



SEcure Decentralised Intelligent Data MARKetplace

D4.5 Data sharing platform and incentives - First version

Document Identification	
Contractual delivery date:	31/12/2023
Actual delivery date:	31/12/2023
Responsible beneficiary:	ATOS
Contributing beneficiaries:	FV, SIE, NUID UCD
Dissemination level:	PU
Version:	1.0
Status:	Final

Keywords:

Marketplace, Data Sharing, User Interface, User Experience, Data Processing.



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.

The dissemination of this document reflects only the authors' view, and the European Commission or UKRI are not responsible for any use that may be made of the information it contains. **This deliverable is subject to final acceptance by the European Commission.**

This document and its content are the property of the SEDIMARK Consortium. The content of all or parts of this document can be used and distributed provided that the SEDIMARK project and the document are properly referenced. Each SEDIMARK Partner may use this document in conformity with the SEDIMARK Consortium Grant Agreement provisions.

Document Information

Document Identification			
Related WP	WP4	Related Deliverables(s):	
Document reference:	SEDIMARK_D4.5	Total number of pages:	60

List of Contributors	
Name	Partner
Maxime Costalonga Arturo Medela	ATOS
Eero Jalo	FV
Elias Tragos Aonghus Lawlor Diarmuid O'Reilly Morgan Erika Duriakova Honghui Du Qinqin Wang Neil Hurley	NUID UCD
Gabriel Danciu Stefan Jarcau	SIE

Document History			
Version	Date	Change editors	Change
0.1	20/07/2023	ATOS	First version of document structure (table of content)
0.2	22/11/2023	NUID UCD	Filled section 8 (Recommender)
0.3	08/12/2023	ATOS	Filled section 3, 4, 5 with frontend mock ups and descriptions Filled section 9 (open data enabler)

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	2 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

Document History			
Version	Date	Change editors	Change
0.4	11/12/2023	SIE, FV	SIE: Filled section 6, 7 (data processing and AI dashboards) FV: Filled section 10 (data sharing incentives)
0.5	15/12/2023	ATOS	Add intro, section 2 (overview) and conclusions Rework all references to figures and bibliography
0.6	22/12/2023	ATOS	Apply review comments
0.7	28/12/2023	ATOS	Quality Review Format
1.0	29/12/2023	ATOS	FINAL VERSION TO BE SUBMITTED

Quality Control		
Role	Who (Partner short name)	Approval date
Reviewer 2	Tarek Elsaleh (SURREY)	20.12.2023
Reviewer 1	Gilles Orazi (EGM)	22.12.2023
Quality manager	María Guadalupe Rodríguez (ATOS)	29.12.2023
Project Coordinator	Arturo Medela (ATOS)	29.12.2023

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	3 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

Table of Contents

Document Information	2
Table of Contents	4
List of Tables	6
List of Figures	7
List of Acronyms	8
Executive Summary	9
1 Introduction	10
1.1 Purpose of the document	10
1.2 Relation to another project work	10
1.3 Structure of the document	11
2 Marketplace overview	12
2.1 Scope and personas	12
2.2 Architecture	14
3 Onboarding and authentication	16
3.1 Home page	16
3.2 Sign in/out	17
3.3 Participant registration	18
4 Catalogue browsing	21
5 Offering provision and consumption	24
5.1 New Offering registration	24
5.1.1 Asset definition	24
5.1.2 Asset access	25
5.1.3 Pricing & policies	27
5.2 Offering management dashboard	27
5.2.1 Overview	28
5.2.2 Assets	28
5.2.3 Contracts	30
6 Data processing dashboard	33
7 AI/ML dashboard	37
8 Recommender	41
8.1 Overview	41
8.2 Overview of Recommender Systems	41
8.3 Design of the recommender module	42

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	4 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

