | To be able to run the data cleaning mechanisms of the SEDIMARK platform on any device without major modification, data cleaning mechanisms should be adaptive to the hardware requirements and the proper mechanism with the correct configuration should be selected at any given time. |
| Validation Criteria | Performance metrics |
| Constraints: | Requires more effort on standardizing data cleaning mechanisms |
| Dependencies: | Data cleaning modules; device hardware requirements |
| Conflicts: | Performance of data cleaning mechanisms; |
| Relevant Use Cases | All use cases |

| Section | Description |
|---|---|
| Pros/Cons | Pros: Minimal modification or configuration effort when running on any device<br>Cons: difficulty of creating generic data cleaning module; mechanisms on low-end devices may have lower performance |
| Category | Data processing |

Data cleaning is not an easy task to perform and usually requires significant domain knowledge in order to be able to identify which are the outliers or which data values are faulty. In this respect, SEDIMARK should allow data providers to configure the cleaning mechanisms as much as possible providing some ground truth for the data cleaning mechanisms in order to allow them to take proper decisions.

| Section | Description |
|---|---|
| Id | REQ-DP-06 |
| Short Name | Ground truth for data quality metrics |
| Type | Functional |
| Priority | Low |
| Requirement Level | Optional |
| Description | Data quality metrics need to have ground truth for proper calculation based on domain knowledge and should be parameterizable by the users |
| Additional Information | Data quality metrics can have user-set explicit constraints per data feature to compute the quality.<br>Users should be able to parameterize data quality modules to set explicit constraints per data feature to compute the quality.<br>Data cleaning modules require domain knowledge to set thresholds for i.e., identifying noise or outliers. |
| Purpose/Goal | To be able to provide proper calculation based on domain knowledge for the data quality metrics of the SEDIMARK platform, data cleaning modules should allow the addition of ground truth. This will help the modules to identify properly outliers, faults, noise, etc. |
| Validation Criteria | Performance metrics |
| Constraints: | Need to have some domain knowledge; users might need to have technical expertise |
| Dependencies: | Data quality modules |
| Conflicts: | N/A |
| Relevant Use Cases | All use cases |

| Section | Description |
|---|---|
| Pros/Cons | Pros: More accurate metrics<br>Cons: requires human intervention |
| Category | Data processing |

As continuation of the requirement above considering that every provider might want to handle duplicates or errors differently based on their preference, the cleaning modules should be configurable to allow the users to define themselves what constitutes a duplicate or an error.

| Section | Description |
|---|---|
| Id | REQ-DP-07 |
| Short Name | Data cleaning modules extendable definitions |
| Type | Functional |
| Priority | Medium |
| Requirement Level | Recommended |
| Description | Data cleaning modules should be extendable to different definitions of duplicates and errors |
| Additional Information | N/A |
| Purpose/Goal | To be able to adapt data cleaning modules to different target so that they are more versatile |
| Validation Criteria | Versatility of data cleaning modules; Performance metrics |
| Constraints: | Need to take into account unknown potential definitions |
| Dependencies: | Data cleaning modules |
| Conflicts: | N/A |
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: increase usefulness of data cleaning modules<br>Cons: there will always be new unpredicted definitions to adapt to; requires human intervention |
| Category | Data processing |

SEDIMARK aims to provide a simplified interface for the data providers so that they can execute and monitor the progress of the data processing pipeline. Considering that not all users will have high technical knowledge about the execution of the various data processing tasks, SEDIMARK aims to automate the pipeline so that all the tasks can run easily with minimum human intervention. Techniques for Auto AI and Auto ML might also be employed in this respect, aiming to provide maximum performance with minimum human intervention.

| Section | Description |
|---|---|
| Id | REQ-DP-08 |
| Short Name | Automate execution of tasks |
| Type | Functional |
| Priority | Medium |
| Requirement Level | Recommended |
| Description | Automate execution of various tasks where possible, to minimise human intervention, allowing also non-technical users to be able to execute tasks. |
| Additional Information | N/A |
| Purpose/Goal | To minimise human intervention and reduce time spent on configuration |
| Validation Criteria | User satisfaction; performance metrics; |
| Constraints: | Need to keep track of possible error of failure in automatic tasks |
| Dependencies: | Hardware capabilities; data cleaning modules |
| Conflicts: | REQ-DP-06; REQ-DP-07 |
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: reduce need of human intervention<br>Cons: may increase risk of error of failure in task execution |
| Category | Data processing |

For proper training of Machine Learning models, dataset size, but also dataset characteristics like diversity, bias and fairness are very important. Datasets that are of very small size may not have enough data points to generalize a ML model. Datasets that are unbalanced towards some labels may also include bias and may produce models that aren't fair to the whole population of the labels. Thus, one requirement for SEDIMARK is to provide datasets that are balanced and fair and of "enough" size so that users can train good ML models on top of them. To achieve this, SEDIMARK will work on data augmentation mechanisms for providing synthetic data points.

| Section | Description |
|---|---|
| Id | REQ-DP-09 |
| Short Name | Dataset augmentation |
| Type | Functional |
| Priority | Low |
| Requirement Level | Optional |

| Section | Description |
|---|---|
| Description | Datasets might need to be augmented with synthetic data before shared to remove biases and produce more balanced datasets. |
| Additional Information | N/A |
| Purpose/Goal | To increase size of small dataset and make them more relevant for training ML models on top of them. |
| Validation Criteria | User satisfaction; ML training accuracy; data quality metrics; |
| Constraints: | Need to make sure the augmentation is relevant and doesn't worsen the data quality metrics. |
| Dependencies: | Data quality metrics; |
| Conflicts: | N/A |
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: increase value of small dataset<br>Cons: potential denaturation of the initial dataset; might increase processing time; |
| Category | Data processing |

To be aligned with privacy regulations, the datasets within SEDIMARK should be properly anonymised. This should be done at the provider level before they are processed by SEDIMARK. In rare cases that this is not done, SEDIMARK may provide some simple mechanisms for the providers to perform simple anonymisation.

| Section | Description |
|---|---|
| Id | REQ-DP-10 |
| Type | Functional |
| Priority | High |
| Requirement Level | Recommended |
| Description | Private information in assets must be anonymised during pre-processing in the Data Owners Domain |
| Additional Information | N/A |
| Purpose/Goal | To protect private information of the user and be compliant with data protection policies |
| Validation Criteria | Quality of the anonymisation |

| Section | Description |
|---|---|
| Constraints: | Need to develop an anonymisation module that will be performant and compliant with SEDIMARK data format |
| Dependencies: | Data anonymisation module |
| Conflicts: | N/A |
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: protection of user personal data<br>Cons: potential loss of some useful data if too much data is anonymized |
| Category | Data processing |

## 6.4 AI

Intelligence is one of the main pillars of SEDIMARK. The system will support the training of efficient ML models either locally at the user's or provider's premises or in a distributed way. For training ML models effectively, though, data have to be converted to common formats and have to be cleaned, without errors, duplicates, noise or missing fields and have to be of specific size, otherwise the model performance will be very poor (REQ-ML-01).

| Section | Description |
|---|---|
| Id | REQ-ML-01 |
| Short Name | Model input data cleaning and formatting |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Data need to be cleaned, in a standard common format and of a specific size for proper training of the models |
| Additional Information | This requirement has the following sub-requirements:<br>• Data need to be structured for proper training of the models.<br>• Data need to be properly annotated and labelled for supervised models.<br>• Data need to be cleaned for training efficient models.<br>• Data need to be of a specific size for achieving high accuracy |
| Purpose/Goal | Dirty data or data in wrong format impact severely the performance of ML models, slowing down training and reducing accuracy. |
| Validation Criteria | model performance (accuracy, etc.) |
| Constraints: | Data size |

| Section | Description |
|---|---|
| Dependencies: | REQ-DP-01 ; REQ-DP-04 ; REQ-DP-05 ; REQ-INT-02 ; REQ-INT-03 ; |
| Conflicts: | No. |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: model performance improvement.<br>Cons: requires data processing and annotation, thus increasing complexity; |
| Category | Artificial Intelligence |

The system also aims to provide tools for training and running machine learning models in a decentralised way (REQ-ML-02) to cater for user privacy, allowing users to retain their data locally, without exchanging raw or sensitive data, but only model weights. To ensure that decentralised training will be done in an efficient way, only trusted peers should be allowed to participate in the training to avoid issues of model poisoning and to improve the security of the training process (REQ-ML-03). We assume that trusted peers are fellow participants in the training process (namely other users who have similar data and want to train the same model) who have been certified as trusted by some trust mechanism.