8.3.1	Internal Structure of the Recommender module.....	42
8.3.2	Recommendation data flow	45
8.3.3	User and item profiling.....	46
8.4	Implementation.....	47
8.4.1	Overview	47
8.4.2	Query based recommendation.....	47
8.4.3	Item-based recommendation	49
8.5	Results.....	49
8.5.1	Query based recommendation examples.....	50
8.5.2	Item based recommendation examples	51
8.6	Future work	52
9	Open Data enabler.....	53
9.1	Architecture	53
9.2	Kaggle data offering	53
10	Data sharing incentives.....	55
10.1	City of Helsinki as a participant in the marketplace	55
10.2	Marketplace frontend features to foster data sharing	55
11	Conclusions	57
12	References	59

Document name:	D4.5 Data sharing platform and incentive – First version			Page:	5 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0
				Status:	Final



List of Tables

Table 1. SEDIMARK marketplace user stories12

Document name:	D4.5 Data sharing platform and incentive – First version			Page:	6 of 60		
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0	Status:	Final

List of Figures

Figure 1 - Relationship between D4.3 and other deliverables, tasks, and work packages. ...	11
Figure 2 - High level view of the SEDIMARK marketplace architecture (from Figure 15 in D2.2 [1])	14
Figure 3 - Marketplace home page	16
Figure 4 - Marketplace sign in (detail).....	17
Figure 5 - Marketplace registration form: account creation	18
Figure 6 - Marketplace registration form: user details	19
Figure 7 - Marketplace registration form: getting verifiable credentials	20
Figure 8 - Marketplace Catalogue browsing page.....	21
Figure 9 - Example of an Offering description page.....	23
Figure 10 - Marketplace Offering publication first step: asset definition (from scratch).....	24
Figure 11 - Marketplace Offering publication first step: Asset definition (reusing an Asset) ..	25
Figure 12 - Marketplace Offering publication second step: asset access	26
Figure 13 - Marketplace Offering publication third step: pricing and policies	27
Figure 14 - Marketplace Offering management dashboard: overview	28
Figure 15 - Marketplace Offering management dashboard: Assets	29
Figure 16 - Marketplace Offering management dashboard: Contracts (provided)	30
Figure 17 - Marketplace Offering management dashboard: Contracts (consumed)	31
Figure 18 - Date processing pipeline	33
Figure 19 - One block of the processing code	34
Figure 20 - Date processing pipelines UI representation	35
Figure 21 - Date processing variable selection form	36
Figure 22 - The AI processing pipeline	37
Figure 23 - The train block code	38
Figure 24 - The UI current implementation of the AI pipeline	39
Figure 25 - The configuration parameters for the Anomaly detection algorithm	40
Figure 26 - Stakeholders in RS (adapted from [5]).....	41
Figure 27 - Internal structure of the recommender module and its interactions with external modules.	43
Figure 28 - Recommendation service data flow example.....	45
Figure 29 - Open Data enabler architecture.....	53

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	7 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

List of Acronyms

Abbreviation / acronym	Description
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CC-BY	Creative Commons Attribution license
CKAN	Comprehensive Knowledge Archive Network
CSV	Comma-separated values file
DID	Decentralized Identifier
DNN	Deep Neural Networks
Dx.y	Deliverable number y belonging to WP x
EU	European Union
FAIR	To find, accessible, interoperable and reusable
GDPR	General Data Protection Regulation
GUI	Graphical User Interface
HTTP	HyperText Transfer Protocol
JSON	JavaScript Object Notation
LSI	Latent Semantic Indexing
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Natural Language Processing
ODRL	Open Digital Rights Language
RS	Recommender System
SVD	Singular Value Decomposition
URL	Uniform Resource Locator
WPx	Work Package x

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	8 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

Executive Summary

This document corresponds to the deliverable SEDIMARK_D4.5 of the SEDIMARK project, named “Data sharing platform and incentives – First version”. Its goal is to describe the SEDIMARK marketplace, a web frontend application constituting the entry point for users to the SEDIMARK ecosystem and all the functionalities it offers.

After a brief introduction of the scope of the marketplace and the expected users’ persona in **Chapter 2**, the subsequent chapters of this deliverable describe the graphical user interfaces of the marketplace, grouped by functionality:

- **Chapter 3**: onboarding of new users, as well as signing in or out.
- **Chapter 4**: browsing the Offering catalogue.
- **Chapter 5**: how users can manage their provided or consumed offerings.
- **Chapter 6**: accessing the data processing toolbox.
- **Chapter 7**: accessing the AI toolbox.
- **Chapter 8**: how the system to provide offering recommendations to users works.
- **Chapter 9**: how open data portals are made accessible in the Marketplace.
- **Chapter 10**: perspectives to incentivise data sharing.