| Section | Description |
|---|---|
| Id | REQ-ML-02 |
| Short Name | Decentralised ML |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Decentralised training of ML models should be employed for increased privacy without exchanging raw data. |
| Additional Information | This requirement has the following sub-requirements:<br>• Decentralised training of ML models should be employed for increased privacy.<br>• Decentralised ML models must be privacy preserving not disclosing user sensitive data.<br>• Decentralised ML models must be trained without exchanging raw data |
| Purpose/Goal | SEDIMARK is a decentralised system, where users keep their data locally and there is no central or cloud server that stores data. As such, it's not possible to train AI models centrally. Thus, SEDIMARK should provide tools to enable decentralised training of models in a privacy preserving way, so that users don't exchange data, but only model weights or gradients. |

| Section | Description |
|---|---|
| Validation Criteria | User privacy; model performance |
| Constraints: | Enough similar nodes should be participating in the training process for efficient training |
| Dependencies: | REQ-ML-03 |
| Conflicts: | No |
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: user privacy; efficient training of models in a decentralised system.<br>Cons: could require increased communication in cases of asynchronous implementation |
| Category | Artificial Intelligence |

| Section | Description |
|---|---|
| Id | REQ-ML-03 |
| Short Name | Trusted participation in decentralised training |
| Type | Functional |
| Priority | Medium |
| Requirement Level | Recommended |
| Description | Decentralised training should be done only using trusted peers and with secure connections for improved performance not allowing tampering with data or man in the middle attacks. |
| Additional Information | N/A |
| Purpose/Goal | Training a ML model in a decentralised way requires finding related peers that can participate in the training process, having similar datasets and training the same model. These peers should be trusted to avoid having malicious nodes tampering with the model. Also, connections for exchanging the model weights need to be secured so that no third parties can identify what is being exchanged, which could in some cases reveal some user information. |
| Validation Criteria | Model performance; secure connectivity; |
| Constraints: | Not being able to verify the trustworthiness of a peer. |
| Dependencies: | Req-SEC-01; Req-SEC-04; Req-UI-01; |
| Conflicts: | N/A |

| Section | Description |
|---|---|
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: enhances model reliability and robustness; protects users' privacy.<br>Cons: restricts the possible number of peers |
| Category | Artificial Intelligence |

Additionally, considering the fact that there are many platforms for training models, SEDIMARK should provide tools to convert models to various formats, so that models can be trained or executed at various types and on multiple frameworks, which will also allow them to be trained jointly (REQ-ML-04).

| Section | Description |
|---|---|
| Id | REQ-ML-04 |
| Short Name | Models' agnostic to platforms |
| Type | Functional |
| Priority | Medium |
| Requirement Level | Recommended |
| Description | ML models should be able to be trained regardless of the underlying platform the peers use even on low-spec machines for maximum interoperability and efficiency. |
| Additional Information | This requirement has the following sub-requirements:<br>• ML models should be runnable on low-spec machines too for edge processing.<br>• ML models should be able to be trained in a distributed manner regardless of the underlying platform the peers use |
| Purpose/Goal | Decentralised training requires that users train their models locally and then exchange model weights for jointly learning a global model. Ensuring that all SEDIMARK users will use the exact ML platform for training the model and the exact same machines is unrealistic, so, SEDIMARK models should be agnostic to underlying platforms and SEDIMARK should provide tools to support models to run on machines of various capabilities and on various platforms. |
| Validation Criteria | Model performance; model interoperability; model training efficiency; energy efficiency |
| Constraints: | Model size should be kept small for running on low end machines. |
| Dependencies: | REQ-ML-01; REQ-ML-02 |
| Conflicts: | N/A |
| Relevant Use Cases | All Use cases |

| Section | Description |
|---|---|
| Pros/Cons | Pros: extends the number of possible peers to participate in decentralised training; allows for energy efficiency.<br>Cons: could restrict model performance; |
| Category | Artificial Intelligence |

| Section | Description |
|---|---|
| Id | REQ-ML-05 |
| Short Name | Models' persistence mechanisms |
| Type | Functional |
| Priority | High |
| Requirement Level | Recommended |
| Description | Model persistence: models should be persisted in a centralized/decentralized manner to be easily accessible and sharable. |
| Additional Information | This requirement has the following sub-requirements:<br>• Models should be easy to save and to load and shared between third party modules |
| Purpose/Goal | Can organise better the ML flow |
| Validation Criteria | Models are accessible, the solution provides sufficient storage |
| Constraints: | N/A |
| Dependencies: | REQ-ML-01; REQ-ML-02 |
| Conflicts: | N/A |
| Relevant Use Cases | All Use cases |
| Pros/Cons | Pros: better organisation<br>Cons: effort to maintain |
| Category | Artificial Intelligence |

| Section | Description |
|---|---|
| Id | REQ-ML-06 |
| Short Name | Event generation from pattern extraction |
| Type | Functional |
| Priority | High |

| Section | Description |
|---|---|
| Requirement Level | Recommended |
| Description | Generate events detected through pattern extraction techniques. |
| Additional Information | Services that will provide added value to the datasets from the providers will include the generating of events of interest to consumers. This will involve distributed inference over datasets and streams. |
| Purpose/Goal | Semantic enrichment of datasets for knowledge extraction. |
| Validation Criteria | Successful detection of events of interest. |
| Constraints: | Training data and initial domain knowledge required prior. |
| Dependencies: | Req-ML-01; Req-ML-04; Req-ML-05; Model inference |
| Conflicts: | None. |
| Relevant Use Cases | All |
| Pros/Cons | Pros: alleviates burden of processing for consumers that require direct access to events for their application scenario.<br>Cons: Issues with quality of data could be interpreted as events. |
| Category | AI-ML |

Decentralised training of ML models in a real-world scenario is usually a difficult task because of the many uncertainties that are included in the process. For example, users may go offline or may become unusable from time to time, other users may have older versions of the model and updating those old versions, it's not easy to synchronise all the users to send updates at the same moment, etc. To mitigate these issues, researchers have worked not only on synchronous versions of decentralised ML training, but also on asynchronous ones. SEDIMARK should build a decentralised ML training system that caters for both scenarios, able to run both synchronous and asynchronous training depending on the task at hand.

| Section | Description |
|---|---|
| Id | REQ-ML-07 |
| Short Name | Synchronous and asynchronous training of models |
| Type | Functional |
| Priority | Low |
| Requirement Level | Optional |
| Description | SEDIMARK should provide a decentralised training framework that supports training of ML models in both synchronous and asynchronous ways. |

| Section | Description |
|---|---|
| Additional Information | N/A |
| Purpose/Goal | Considering that it's not always possible to synchronise model updates between remote users, asynchronous model training may also be required to ensure that model updates will be done correctly, without affecting the model performance. |
| Validation Criteria | Model performance; model training efficiency; |
| Constraints: | In asynchronous training, model performance may degrade if updates come from very older versions of the model or from users that aren't representing properly the population. ML training toolbox will need to decide which version to use. |
| Dependencies: | REQ-ML-02; REQ-ML-03; REQ-ML-04; |
| Conflicts: | N/A |
| Relevant Use Cases | All Use cases |
| Pros/Cons | Pros: asynchronous training doesn't require synchronisation and allows parallel updates; synchronous training requires fewer rounds of training.<br>Cons: asynchronous training could restrict model performance and parameters may drift away; synchronous training requires synchronisation of the updates, which increases complexity; |
| Category | Artificial Intelligence |

As a user-targeted marketplace, SEDIMARK also aims to provide user-friendly recommendations for data, services, models and offerings in general, considering user preferences and past activities. As such, various AI requirements are directly related with supporting a decentralised recommendation platform and for providing efficient recommendations to end users. As such, the SEDIMARK recommendation platform should be able to compute user preferences based on past interactions with the system (REQ-RS-01) so that the recommender is able to understand what a user might like and provide only relevant items for recommendation. Additionally, the SEDIMARK recommender should be able to know locally at the user's server either the full set of the available offerings including their metadata, or a filtered set based on user defined filters, as well as statistics about the offering, such as trending ones, most downloaded, highly rated, etc. (REQ-RS-02). One of the main problems in Recommender Systems is the "cold start" problem, when the system doesn't have any history for the user and can't build a preference profile. To be able to provide recommendations even in this case, the SEDIMARK recommender should also be capable of addressing this cold start problem (REQ-RS-03). Finally, since SEDIMARK is a decentralised system, the recommendation platform should also be trained and executed in a decentralised way, not leaking user information and should be run completely locally at the user's computer (REQ-RS-04).

| Section | Description |
|---|---|
| Id | REQ-RS-01 |
| Short Name | User profiling |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Recommender system needs access to user profile logs and user interaction history to be able to effectively compute user preferences and build a user profile. |
| Additional Information | This requirement has the following sub-requirements:<br>• Recommender models need to be able to compute user profiles from their history.<br>• Recommender system requires to have access to user profile information and historical info for their interaction with datasets. |
| Purpose/Goal | Recommender systems provide personalised services to user based on their preferences. To be able to compute user preferences, a recommender requires access to user interactions with the system at hand. The more information the recommender has the better will be the user profile and the more personalised the recommendations will be. |
| Validation Criteria | Model accuracy; |
| Constraints: | Access to user logs for tracking the user interactions with the system. |
| Dependencies: | Req-UI-01 ; Req-UI-07 ; Req-UI-08 ; |
| Conflicts: | N/A; |
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: improved personalisation<br>Cons: requires user action logging; |
| Category | Artificial Intelligence |

| Section | Description |
|---|---|
| Id | REQ-RS-02 |
| Short Name | Rich item information |
| Type | Functional |
| Priority | High |

| Section | Description |
|---|---|
| Requirement Level | Required |
| Description | Recommender models need to know the available item set and have as many features, metadata, and statistics as possible |
| Additional Information | This requirement has the following sub-requirements:<br>• Recommender models need to know the available item set to find the best item(s) to recommend to users.<br>• Recommendation platform must have as many features about the users and items as possible.<br>• Recommendation platform must be capable of exploiting dataset metadata.<br>• Recommendation platform should have access to marketplace statistics.<br>• Recommender system should consider item value/cost when making recommendations |
| Purpose/Goal | To provide accurate recommendations to users and to be trained efficiently, a recommender system should have as much information about the candidate items as possible. This helps to construct rich item features that will be used to describe the items. Thus, item metadata, statistics, etc. should be available to the recommender system. |
| Validation Criteria | Model accuracy |
| Constraints: | N/A |
| Dependencies: | Req-P&D-01; Req-P&D-03; REQ-INT-02; REQ-INT-03; Req-UI-05; Req-UI-07; Req-UI-08 |
| Conflicts: | N/A |
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: improves model accuracy; improves personalisation<br>Cons: model complexity; metadata complexity |
| Category | Artificial Intelligence |

| Section | Description |
|---|---|
| Id | REQ-RS-03 |
| Short Name | Decentralised Recommender system |
| Type | Functional |
| Priority | High |

| Section | Description |
|---|---|
| Requirement Level | Required |
| Description | Recommendation platform must be trained and run in a decentralised way, not leaking user private information. |
| Additional Information | N/A |
| Purpose/Goal | SEDIMARK is a decentralised system, so training the recommender system model should also be done in a decentralised way, considering also that users run the recommender locally and all of their data including their interactions are also stored locally. |
| Validation Criteria | Model privacy |
| Constraints: | Not enough peers that participate in training; not enough item features; |
| Dependencies: | REQ-ML-02 ; REQ-RS-01 ; RES-RS-02 |
| Conflicts: | N/A |
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: user privacy; model efficiency.<br>Cons: requires local training. |
| Category | Artificial Intelligence |

| Section | Description |
|---|---|
| Id | REQ-RS-04 |
| Short Name | Cold start problem |
| Type | Functional |
| Priority | Medium |
| Requirement Level | Recommended |
| Description | Recommendation platform should be capable of addressing the cold-start problem for users. |
| Additional Information | N/A |
| Purpose/Goal | New users in the system have no prior history or interaction log. The SEDIMARK recommender system should be able to cater for these users too and be able to provide them with recommendations of sufficient quality until it captures enough user information to be able to properly compute their preferences. |

| Section | Description |
|---|---|
| Validation Criteria | Model accuracy |
| Constraints: | Requires marketplace statistics and users properly using filtering criteria. |
| Dependencies: | Req-UI-05; Req-UI-08 |
| Conflicts: | N/A |
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: increased user satisfaction from the first interaction with the system<br>Cons: increases system complexity for tracking statistics |
| Category | Artificial Intelligence |

## 6.5 Energy efficiency

The SEDIMARK platform put a strong emphasis on energy efficiency. Modules will be designed to be as lightweight as possible while still aiming for a top performance. This will be applied to data processing modules and machine learning models, both for the training and inference parts

| Section | Description |
|---|---|
| Id | REQ-EE-01 |
| Short Name | Lightweight and energy efficient DP modules |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Data processing modules must be lightweight and energy efficient |
| Additional Information | N/A |
| Purpose/Goal | To reduce energy consumption while maintaining the same level of performance of the data processing modules |
| Validation Criteria | Data processing modules energy efficiency indicator |
| Constraints: | Requires more effort on development of energy efficient technique |
| Dependencies: | Req-DP-05; |
| Conflicts: | N/A |
| Relevant Use Cases | All use cases |

| Section | Description |
|---|---|
| Pros/Cons | Pros: less energy consumption, in line with economic and environmental concerns<br>Cons: may limit possibility and power of the modules |
| Category | Energy Efficiency |

| Section | Description |
|---|---|
| Id | REQ-EE-02 |
| Short Name | Lightweight and energy efficient AI/ML models |
| Type | Functional |
| Priority | Medium |
| Requirement Level | Recommended |
| Description | AI models must be small and lightweight to consume less energy both during training and inference. Energy efficiency techniques on AI models must not impact accuracy/performance |
| Additional Information | Quantisation and low rank approximation of ML models should be applicable to reduce energy consumption of ML and allow fast inference |
| Purpose/Goal | To reduce energy consumption while maintaining the same level of accuracy on the AI/ML models |
| Validation Criteria | AI/ML model energy efficiency indicator (on training and inference) |
| Constraints: | Requires more effort on development of energy efficient technique |
| Dependencies: | AI/ML models |
| Conflicts: | N/A |
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: less energy consumption, in line with economic and environmental concerns<br>Cons: may increase training time and complexity |
| Category | Energy Efficiency |

| Section | Description |
|---|---|
| Id | REQ-EE-03 |
| Short Name | Energy efficient decentralized training of ML model |
| Type | Functional |

| Section | Description |
|---|---|
| Priority | High |
| Requirement Level | Required |
| Description | Decentralised training of ML models must be communication and energy efficient |
| Additional Information | N/A |
| Purpose/Goal | To reduce energy consumption while maintaining the same level of accuracy on the AI/ML models |
| Validation Criteria | Decentralized training of ML model energy efficiency indicator |
| Constraints: | Requires more effort on development of energy efficient technique |
| Dependencies: | Decentralized AI/ML models training |
| Conflicts: | N/A |
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: less energy consumption, in line with economic and environmental concerns<br>Cons: may increase training time and complexity |
| Category | Energy Efficiency |

| Section | Description |
|---|---|
| Id | REQ-EE-04 |
| Short Name | Usage of compiler optimizations for ML model |
| Type | Functional |
| Priority | Low |
| Requirement Level | Optional |
| Description | ML models should make use of compiler optimizations (e.g., Torch.compile() vs vanilla pytorch) |
| Additional Information | N/A |
| Purpose/Goal | To reduce energy consumption while maintaining the same level of accuracy on the AI/ML models |
| Validation Criteria | Decentralized training of ML model energy efficiency indicator |

| Section | Description |
|---|---|
| Constraints: | Requires more effort on development of energy efficient technique |
| Dependencies: | AI/ML training |
| Conflicts: | N/A |
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: less energy consumption, in line with economic and environmental concerns<br>Cons: N/A |
| Category | Energy Efficiency |

## 6.6 Interoperability

SEDIMARK aims at integrating data and related actors and assets of very heterogeneous origins. Hence, ensuring interoperability among such heterogeneous entities is of vital importance. An essential element towards this is introducing a common information model that can represent all SEDIMARK entities (participants, data, services, AI models, etc.) in a unified way, thus allowing their identification, processing and interaction inside the SEDIMARK platform. The requirements detailed in the following address first the wide applicability of the information model (REQ-INT-01) and then focus on the modelling of data and their related metadata (REQ-INT-02, REQ-INT-03, REQ-INT-04, REQ-INT-05). Lastly, an examination is conducted to assess how the information model can ensure interoperability of AI (ML) models, specifically addressing the requirement REQ-INT-06

| Section | Description |
|---|---|
| Id | REQ-INT-01 |
| Short Name | Information model for interoperability |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | The information model should allow a wide application scope and enable interoperability |

| Section | Description |
|---|---|
| Additional Information | This requirement has the following sub-requirements:<br>• The information model should prescribe a unified representation of SEDIMARK entities (participants, data, services, AI models, etc.).<br>• The information model should use standard ontology languages.<br>• The information model should allow incorporation of domain models (ontologies/taxonomies) to support use case data loading, processing and enrichment.<br>• Comprehensive documentation/names should be provided for the entire information model and its extensions with domain models. |
| Purpose/Goal | SEDIMARK aims at integrating data and related actors and assets of various origins. An information model that can embrace this heterogeneity and enable interoperability is of highest importance. Coverage of all the various SEDIMARK entities, reliance on standards, clarity and extensibility are key elements towards this goal. |
| Validation Criteria | Effective entity representation and interoperability |
| Constraints: | Alignment with ongoing standardization / wide collaboration initiatives |
| Dependencies: | Information model has an impact on almost all the other requirement categories |
| Conflicts: | N/A |
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: Considerably increases SEDIMARK's scope<br>Cons: Handling of heterogeneity increases complexity |
| Category | Interoperability (Information model) |

| Section | Description |
|---|---|
| Id | REQ-INT-02 |
| Short Name | Information model for data and their metadata |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | The information model should prescribe aspects of the data and the metadata |

| Section | Description |
|---|---|
| Additional Information | This requirement has the following sub-requirements:<br>• The information model should prescribe:<br>  ○ a syntactic format and a (high-level) data schema for the data.<br>  ○ a syntactic format and a data schema for the metadata<br>• Metadata may be applied at different data resolutions: single data point, data batch, whole dataset/data stream.<br>• Data and metadata may be evolvable over time |
| Purpose/Goal | To enable interoperability of heterogeneous data, the information model should prescribe a minimum set of syntactic and semantic aspects of these data and their metadata. |
| Validation Criteria | Effective data interoperability |
| Constraints: | Alignment with ongoing standardization / wide collaboration initiatives |
| Dependencies: | Information model has an impact on almost all the other requirement categories |
| Conflicts: | N/A |
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: Data interoperability |
| Category | Interoperability (Information model) |

| Section | Description |
|---|---|
| Id | REQ-INT-03 |
| Short Name | Metadata fields |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | The information model should prescribe specific metadata fields for the data |