At this stage of the project, the Marketplace, along with most of the other components of SEDIMARK, are currently being implemented. Consequently, all figures displaying user interfaces in this document are mock-ups used to guide the implementation of the frontends. As the project evolves, newly identified needs or constraints may arise and will likely require adaptations to the user interfaces. This is why a second version of this document will be provided by July 2025 of the project, in the form of new deliverable, SEDIMARK_D4.6.

Document name:	D4.5 Data sharing platform and incentive – First version			Page:	9 of 60	
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0	Status: Final



1 Introduction

1.1 Purpose of the document

This report represents the first version of the SEDIMARK's approach on what will be its data sharing platform, as the main entry point to the system from the outside world. Hence, it must touch base not only on the front-end users will interact with but also on added features such as the Recommender system and the Open Data enabler which are at the essence of the solution. Given the stage on the project execution, the contents hereby presented will be subjected to an evolution and thus a new version of the SEDIMARK data sharing platform will be provided in Month 34 (July 2025) in the Deliverable 4.6 (Data sharing platform and incentives. Final version).

Therefore, this document does not offer a fully functional depiction of this platform, just a high-level presentation of its constitutive components instead. In fact, in what refers to the Marketplace front-end a description will appear in the report, while as only the Recommender system and the open data enabler will be described also from a backend perspective. Thus, it is intended for a certain audience, mainly for members in the project consortium to employ it as the template to drive specific technical activities from other work packages within SEDIMARK.

1.2 Relation to another project work

This deliverable represents the main output that emanates from the work carried out in Task 4.5 (Open Data enabler) during the first period of the project execution (from M10, July 2023, to M15, December 2023). Figure 1 depicts the interaction of the activities within WP4 (Secure data sharing in a decentralized Marketplace) and the relationships with other work packages. As may be inferred from the graph, the work presented in this report relates to the outputs of the work done in Tasks 4.1 (decentralised infrastructure and APIs for Data Spaces), 4.2 (edge data processing and sharing), 4.3 (digital identities and data confidence) and 4.4 (ethical data sharing platform and services), which output appears in Deliverables 4.1 (Data sharing platform and incentives. Final version) and 4.3 (Edge data processing and service certification. First version), both of them issued as well in M15, December 2023.

In addition, the output of the work presented in this deliverable will represent a valuable input to the other technical work packages to conduct their very own activities related to development and ulterior validation via testing.

On the other hand, those work packages and their corresponding tasks will in turn provide their feedback to task 4.5 during the next stages of the project evolution and thus represent a solid contribution to drive its activities and present their results in Deliverable 4.6 (M34, July 2025).

Document name:	D4.5 Data sharing platform and incentive – First version			Page:	10 of 60		
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0	Status:	Final