| Section | Description |
|---|---|
| Additional Information | This requirement has the following sub-requirements:<br>• Metadata should include identification of the data: e.g., what it is about, dataset/stream info (e.g., sampling strategy), provenance, provider, owner.<br>• Metadata should include metadata for data publication and discovery.<br>• Metadata should include semantic annotations of the data and the metadata referring to external well-known ontologies.<br>• Metadata should include provider's conditions of data use.<br>• Metadata should include payment-for-data-use policies.<br>• Metadata should include data quality indicators or metrics.<br>• Metadata should include data quality indicators or metrics qualifying data suitability for ML.<br>• Metadata should include metadata for enabling ML on the data (e.g., labels, semantic annotations)<br>• Metadata should include policies for data privacy, authorization and access control for data use.<br>• Metadata should include trust anchors for the data |
| Purpose/Goal | Metadata should provide sufficient information so that data can be handled inside SEDIMARK |
| Validation Criteria | Effective handling of data inside SEDIMARK |
| Constraints: | Alignment with ongoing standardization / wide collaboration initiatives |
| Dependencies: | Information model has an impact on almost all the other requirement categories |
| Conflicts: | N/A |
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: Rich metadata allows effective handling of data inside SEDIMARK<br>Cons: Rich metadata goes with increased processing complexity |
| Category | Interoperability (Information model) |

| Section | Description |
|---|---|
| Id | REQ-INT-04 |
| Short Name | Data compliance with the information model |
| Type | Functional |
| Priority | High |

| Section | Description |
|---|---|
| Requirement Level | Required |
| Description | Compliance with the information model should be a (partial) prerequisite for SEDIMARK data |
| Additional Information | This requirement has the following sub-requirements:<br>• Data should be annotated with metadata that comply with the information model in order to join SEDIMARK.<br>• Data should be annotated with metadata, and both data and metadata should comply with the information model, in order to be processed with SEDIMARK tools. |
| Purpose/Goal | Two cases are identified:<br>• Providers may maintain their original data. This data should at least be annotated with metadata as prescribed by the information model in order to be publishable inside SEDIMARK.<br>• Providers may both convert their data and annotate them with metadata as prescribed by the information model in order to be able to use the data processing tools provided by SEDIMARK. |
| Validation Criteria | Effective support of the two cases of incorporating data into SEDIMARK |
| Constraints: | N/A |
| Dependencies: | Information model has an impact on almost all the other requirement categories; REQ-INT-05 |
| Conflicts: | N/A |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: Offers freedom of choice to providers<br>Cons: Allowing both cases increase handling complexity |
| Category | Interoperability (Information model) |

| Section | Description |
|---|---|
| Id | REQ-INT-05 |
| Short Name | Tools for enforcing data compliance with the information model |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |

| Section | Description |
|---|---|
| Description | Tools should be developed for enforcing compliance of data with the information model. |
| Additional Information | This requirement has the following sub-requirements:<br>• Data converters should be developed for converting providers' original data so that they comply with the information model.<br>A data annotator should be developed for annotating data with metadata that comply with the information model.<br>• Data validator should be developed to assess compliance with the SEDIMARK information model. |
| Purpose/Goal | The importance of data compliance with the SEDIMARK information model and the two possible cases for providers are discussed in REQ-INT-04. Ready-to-use or customizable data tools should be offered to providers so that they ensure such compliance for their data. |
| Validation Criteria | Development of generic and/or ready-to-use tools that will at least cover the SEDIMARK use cases |
| Constraints: | Potentially great number of different data syntactic formats. |
| Dependencies: | Information model has an impact on almost all the other requirement categories. |
| Conflicts: | N/A |
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: Ensures compliance<br>Cons: Effort for supporting a potentially great number of different data syntactic formats |
| Category | Interoperability (Information model) |

| Section | Description |
|---|---|
| Id | REQ-INT-06 |
| Short Name | Information model for AI models |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | The information model should prescribe a common format for AI (ML) models that are processed with SEDIMARK tools |
| Additional Information | This requirement has the following sub-requirement:<br>Tools are required for converting AI models to the common format. |

| Section | Description |
|---|---|
| Purpose/Goal | AI models could be different, given that not all providers / consumers will use the same software to train/run their AI models. Setting a common format will allow AI models to work in a variety of frameworks and compilers. |
| Validation Criteria | Effective interoperability of AI models |
| Constraints: | Potentially great number of different AI model formats and platforms |
| Dependencies: | AI requirements |
| Conflicts: | N/A |
| Relevant Use Cases | All use cases |
| Pros/Cons | Pros: AI model interoperability |
| Category | Interoperability (Information model) |

## 6.7 Data storage

As a core functional requirement for most data management lifecycles, SEDIMARK's decentralised framework will depend on distributed storage of metadata for data from providers to support the provision of the catalogue for search and discovery of datasets and the storage of intermediate artefacts during a data processing pipeline.

| Section | Description |
|---|---|
| Id | Req-STR-01 |
| Short Name | Default data hosting location |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Datasets must be stored within the provider's domain by default. |
| Additional Information | By default, any datasets that will be available through the marketplace, must be stored within the provider's domain. |
| Purpose/Goal | Information governance for datasets must be handled by the provider to minimize external liability for handling possible sensitive data. |
| Validation Criteria | Persistence hosted within provider's domain. |
| Constraints: | Not all providers are expected to have storage capability |
| Dependencies: | Data Storage; |
| Conflicts: | No |

| Section | Description |
|---|---|
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: Ensures information governance is handled by the provider.<br>Cons: Limit types of providers who can register with the marketplace. |
| Category | Storage; |

| Section | Description |
|---|---|
| Id | Req-STR-02 |
| Short Name | Distributed metadata storage |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Offerings' descriptions (metadata) must be stored in the distributed catalogue provided by SEDIMARK. |
| Additional Information | The store will need to also cater to changes in Offerings. |
| Purpose/Goal | To support consumers with search and discovery. |
| Validation Criteria | Consumers able to retrieve offerings' descriptions. |
| Constraints: | Offering descriptions need to be retrievable from any consumer node.<br>Offering descriptions should remain in respective provider's domain. |
| Dependencies: | REQ-P&D-01 ; REQ-INT-02 |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: A decentralised approach to metadata storage can allow search queries to be more targeted.<br>Cons: requires multiple queries as opposed to centralised approach. |
| Category | Storage; Publication and Discovery |

| Section | Description |
|---|---|
| Id | Req-STR-03 |
| Short Name | Pipeline temporary storage |

| Section | Description |
|---|---|
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Intermediate data through data processing pipeline needs to be temporarily stored on disk or in-memory. |
| Additional Information | Depending on the data processing pipeline that has been orchestrated, storage for different artefacts will be needed temporarily throughout the process. |
| Purpose/Goal | Support data processing checkpoints. |
| Validation Criteria | Data processing pipeline is able to store and retrieve intermediate processing artefacts. Any intermediate artefacts remaining should be deleted if not required anymore. |
| Constraints: | No |
| Dependencies: | REQ-DP-01; |
| Conflicts: | No. |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: Intermediate artefacts remain with provider's domain.<br>Cons: Provider's would be expected to provide compute resources for storage. |
| Category | Storage; |

| Section | Description |
|---|---|
| Id | Req-STR-04 |
| Short Name | Storage for pipeline artefact outputs |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Data from final output of data processing pipeline needs to be stored in consumable manner. |
| Additional Information | The final product from a data processing pipeline could differ in terms of format and schema, and therefore would need to be packaged and stored appropriately. |

| Section | Description |
|---|---|
| Purpose/Goal | Provide alternative data to original dataset, in terms of quality, aggregation, enrichment. |
| Validation Criteria | Data processing pipeline outputs are stored. |
| Constraints: | Provider would need to be able to accommodate for output which could be different from conventional format. |
| Dependencies: | Req-STR-02, Req-STR-03 |
| Conflicts: | None. |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: facilitates the consumption of processed data<br>Cons: requires considering many different output formats |
| Category | Storage; |

| Section | Description |
|---|---|
| Id | Req-STR-05 |
| Short Name | Scalable provision for large data generated. |
| Type | Functional |
| Priority | Low |
| Requirement Level | Required |
| Description | Providers collecting huge amounts of data require distributed data management for scalable provision. |
| Additional Information | Accumulation of large datasets can gradually cause scalability issues which can affect the quality of service in provision. |
| Purpose/Goal | Better provision of datasets upon exchange. |
| Validation Criteria | QoS |
| Constraints: | Provider resource capabilities. |
| Dependencies: | Data Storage |
| Conflicts: | No. |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: improves providers quality of service<br>Cons: adds complexity to data management |

| Section | Description |
|---|---|
| Category | Storage; |

| Section | Description |
|---|---|
| Id | Req-STR-06 |
| Short Name | Outsourced Data Storage |
| Type | Functional |
| Priority | Medium |
| Requirement Level | Optional |
| Description | Data from providers who cannot host data onsite needs to be stored on other edge domains through an external service. |
| Additional Information | The service will allow resource-constrained providers to store their data with providers at the edge with extra storage capacity. |
| Purpose/Goal | Support resource-constrained data providers. |
| Validation Criteria | Data from one provider is transferred and securely stored in the other provider's domain. |
| Constraints: | Data stored would normally need to be encrypted and only accessible to consumers. |
| Dependencies: | Data Storage |
| Conflicts: | Req-STR-01 |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: Allow more diverse data providers. |
| Category | Cons: Data protection measures could be compromised. |

| Section | Description |
|---|---|
| Id | Req-STR-07 |
| Short Name | Knowledge graph persistence |
| Type | Functional |
| Priority | Medium |
| Requirement Level | Recommended |
| Description | Services creating knowledge graphs need to be able to store in appropriate persistence. |

| Section | Description |
|---|---|
| Additional Information | Datasets can be supplemented by knowledge graphs for building domain knowledge relating to use cases, which can link to different datasets. |
| Purpose/Goal | Provision of knowledgebases. |
| Validation Criteria | Semantic storage is available. |
| Constraints: | Semantic enrichment of knowledge graphs would need persistence that supports graph-based annotation and querying. |
| Dependencies: | Data Storage |
| Conflicts: | No. |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: Build domain knowledge for various specific scenarios.<br>Cons: can significantly increase size of data if annotation is not handled carefully. |
| Category | Storage; |

| Section | Description |
|---|---|
| Id | Req-STR-08 |
| Short Name | Appropriate storage for offering type. |
| Type | Functional |
| Priority | Low |
| Requirement Level | Recommended |
| Description | Storing options for each type of offering (data, features, model, configuration etc) or configuration need to be consistent. |
| Additional Information | To support utilisation from services and consumers |
| Purpose/Goal | Support distributed data analysis operations. |
| Validation Criteria | Provision of multi-type storage |
| Constraints: | Provider resource capabilities |
| Dependencies: | REQ-INT-06 |
| Conflicts: | No. |
| Relevant Use Cases | All use cases. |

| Section | Description |
|---|---|
| Pros/Cons | Pros: enables easier access to artifact.<br>Cons: adds complexity to storage mechanism. |
| Category | Storage |