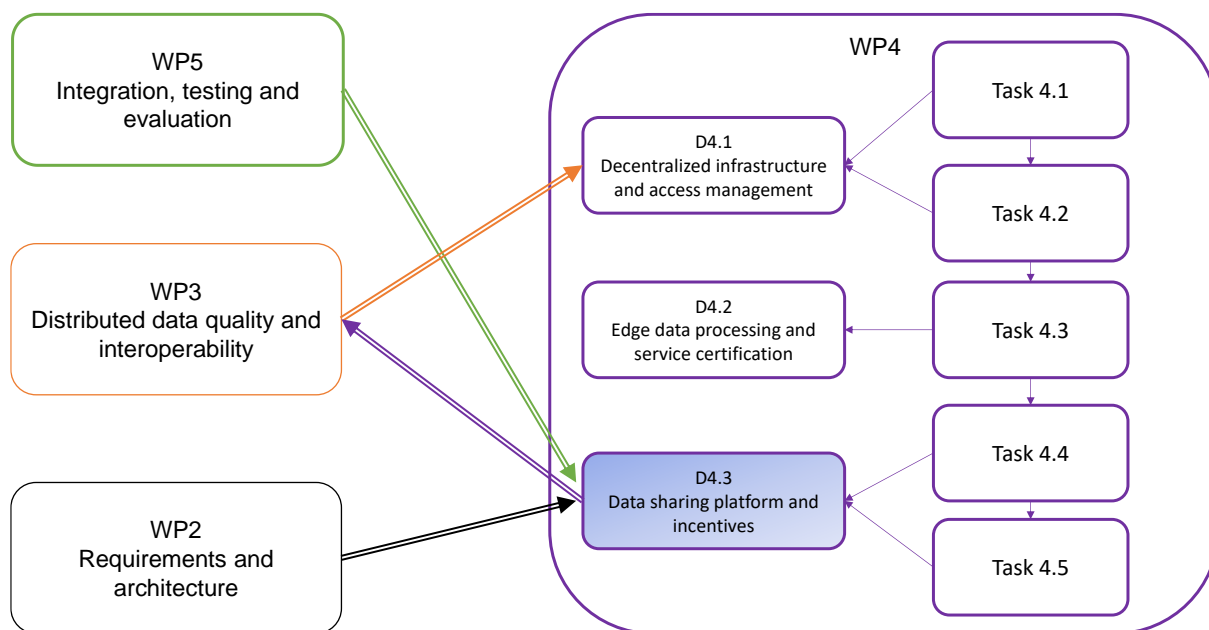


Figure 1 - Relationship between SEDIMARK_D4.3 and other deliverables, tasks, and work packages.

1.3 Structure of the document

This document is structured in 11 major chapters:

Chapter 1 is the current chapter and presents the introduction to the report.

Chapter 2 presents an overview on the SEDIMARK Marketplace, its constitutive parts and what kind of users it may have.

Chapter 3 introduces the way to proceed with the user onboarding and authentication into the SEDIMARK data sharing platform.

Chapter 4 includes an initial view on the Catalogue that will be offered by the platform and the options the user may find upon browsing its contents.

Chapter 5 depicts the appearance of the Offerings that will be shared by the data sharing platform, including an overview of its registration and further management.

Chapter 6 delves into the data processing dashboard incorporated into the data sharing platform to make it work as desired.

Chapter 7 discusses the Artificial Intelligence (AI) and Machine Learning (ML) dashboard incorporated into the operational flux of SEDIMARK's data sharing platform and how it will be used.

Chapter 8 devotes itself to analyse the Recommender system that is a constitutive part of the data sharing platform and will help users to find information of their interest within the catalogue.

Chapter 9 presents an overview on the Open Data enabler, its projected architecture and the kind of data offerings that will be presented in its first stages.

Chapter 10 establishes an initial debate on how the sharing process will take place as well as its implications.

Chapter 11 presents the conclusions of the report debating the main outcomes and the expected next steps.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	11 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

2 Marketplace overview

2.1 Scope and personas

The conception of SEDIMARK's Marketplace implies for it to act as the entry point to SEDIMARK functionalities and resources. Its architecture joins together a vast series of cooperating services and tools that provide various functionalities. This means the developments provided by technical work packages can be deployed independently and expose their functionalities through the marketplace itself.

To do so, the SEDIMARK Marketplace relies on a graphical user interface (GUI) that displays such set of services and offers users an easy access to log in or out of the application, perform the registration of both Participants and/or Offerings, carry out Contract negotiations, receive recommendations based on their profiles and much more actions that will be covered in further sections of this report. Marketplace users are expected to present diverse profiles and thus enjoy different roles and permissions to manage and operate with the resources available in the Catalogue, and also to configure and exploit the services offered. They fall in three groups:

- **Participants:** corresponding to users who registered to SEDIMARK and have been approved by the administrators. Participants can be *Providers* or *Consumers* of Offerings in the Marketplace Catalogue.
- **Visitors:** referring to non-registered users. They can only view the public Offerings of the Catalogue, and request to register in the SEDIMARK ecosystem.
- **Administrators:** defining the policies ruling the usage of SEDIMARK.

Table 1 below captures a brief description of what each one of those kinds of users could do in the platform through the depiction of basic user stories.

Table 1. SEDIMARK marketplace user stories

Ref. No.	As a (stakeholder)	I want to	So that	And is considered 'done' when
SM-US1	Participant, Visitor	Register an account in SEDIMARK Marketplace.	I may access the SEDIMARK Marketplace resources (provide/consume Offerings).	I receive the confirmation of a verified account from the Identity Manager component.
SM-US2	Participant	Check my user account information role and permissions.	I can manage my user account and check the permissions granted.	I receive the confirmation of a verified account from the Identity Manager component.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	12 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

Ref. No.	As a (stakeholder)	I want to	So that	And is considered 'done' when
SM-US3	Participant, Visitor	Browse the Catalogue of Offerings.	I can see datasets or services corresponding to my permissions.	I can access / search / filter the Offerings in the Catalogue via the marketplace GUI.
SM-US4	Participant (Provider)	Register a dataset Offering.	I can put make my Offering accessible to other Participants, with the access and usage policies of my choice.	My Offering is accessible in the Catalogue, and its usage policies are enforced.
SM-US5	Participant (Consumer)	Purchase a dataset Offering.	I can access the content of the Offering.	I can transfer the data on my premises.
SM-US6	Participant (Consumer)	Get recommendations of Offerings.	I can quickly get informed of Offerings matching my interests.	The Marketplace provides me with such recommendations upon browsing the Catalogue.
SM-US7	Participant (Provider)	Monitor the status of my provided Offerings.	I can keep track of my contracts, active or expired, and get statistics for all of them.	I can access a dashboard listing all the Contracts corresponding to Offerings I provide.
SM-US8	Participant (Consumer)	Monitor the status of my consumed Offerings.	I can keep track of my Contracts, active or expired, and get statistics for all of them.	I can access a dashboard listing all the Contracts corresponding to Offerings I consume.
SM-US9	Participant	Control the data transfer of my active Contracts.	I can start or cancel the transfer of data.	I can start or cancel my data transfer in the Marketplace GUI.
SM-US10	Participant	Access the data processing dashboard.	I can use the SEDIMARK data processing toolbox.	I can access the data processing toolbox from the Marketplace GUI.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	13 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

Ref. No.	As a (stakeholder)	I want to	So that	And is considered 'done' when
SM-US11	Participant	Access the AI dashboard.	I can use the SEDIMARK AI toolbox.	I can access the AI toolbox from the Marketplace GUI.
SM-US12	Administrator	Validate new users coming into the Marketplace.	I can assign roles and permissions and check existing ones.	I assign new roles and permissions to incoming users.

It is worth noting that both the functionalities and roles identified at this phase will be expanded and improved in further stages of the project execution, according to the progress in the specification and development of the diverse services and taking into consideration requirements posed by SEDIMARK use cases and their end users. For instance, at this stage of the SEDIMARK development, only dataset offerings will be supported, and the administrator controls in the GUI very limited.

2.2 Architecture

In here readers will find a succinct description of the components which comprises the SEDIMARK Marketplace, which constitutes the soul of its data sharing platform. As described in deliverable SEDIMARK_D2.2 “Architecture and Interfaces. First version” [1], this is one of the constitutive parts of the overall SEDIMARK architecture and relates to the items presented in the green box in Figure 2.