## 6.8 Publication and discovery

In order to enable a proper exchange of information among participants in the SEDIMARK marketplace, several requirements for publication and discovery of the different assets provided in the marketplace need to be fulfilled.

| Section | Description |
|---|---|
| Id | REQ-P&D-01 |
| Short Name | Assets described as part of offerings |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Assets (e.g., datasets, data streams, services, …) have to be uniformly described as offerings to support its homogeneous publication and easy discovery |
| Additional Information | This requirement includes the following sub-requirements:<br>• Establish a standardized metadata schema for describing assets. This will help users/services understand the content, format, and context of the data, and facilitate its discovery and integration.<br>• Tools should be provided to simplify the offering generation stage.<br>• Data processing pipeline should provide additional metadata related to the data quality domain which should be included in the offering as part of the asset description. |
| Purpose/Goal | SEDIMARK aims to provide a marketplace in which participants can exchange heterogeneous assets. Therefore, it is key to generate uniform descriptions able to support a diverse set of assets, as this sets the foundations for their publication and subsequent discovery.<br>The goal is to offer providers a flexible model to describe what they want to sell in the marketplace, allowing then to model their offerings imposing the minimum set of restrictions on how to package and sell their available assets. |
| Validation Criteria | Generated offering description are validated against the defined information model. |
| Constraints: | Too many possible scenarios to support, as the set of all existing assets is unbounded by nature. |

| Section | Description |
|---|---|
| Dependencies: | REQ-INT-01, REQ-INT-02, REQ-INT-03 |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: information model handled in the SEDIMARK marketplace is known in advance.<br>Cons: possible information duplicities. |
| Category | Publication and discovery |

| Section | Description |
|---|---|
| Id | REQ-P&D-02 |
| Short Name | Offerings' registry |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Offerings have to be published on a common distributed registry from which they can be listed |
| Additional Information | N/A |
| Purpose/Goal | The marketplace needs to have a distributed registry able to provide trustworthiness to the set of existing heterogeneous offerings. This registry will serve as reference point for any other module/component willing to provide additional functionalities on top of offerings. |
| Validation Criteria | Registry is available |
| Constraints: | Amount of information to be stored on a distributed ledger versus performance. |
| Dependencies: | REQ-SEC-04 |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: common reference point for offering information<br>Cons: information duplicity (i.e., registry versus catalogue(s)) |
| Category | Decentralization, Publication and discovery |

| Section | Description |
|---|---|
| Id | REQ-P&D-03 |
| Short Name | Generic offering metadata |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Metadata included in offerings must be generic enough to avoid imposing restrictions on the assets data format nor the mechanisms used for its provision |
| Additional Information | N/A |
| Purpose/Goal | In essence, this requirement pretends to detach the assets that any provider wants to sell in the marketplace from their description within offerings. While this requirement deals with the information model definition and could also be considered part of the interoperability requirements, the goal is to have a generic enough offering information model to allow providers to describe any kind of asset, without being only restricted to the specific ones included on the SEDIMARK use cases. |
| Validation Criteria | Offering information model |
| Constraints: | Too many possible scenarios to support, as the set of all existing assets is unbounded by nature. |
| Dependencies: | REQ-INT-01, REQ-INT-02, REQ-INT-03 |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: flexibility on the offering information model<br>Cons: some metadata included in the offering might be too generic |
| Category | Publication and discovery, Openness |

| Section | Description |
|---|---|
| Id | REQ-P&D-04 |
| Short Name | Open Data portal discovery |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |

| Section | Description |
|---|---|
| Description | Some Open Data should be discoverable as part of the offering catalogue |
| Additional Information | The SEDIMARK catalogue should be populated with some offerings from Open Data portals, to provide users with access to open datasets from CKAN (https://ckan.org/showcase) or Kaggle (https://www.kaggle.com/) for instance. |
| Purpose/Goal | The goal of this requirements is to populate the SEDIMARK catalogue with some offerings from Open Data portals, accessible to any user. |
| Validation Criteria | Some selected offerings (datasets) from Open Data portals are discoverable in the catalogue.<br>These offerings are accessible to any user of SEDIMARK.SEDIMARK |
| Constraints: | Offerings from Open Data portals should not appear different to SEDIMARK users. |
| Dependencies: | REQ-P&D-01<br>REQ-P&D-02 |
| Conflicts: | REQ-INT-04 |
| Relevant Use Cases | None |
| Pros/Cons | Pros: ensures a minimum set of offerings accessible directly from SEDIMARK inception.<br>Cons: from the users' perspective, there should not be any difference between accessing an open dataset and a paid one from another SEDIMARK participants. This could be challenging to address. |
| Category | Publication and Discovery, Openness |

| Section | Description |
|---|---|
| Id | REQ-P&D-05 |
| Short Name | Offerings' catalogue for queries |
| Type | Functional |
| Priority | High |
| Requirement Level | Recommended |
| Description | Catalogue(s) should be constructed by fetching offering metadata and employ different indexing strategies based on provider/consumer requirements |
| Additional Information | N/A |

| Section | Description |
|---|---|
| Purpose/Goal | Different searchable catalogues might exist in the marketplace to provide different optimizations on the discovery stage. In order to do so, catalogues need to rely on different technologies and/or indexing techniques on top of the offering related information available on the distributed registry. |
| Validation Criteria | Catalogue is available |
| Constraints: | Query performance and data storage limitations. Distributed queries. |
| Dependencies: | REQ-STR-02 |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: support different optimizations at search engine layer, possibility to explore distributed queries<br>Cons: information duplicity (i.e., multiple catalogues on the marketplace) |
| Category | Decentralization, discovery |

## 6.9 User requirements (Marketplace User Interface)

| Section | Description |
|---|---|
| Id | REQ-UI-01 |
| Short Name | Logging in UI |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Users should be able to log in in the marketplace UI. |
| Additional Information | N/A |
| Purpose/Goal | The goal of this requirement is to add a log in portal to the marketplace UI |
| Validation Criteria | Log in portal should be in place in the marketplace.<br>Protected pages should not be accessible to non-authenticated users. |
| Constraints: | N/A |
| Dependencies: | REQ-SEC-01 |
| Conflicts: | No |

| Section | Description |
|---|---|
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: offers security<br>Cons: extra step required, implies development efforts |
| Category | UI, Security |

| Section | Description |
|---|---|
| Id | REQ-UI-02 |
| Short Name | Offerings discoverability |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Users must be able to discover offerings in the marketplace based on identity-based policies defined by the asset provider. |
| Additional Information | The user can perform searches and filter the list of offerings she/he can access via the marketplace UI. |
| Purpose/Goal | The goal of this requirement is to ensure that all offerings posted by providers can be discovered by other participants in the SEDIMARK ecosystem, depending on the policies set by the providers. |
| Validation Criteria | Users can discover offerings corresponding to their role/identity in the marketplace UI. |
| Constraints: | - |
| Dependencies: | REQ-P&D-02 (more generally, depends on Publication & Discovery requirements)<br>REQ-SEC-02 |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: Visibility and searchability of offerings in a UI.<br>Cons: Access policies must be well defined in advance. |
| Category | UI, Publication and Discovery |

| Section | Description |
|---|---|
| Id | REQ-UI-03 |

| Section | Description |
|---|---|
| Short Name | Users Identity |
| Type | Non-Functional/Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Users have to be able to create, own and manage their own SSI identity |
| Additional Information | This requirement allows the user to own a digital identity that can be employed to access the services within SEDIMARK. Users must be able to create, modify own and manage their own SSI-based identity. |
| Purpose/Goal | The target of this requirement is to provide every user of SEDIMARK with its own identity. Employing SSI framework let the user maintains control of its own identity. |
| Validation Criteria | - |
| Constraints: | No |
| Dependencies: | Req-SEC-01 |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: Each user has full control on its own identity<br>Cons: Need to define the roles of Issuer, Holder, Verifier among the actors. |
| Category | UI, Security |
| Id | REQ-UI-04 |
| Short Name | Offerings management |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Providers should be able to manage their offerings (adding / removing them) |
| Additional Information | When adding new offerings, providers can configure their privacy levels, authorization policies and potentially set a licence.<br>Providers can also remove their offerings, making them not discoverable anymore in the catalogue UI. |