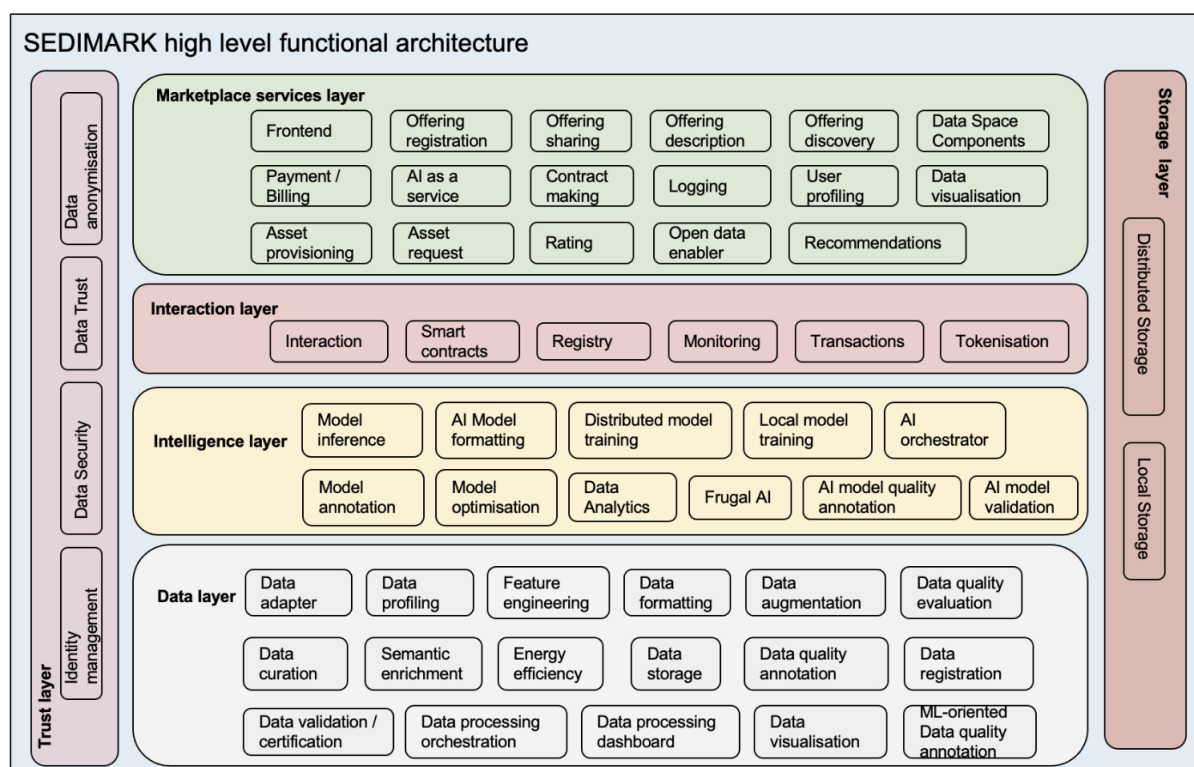


Figure 2 - High level view of the SEDIMARK marketplace architecture (from Figure 15 in SEDIMARK_D2.2 [1])

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	14 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

Even though the picture does not delve into the fine details on how each component in such Marketplace services layer interfaces with the rest of them, neither goes into the close definition of connections, the descriptions coming in following chapters of the report make explicit how they all fit together, as well as how they interact with the other layers of the overall SEDIMARK platform. In fact, and starting from the premises established in SEDIMARK_D2.2, it is possible to divide the services layer into two parts, namely: Offering management and Marketplace.

Hence, within the former group the modules included go by the naming:

- **Offering description** to create and/or edit descriptions of data and services apt to be exchanged.
- **Offering registration** to perform the embarkment of data/services into the local Catalogue and confirm they comply with the SEDIMARK rules.
- **Offering discovery** to let users navigate through the Marketplace Catalogue and find the data and services closest to their interests.
- **Offering sharing** to ease the way data and services will be distributed.
- **Local Catalogue** to compile the complete collection of Offerings.
- **Open Data enabler** to make available in the SEDIMARK Marketplace datasets coming from diverse open data portals.

While as the latter cluster embarks services such as:

- **Marketplace GUI** offers the window into SEDIMARK from the outside world.
- **Logging** establishes the protocol to let registered users (Providers, Consumers, Administrators) get into the platform.
- **Contracting** sets up the policies to proceed with the acquisition of a certain Offering by an interested user.
- **User profiling** keeps a log of users' activity within the Marketplace to provide relevant input to the Recommendation service.
- **Service provisioning** to make resources available to conduct a specific activity.
- **Service request** eases the procedure to present a formal request for certain service to be provided.
- **AI as a Service** interacts with the Intelligence layer to interface with the AI orchestrator with the aim to let users perform machine learning tasks on their Offerings or as a service.
- **Payment** connects with the corresponding gateway to complete the transaction once a contracting takes place.
- **Recommendations** suggest users particular Offerings that may be of their interest.