| Section | Description |
|---|---|
| Purpose/Goal | The goal of this requirement is to grant providers full control on the management of their offerings via the marketplace UI, so they can add new offerings or remove existing ones. |
| Validation Criteria | The marketplace should provide an interface to add new offerings and configure them.<br>The marketplace should provide an interface to providers to see their offerings, edit them and remove them. |
| Constraints: | Editing an offering, if possible, should be clarified, since information about offerings are stored in an immutable way. |
| Dependencies: | REQ-P&D-01<br>REQ-P&D-02<br>REQ-SEC-03 |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: User friendly graphical interface for offering management<br>Cons: requires user intervention for managing their offerings |
| Category | UI |

| Section | Description |
|---|---|
| Id | REQ-UI-05 |
| Short Name | Offering description page |
| Type | Functional |
| Priority | High |
| Requirement Level | Required |
| Description | Users should be able to view detailed information about offerings |
| Additional Information | Upon browsing the available offerings, a user should be able to select one and access its more detailed description (dataset size, service description, authorization policies, price, contract duration, delivery method, …) |
| Purpose/Goal | The goal of this requirement is to provide a user interface gathering all information about an offering and expose them in a clear and intuitive way to future consumers. |
| Validation Criteria | The marketplace should provide an interface describing exhaustively a selected offering in the catalogue. |

| Section | Description |
|---|---|
| Constraints: | Editing an offering, if possible, should be clarified, since information about offerings are stored in an immutable way. |
| Dependencies: | REQ-P&D-01<br>REQ-P&D-03 |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: User friendly graphical interface to visualize an offering description<br>Cons: Supporting other information than offering metadata can be challenging if not standardized |
| Category | UI |

| Section | Description |
|---|---|
| Id | REQ-UI-06 |
| Short Name | SEDIMARK toolbox access in UI |
| Type | Functional |
| Priority | Medium |
| Requirement Level | Recommended |
| Description | Users should be able to interact with their offerings in the marketplace UI (visualize them, run them through the SEDIMARK toolbox, …) |
| Additional Information | SEDIMARK's added value relies on providing advanced tools for data processing, directly embedded in its ecosystem. This requirement aims at making this toolbox accessible from a graphical interface, seamlessly from the marketplace UI.<br>For example, the marketplace UI could provide access to a Jupiter Hub where users could import their data and SEDIMARK toolbox. |
| Purpose/Goal | The goal of this requirement is to provide users with an access to SEDIMARK's data processing tools directly from the marketplace UI. |
| Validation Criteria | The marketplace should provide a UI, or a hook, to access SEDIMARK's toolbox. |
| Constraints: | Some offerings may not be compatible with the SEDIMARK's data processing tools. |
| Dependencies: | REQ-DP-01, and in general data processing requirements |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |

| Section | Description |
|---|---|
| Pros/Cons | Pros: Values SEDIMARK data processing toolbox and make it easier for users to access it.<br>Cons: Adapting to the diversity of offerings can be challenging |
| Category | UI |

| Section | Description |
|---|---|
| Id | REQ-UI-07 |
| Short Name | Rating offerings in UI |
| Type | Functional |
| Priority | Low |
| Requirement Level | Required |
| Description | Users should be able to rate and review offerings in the marketplace UI. |
| Additional Information | Star rating/written feedback |
| Purpose/Goal | The goal of this requirement is to enable consumers to review and rate offerings they used |
| Validation Criteria | Ratings should be visible with offerings in the catalogue UI.<br>Ratings and reviews can be created by consumers in the marketplace.<br>Detailed reviews can be seen from offering description pages |
| Constraints: | Requires a mechanism for providing ratings in a trusted manner. |
| Dependencies: | REQ-UI-05 |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: feedback can provide insights about offers<br>Cons: need of a moderator, which implies human assistance |
| Category | UI |

| Section | Description |
|---|---|
| Id | REQ-UI-08 |
| Short Name | Offering statistics in UI |
| Type | Functional |
| Priority | Low |

| Section | Description |
|---|---|
| Requirement Level | Optional |
| Description | Providers should be able to see statistics related to the access of their offerings |
| Additional Information | Statistics like histograms, charts can be provided |
| Purpose/Goal | The goal of this requirement is to enable providers to visualize, in the marketplace UI, some statistics about their offerings |
| Validation Criteria | Statistics must be relevant for the users |
| Constraints: | Requires a mechanism for gathering statistics regarding offering usage |
| Dependencies: | REQ-DP-01, REQ-DP-02, REQ-DP-03, REQ-DP-04 |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: statistics can offer insights<br>Cons: too many statistics can cause overload |
| Category | UI |

## 6.10 Smart Contract and Tokenization

Trading of assets in the SEDIMARK Marketplace is regulated with Smart Contracts, enabled by an additional level on top of the Distributed Ledger. The following requirements summarize the necessary features that must be exposed by the smart contracts that will be deployed to construct and manage the SEDIMARK Marketplace.

| Section | Description |
|---|---|
| Id | REQ-SCT-01 |
| Short Name | Smart Contracts support |
| Type | Non-Functional/Functional |
| Priority | Medium |
| Requirement Level | Required |
| Description | DLT shall support Smart Contracts |
| Additional Information | IOTA Tangle enables the usage of Smart Contracts by leveraging an additional framework on top of the Tangle itself. The framework to be employed for the smart contracts is denoted as “IOTA Smart Contract (ISC)”, usually referenced as “Level 2”. |

| Section | Description |
|---|---|
| Purpose/Goal | Smart Contracts are a mandatory feature that enables a form of secure trading of digital assets exchanged in the SEDIMARK Marketplace. |
| Validation Criteria | The group of technologies employed, i.e., IOTA Tangle, ISC framework and WASP validator nodes. |
| Constraints: | No |
| Dependencies: | No |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: allows to regulate buy and sell assets through Smart Contracts<br>Cons: limits the choice of DLTs to those that support Smart Contracts |
| Category | Smart Contract and Tokenization |

| Section | Description |
|---|---|
| Id | REQ-SCT-02 |
| Short Name | Tokenization of Assets |
| Type | Non-Functional/Functional |
| Priority | Medium |
| Requirement Level | Required |
| Description | Assets shall be tokenized |
| Additional Information | A representation of the asset in the form of an NFT: the asset together with additional metadata are uniquely created and identified by its corresponding smart contract. Being the NFT a non-fungible token, it must be associated with a fungible token to regulate the trading of the respective asset. Moreover, based on the consumer credential (containing the authorization policies), the Marketplace automatically filters the offerings that the consumer is authorized to buy. |
| Purpose/Goal | The purpose is to enable the trusted exchange of assets by leveraging the combination of the asset related NFTs together with the Fungible Tokens. The ownership of the correct number of fungible tokens (related to the respective NFT) proves that a consumer has the rights to access the related asset. |
| Validation Criteria | - |
| Constraints: | The dataset representation (NFT) must be related to a unique fungible token to enable the exchange of such asset. |

| Section | Description |
|---|---|
| Dependencies: | No |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: allows to exchange of assets through SC framework<br>Cons: overhead due to libraries and operations for Tokenization and SC |
| Category | Smart Contract and Tokenization |

| Section | Description |
|---|---|
| Id | REQ-SCT-03 |
| Short Name | User Digital Wallet |
| Type | Non-Functional/Functional |
| Priority | Medium |
| Requirement Level | Required |
| Description | Users must be able to own and manage a Digital Wallet |
| Additional Information | Every actor must own an appropriate Digital Wallet that allows them to interact with the SEDIMARK Marketplace also in terms of tokens. |
| Purpose/Goal | The Digital Wallet allows to hold, send and receive tokens resulting from the trading of assets. |
| Validation Criteria | - |
| Constraints: | The Digital Wallet must be able to interact within the SEDIMARK domain. Every user needs to maintain the access to its own wallet (including wallet keys). |
| Dependencies: | Req-UI-03 |
| Conflicts: | No |
| Relevant Use Cases | All use cases. |
| Pros/Cons | Pros: Users are able to receive and send funds<br>Cons: overhead due to libraries and framework for the digital wallet |
| Category | Smart Contract and Tokenization |

# 7 Conclusions

This deliverable provides the analysis of the 4 SEDIMARK use cases and more than 70 requirements associated with these use cases.

The SEDIMARK project comprises four use cases, namely Mobility Digital Twin, Urban Bike Mobility Planning, Valorisation of Energy Consumption and Customer Reactions/Complaints, and Valuation and Commercialisation of Water Data. Each use case is elaborated upon with a concise description, identification of involved actors and stakeholders, current status, preconditions, postconditions, exceptions, services to be offered, data generation and sharing, associated data models, key performance indicators (KPIs), and expected flows. The objective is to gain a comprehensive understanding of the offerings and relevance of each use case within the SEDIMARK project. Additionally, this analysis aids in identifying the required modules for each use case and establishing measurable success criteria.