As anticipated, the rest of sections in this report cover the main areas of the SEDIMARK Marketplace and offer detailed views on them all.

Document name:	D4.5 Data sharing platform and incentive – First version			Page:	15 of 60		
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0	Status:	Final

3 Onboarding and authentication

In this chapter, we describe the home page of the Marketplace, as well as how Participants can log in their account and the onboarding procedure for new users.

3.1 Home page

As a showcase of SEDIMARK's Marketplace, the home page's targeted audience is primarily prospective users. Consequently, its main role is to provide an overview of the activities in the Marketplace and encourage visitors to browse the Offering Catalogue. It is also the perfect place to display some news about SEDIMARK's use cases or more broadly about recent developments or outreach of the project.

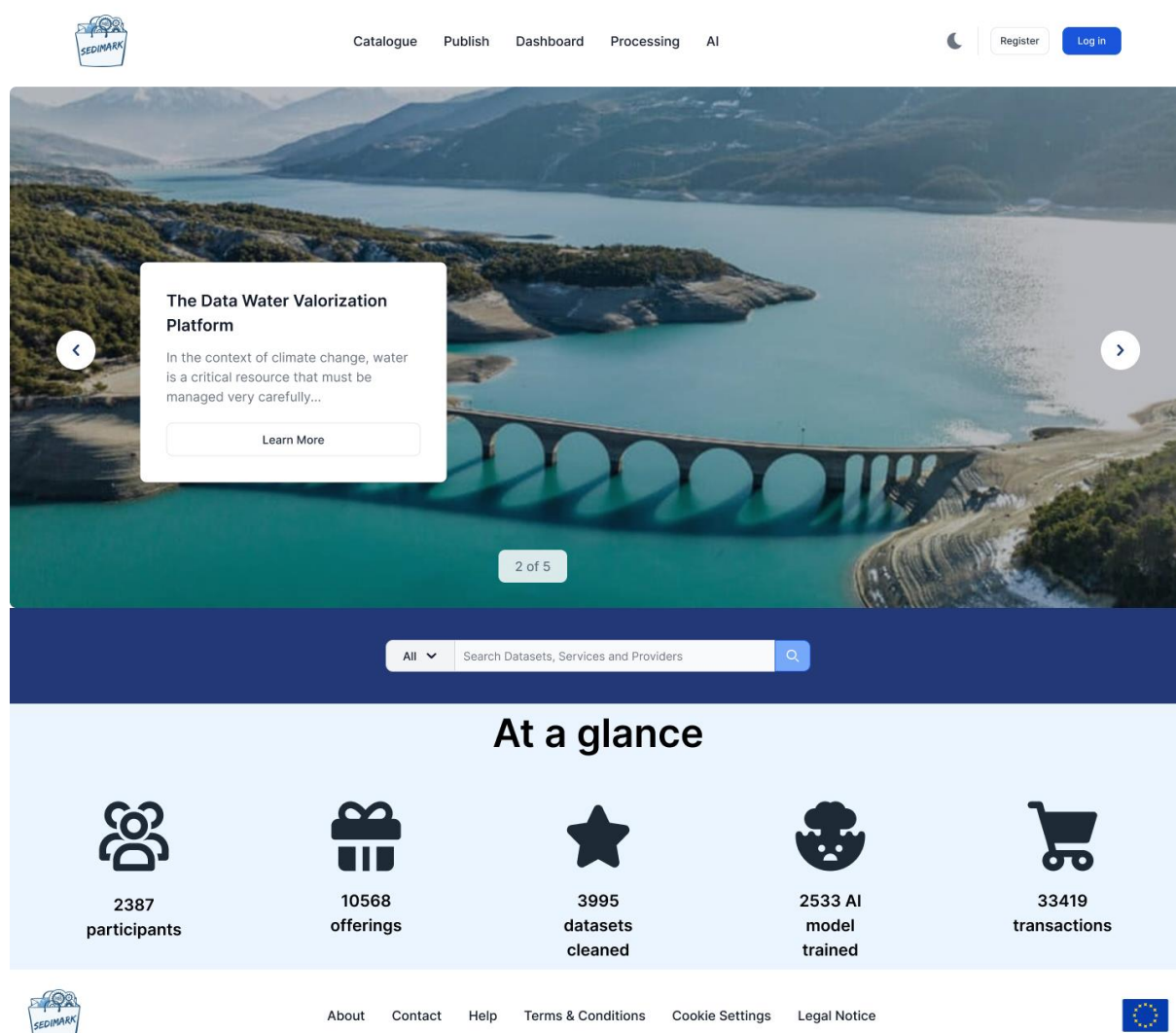


Figure 3 - Marketplace home page

The first version of the Marketplace home page shown in the figure above answers all these calls by showcasing SEDIMARK's latest insights through a carousel. Each of its pages will point towards an article posted in the official SEDIMARK website or to any other relevant resources such as blogposts of project consortium partners, social media article or scientific publication.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	16 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