From the analysis of these use cases, there is a list of requirements that is derived. These requirements are classified in 10 categories, which correspond to different research aspects and implementation activities of the project, namely: Non-functional - system architecture requirements; Security, privacy and trust; Data quality management and data processing; Artificial Intelligence; Energy efficiency; Interoperability; Data storage; Publication and discovery; User interface; Smart Contract and Tokenization. The requirements are described and analysed in terms of an identifier for traceability, a short name, their distinction on functional or non-functional, their priority and their requirement level, their description, the purpose goal, the validation criteria, any constraints/dependencies/conflicts, the relevant use cases, the advantages and disadvantages on having this requirement in the project. The purpose is to guide all design and development activities, in order to make the SEDIMARK marketplace decentralised, trustworthy, interoperable and open to new data (open data), with intelligent AI-based and energy efficient data management tools capable of providing high quality data and services to consumers.

The outcomes of this deliverable encompass an exploration of the business and technology landscape surrounding SEDIMARK, along with an initial analysis of trust and security. This deliverable holds significant importance as it establishes the context for all subsequent activities within the project. The insights gathered from this document will shape the project's architecture, as outlined in D2.2 (SEDIMARK architecture and interfaces- First version). The architecture, developed through a requirement elicitation process based on this deliverable, will serve as the foundation for the design and implementation of activities in WP3 (Distributed data quality management and interoperability) and WP4 (Secure data sharing in a decentralized Marketplace). Furthermore, as the requirements are further analyzed and aligned with architectural patterns, the instantiation of the platform will be created to guide the development process in WP5 (Integration, testing and evaluation). The defined use cases will play a crucial role in WP5, where pilot demonstrators will be developed, and the solution will be evaluated and monitored using performance assessment metrics relevant to the defined activities.

# 8 References


[1] ISO/IEC/IEEE 29148:2018, "Systems and software engineering -- Life cycle processes -- Requirements engineering," https://www.iso.org/standard/72089.html.

[2] Hausmann, Jan Hendrik, Reiko Heckel, and Gabi Taentzer., " "Detection of conflicting functional requirements in a use case-driven approach: a static analysis technique based on graph transformation."," *Proceedings of the 24th international conference on software engineering,* 2002.

[3] Nuseibeh, Bashar, and Steve Easterbrook, " "Requirements engineering: a roadmap."," *Proceedings of the Conference on the Future of Software Engineering,* 2000.

[4] Zave, Pamela. , ""Classification of research efforts in requirements engineering."," *ACM Computing Surveys (CSUR,* pp. 315-321, 1997.

[5] al., Mohammed Odeh et, "A use-case driven approach in requirements engineering: the MammoGrid project," *arXiv,* 2004.

[6] De Lucia, Andrea, and Abdallah Qusef, "Requirements engineering in agile software development.," *ournal of emerging technologies in web intelligence 2.3,* pp. 212-220, 2010.

[7] Scott, Ambler, The agile unified process (aup).", Ambysoft, 2005.

[8] Scott, Kendall, The unified process explained, Addison-Wesley Longman Publishing Co., 2002.

[9] ISO, *Systems and software engineering — Life cycle processes — Requirements engineering ISO/IEC/ STANDARD IEEE 29148,* ISO.

[10] Graves, Alex et al., "Hybrid computing using a neural network with dynamic external memory.," vol. 538, 2016.

[11] Xin J, Afrasiabi C, Lelong S, Adesara J, Tsueng G, Su AI, Wu C., "Cross-linking BioThings APIs through JSON-LD to facilitate knowledge exploration. BMC Bioinformatics.," Vols. doi: 10.1186/s12859-018-2041-5, 2018.

[12] W3C Working Group Note, ""CSV on the Web: A Primer" W3C Recommendation," 14 January 2016. [Online]. Available: https://www.w3.org/TR/tabular-data-primer/.

[13] H. Ed-douibi, J. L. C. Izquierdo and J. Cabot, "Model-driven development of OData services: An application to relational databases," *12th International Conference on Research Challenges in Information Science (RCIS),* 2018.

[14] https://www.gaia-x.eu/, "GAIA-X".

[15] https://internationaldataspaces.org/, "International Data spaces association".

[16] https://www.bdva.eu/, "Big Data Value Association".

[17] https://www.opendei.eu/, "OPEN DEI".

[18] https://www.fiware.org/, "FIWARE".

[19] Kaan Ozbay, Satish V. Ukkusuri, and Moises Sudit, "Intelligent Transportation Systems: New Principles and Architectures," *IEEE Transactions on Intelligent Transportation Systems,* 2011.

[20] Devansh Saxena, S. S. Chauhan, and L. M. Saini, "Data-driven Demand Response: Statistical Techniques for Decentralized Load Control," *2018 International Conference on Recent Advances in Electronics and Communication Technology (ICRAECT),* 2018.

[21] Vaishnavi Suresh, Sujata K. Patil, "Big data analytics in healthcare: promise and potential," *Health Information Science and Systems,* 2019.

[22] Yang Yu, Yada Zhu, and Zhiqiang Ma, "Deep Learning for Financial Text Mining: A Survey," *International Journal of Financial Studies,* 2021.

[23] Monika Ramesh, N. Ramesh Babu, and Venkatesh Babu R, "Industry 4.0: A Comprehensive Survey on Technologies, Use Cases, Challenges, and Opportunities," *Journal of Cleaner Production,* 2021.

[24] Li, H., "Docker. Encyclopedia. Available," 2023.

[25] Hardt, D., Ed., ""The OAuth 2.0 Authorization Framework",," October 2012.

[26] R. V. a. M. D. Wilde, Using OpenRefine (1st. ed.). Packt Publishing, 2013.

[27] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica., "Apache Spark: a unified engine for big data processing," vol. 59, 2016.

[28] Hammersley, Ben, "Developing Feeds with RSS and Atom," Sebastopol: O'Reilly, 2005.

[29] Fedoryszak, M., Tkaczyk, D., Bolikowski, Aalberg, T., Papatheodorou, C., Dobreva, M., Tsakonas, G., Farrugia, C.J. (eds), "Large Scale Citation Matching Using Apache Hadoop," Research and Advanced Technology for Digital Libraries, 2013.

[30] Helsinki City Government, "Helsinki City Government, "3D City Model,"," [Online]. Available: https://www.hel.fi/helsinki/en/administration/information/general/3d.

[31] Helsinki Region Infoshare, "Helsinki Region Infoshare," [Online]. Available: https://hri.fi/en_gb/.

[32] LAMP API, "LAM API Documentation," [Online]. Available: https://lamapi.azurewebsites.net/swagger/index.html?url=/swagger/v1/swagger.json#/Public.

[33] Helsinki Region Environmental Services HSY, "HSY," [Online]. Available: https://www.hsy.fi/en/air-quality-and-climate/geographic-information/open-geographic-information-interfaces/.

[34] https://smartdatamodels.org.

[35] https://github.com/smart-data-models/dataModel.Transportation/tree/db1e11c396a7b5f6e251bab27d56b38a26068a35 .

[36] https://github.com/smart-data-models/dataModel.Environment/tree/6a007de64d6b072c8a1e3bf0dfe59d301b55e8c3 .

[37] https://github.com/smart-data-models/dataModel.Weather.

[38] My protergia, "Myprotergia app," [Online]. Available: https://play.google.com/store/apps/details?id=com.intelen.dig.protergia&hl=en_US.

[39] Popov, Serguei, "The Tangle," *White Paper,* p. 30, 2018.

[40] Alex Preukschat, Drummond Reed, Decentralized digital identity and verifiable credentials, Manning Shelter Island, 2021.

[41] (W3C), The World Wide Web Consortium, "Decentralized Identifiers (DIDs) v1.0," [Online]. Available: https://www.w3.org/TR/did-core/.

[42] W3C, "W3C: Verifiable Credentials Data Model v1.1. W3C Recommendation," [Online]. Available: https://www.w3.org/TR/vc-data-model/. [Accessed 2022].

[43] Hausmann, Jan Hendrik, Reiko Heckel, and Gabi Taentzer., "Detection of conflicting functional requirements in a use case-driven approach: a static analysis technique based on graph transformation."," 2002.

# Annexes

The following table summarizes all the requirements as these were initially collected before their analysis in Chapter 6.

| Identifier | Short name | Title | Priority | Req. Level | Category |
|---|---|---|---|---|---|
| **Security and Authorization** | | | | | |
| Req-SEC-01 | Authentication of users | Users (Providers and Consumers) must be able to authenticate in the SEDIMARK Trust domain | H | REQ | *Security* |
| Req-SEC-02 | Authorization policies of assets | Each asset must be characterized by an authorization policy decided by the respective Provider | H | REQ | *Security* |
| Req-SEC-03 | Origin of assets | Assets origin and integrity must be maintained in the SEDIMARK Trust Domain | H | REQ | *Security* |
| Req-SEC-04 | Trust Metadata on Distributed Ledger | Trust Metadata must be written onto the Distributed Ledger | H | REQ | *Security, Decentralization* |
| Req-SEC-05 | Decentralized provisioning | Assets provisioning must employ a decentralised approach | H | REQ | *Decentralization* |
| Req-SEC-06 | Secure channel of the assets | Assets must be transferred over secure communication channels (e.g., through TLS) | H | REQ | *Security* |
| **Data Processing requirements** | | | | | |
| Req-DP-01 | Data cleaning tools | SEDIMARK must provide a complete toolset for data cleaning, curation and quality improvement | H | REQ | Data processing |

| Identifier | Short name | Title | Priority | Req. Level | Category |
|---|---|---|---|---|---|
| Req-DP-02 | Flexibility to handling both static and streaming data. | Data curation modules should be flexible to handle different types of data and both static and streaming data. | M | REC | Data processing |
| Req-DP-03 | Data processing pipeline configuration flexibility | The data curation pipeline should be configurable as per what modules will be run and how each module will run. | M | REC | Data processing |
| Req-DP-04 | Data quality indicators | Data should include quality indicators or metrics. | M | REC | Data processing |
| Req-DP-05 | Adaptability of data cleaning mechanisms | Data cleaning mechanisms should be adaptable to the technical capabilities of the device that runs them | M | REC | Data processing |
| Req-DP-06 | Ground truth for data quality metrics | Data quality metrics need to have ground truth for proper calculation based on domain knowledge and should be parameterizable by the users | L | OPT | Data processing |
| Req-DP-07 | Data cleaning modules extendable definitions | Data cleaning modules should be extendable to different definitions of duplicates and errors | M | REC | Data processing |

| Identifier | Short name | Title | Priority | Req. Level | Category |
|---|---|---|---|---|---|
| Req-DP-08 | Automate execution of tasks | Automate execution of various tasks where possible, to minimise human intervention, allowing also non-technical users to be able to execute tasks. | M | REC | Data processing |
| Req-DP-09 | Dataset augmentation | Datasets might need to be augmented with synthetic data before shared to remove biases and produce more balanced datasets. | L | OPT | Data processing |
| Req-DP-10 | Anonymization of private information | Private information in assets must be anonymized during pre-processing in the Data Owners Domain | H | REC | *Data Processing* |
| **AI-ML Requirements** | | | | | |
| Req-ML-01 | Model input data cleaning and formatting | Data need to be cleaned, in a standard common format and of a specific size for proper training of the models | H | REQ | AI/ML |
| Req-ML-02 | Decentralised ML | Decentralised training of ML models should be employed for increased privacy without exchanging raw data. | H | REQ | AI/ML |
| Req-ML-03 | Trusted participation in decentralised training | Decentralised training should be done only using trusted peers and with secure connections for improved performance not allowing tampering with data or man in the middle attacks. | H | REQ | AI/ML |
| Req-ML-04 | Models' agnostic to platforms | ML models should be able to be trained regardless of the underlying platform the peers use even on low-spec machines for maximum interoperability and efficiency. | M | REC | AI/ML |

| Identifier | Short name | Title | Priority | Req. Level | Category |
|---|---|---|---|---|---|
| Req-ML-05 | Models' persistence mechanisms | Model persistence: models should be persisted in a centralized/decentralized manner to be easily accessible and sharable | H | REC | AI/ML |
| Req-ML-06 | Event generation from pattern extraction | Pattern Extraction for event detection | H | REC | AI/ML |
| Req-ML-07 | Synchronous and asynchronous training of models | SEDIMARK should provide a decentralised training framework that supports training of ML models in both synchronous and asynchronous ways. | L | OPT | AI/ML |
| **Recommender Requirements** | | | | | |
| Req-RS-01 | User profiling | Recommender system needs access to user profile logs and user interaction history to be able to effectively compute user preferences and build a user profile. | H | REQ | Recom |
| Req-RS-02 | Rich item information | Recommender models need to know the available item set and have as many features, metadata and statistics as possible | H | REQ | Recom |
| Req-RS-03 | Decentralised Recommender system | Recommendation platform must be trained and run in a decentralised way, not leaking user private information. | H | REQ | Recom |
| Req-RS-04 | Cold start problem | Recommendation platform should be capable of addressing the cold-start problem for users. | M | REC | Recom |

| Identifier | Short name | Title | Priority | Req. Level | Category |
|---|---|---|---|---|---|
| **Energy Efficiency Requirements** | | | | | |
| Req-EE-01 | Lightweight and energy efficient DP modules | Data processing modules must be lightweight and energy efficient | H | REQ | Energy efficiency |
| Req-EE-02 | Lightweight and energy efficient AI/ML models | AI models must be small and lightweight to consume less energy both during training and inference | M | REC | Energy efficiency |
| Req-EE-03 | Energy efficient decentralized training of ML model | Decentralised training of ML models must be communication and energy efficient | H | REQ | Energy efficiency |
| Req-EE-04 | Usage of compiler optimizations for ML model | ML models should make use of compiler optimizations (e.g., Torch.compile() vs vanilla pytorch) | L | OPT | Energy efficiency |
| | | **Interoperability** | | | |
| REQ-INT-01 | Enable interoperability in different domains | The information model should allow a wide application scope and enable interoperability | H | REQ | Interoperability |

| Identifier | Short name | Title | Priority | Req. Level | Category |
|---|---|---|---|---|---|
| REQ-INT-02 | Data and metadata for information model | The information model should prescribe aspects of the data and the metadata | H | REQ | Interoperability |
| REQ-INT-03 | Meta data fields | Compliance with the information model should be a (partial) prerequisite for SEDIMARK data | H | REQ | Interoperability |
| REQ-INT-04 | Compliance with the information model | Data tools should be developed for enforcing compliance with the information model | H | REQ | Interoperability |
| REQ-INT-05 | Data tools for the information model | The information model should prescribe specific metadata fields | H | REQ | Interoperability |
| REQ-INT-06 | Interoperability of ML models | Provide ML interoperability by converting the AI models to a common format | H | REQ | Interoperability |
| **Storage** | | | | | |
| Req-STR-01 | Keep datasets in provider domain | Datasets must be stored within the provider domain. | H | REQ | Data Storage |

| Identifier | Short name | Title | Priority | Req. Level | Category |
|---|---|---|---|---|---|
| Req-STR-02 | Distributed metadata storage | The platform needs to be able to store metadata relating to providers, offerings, policies, data in a distributed catalogue. | H | REQ | Data Storage |
| Req-STR-03 | Temporary Storage | Intermediate data through data processing pipeline needs to be temporarily stored on disk or in-memory. | H | REQ | Data Storage |
| Req-STR-04 | Storage for pipeline artefact outputs | Data from final output of data processing pipeline needs to be stored in consumable manner. | H | REQ | Data Storage |
| Req-STR-05 | Scalable provision for large data generated | Providers collecting huge amounts of data require distributed data management for scalable provision. | L | OPT | Data Storage |
| Req-STR-06 | Outsourced Storage | Data from providers who cannot host data onsite needs to be stored via service. | M | OPT | Data Storage |
| Req-STR-07 | Knowledge graph persistence | Services creating knowledge graphs need to be able to store in appropriate persistence. | M | OPT | Data Storage |
| Req-STR-08 | Appropriate storage for offering type. | Storing options for each type of offering (data, features, model, configuration etc) or configuration need to be consistent. | L | REC | Data Storage |
| **Publication and discovery** | | | | | |

| Identifier | Short name | Title | Priority | Req. Level | Category |
|---|---|---|---|---|---|
| Req-P&D-01 | Assets described as part of offerings | Assets (e.g., datasets, data streams, services, …) have to be uniformly described as offerings to support its homogeneous publication and easy discovery | H | REQ | Publication and Discovery |
| Req-P&D-02 | Offerings' registry | Offerings have to be published on a common distributed registry from which they can be listed | H | REQ | Decentralization, Publication and Discovery, Openness |
| Req-P&D-03 | Generic offering metadata | Metadata included in offerings must be generic enough to avoid imposing restrictions on the assets data format nor the mechanisms used for its provision | H | REQ | Publication and Discovery, Openness |
| Req-P&D-04 | Open Data portal discovery | Some Open Data should be discoverable as part of the offering catalogue | H | REQ | Publication and Discovery, Openness |
| Req-P&D-05 | Offerings' catalogue for queries | Catalogue(s) should be constructed by fetching offering metadata and employ different indexing strategies based on provider/consumer requirements | H | REC | Decentralization, Publication and Discovery |
| **User interface (Marketplace UI / GUI Tools / whatever)** | | | | | |
| Req-UI-01 | Logging in UI | Users should be able to log in in the marketplace UI. | H | REQ | UI, Security |
| Req-UI-02 | Offerings discoverability | Users must be able to discover offerings in the marketplace based on identity-based policies defined by the asset provider. | H | REQ | UI, Publication and Discovery |

| Identifier | Short name | Title | Priority | Req. Level | Category |
|---|---|---|---|---|---|
| Req-UI-03 | Users Identity | Users have to be able to create, own and manage their own SSI identity | H | REQ | UI, Security |
| Req-UI-04 | Offerings management | Providers should be able to manage their offerings (adding / removing them) | H | REQ | UI |
| Req-UI-05 | Offering description page | Users should be able to view detailed information about offerings | H | REQ | UI |
| Req-UI-06 | SEDIMARK toolbox access in UI | Users should be able to interact with their offerings in the marketplace UI (visualize them, run them through the SEDIMARK toolbox, …) | M | REC | UI |
| Req-UI-07 | Rating offerings in UI | Users should be able to rate and review offerings in the marketplace UI. | L | REQ | UI |
| Req-UI-08 | Offering statistics in UI | Providers should be able to see statistics related to the access of their offerings | L | OPT | UI |
| **Smart Contract and Tokenization** | | | | | |
| Req-SCT-01 | Smart Contracts support | DLT shall support Smart Contracts | M | REQ | |

| Identifier | Short name | Title | Priority | Req. Level | Category |
|---|---|---|---|---|---|
| Req-SCT-02 | Tokenization of Assets | Assets shall be tokenized | M | REQ | |
| Req-SCT-03 | User Digital Wallet | Users must be able to own and manage a Digital Wallet | M | REQ | |