In addition to its possible access via the top navigation bar, a quick search input has been placed in the centre of the page to foster Catalogue browsing by prospective users. This search is kept as simple as possible: it just requires a text input to match with Offering names or keywords, and can optionally filter the Offerings via a dropdown menu, in case users would want to browse only datasets or services.

Finally, the dynamism of the Marketplace is emphasized using a set of quantitative facts: number of Participants, Offerings, transactions. This overview will be refined as the Marketplace gets implemented.

3.2 Sign in/out

Participants of the SEDIMARK ecosystem can authenticate themselves upon hitting the always accessible *log in* button in the navigation bar. This redirects them towards the very simple sign in form shown below.

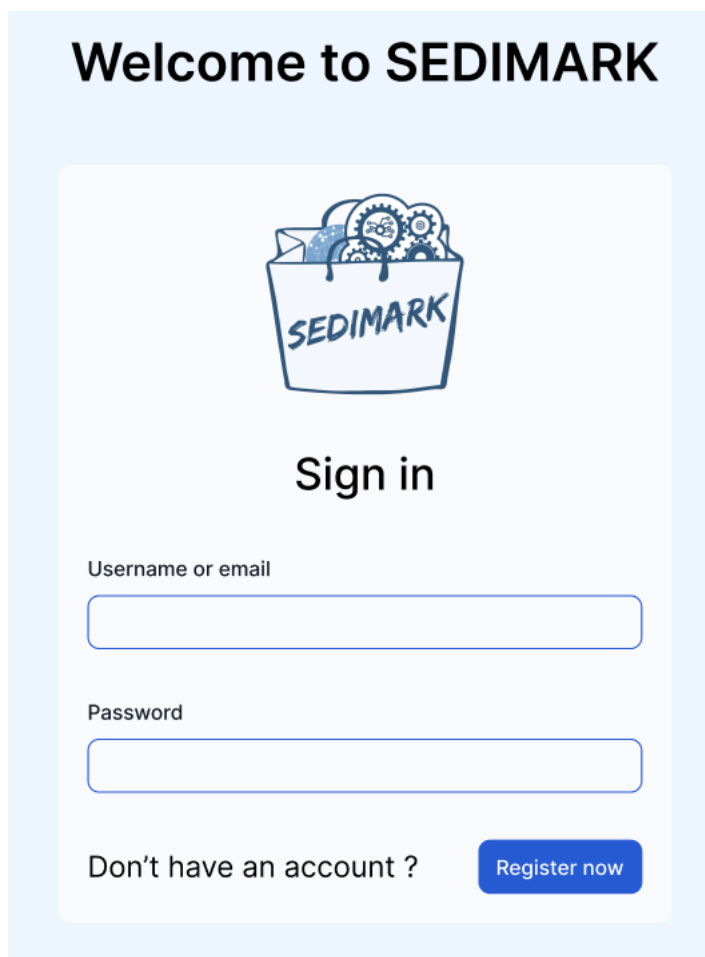


Figure 4 - Marketplace sign in (detail)

Although already present in the navigation bar, another button to enter the registration process has been placed at the bottom of this form for new users' convenience.

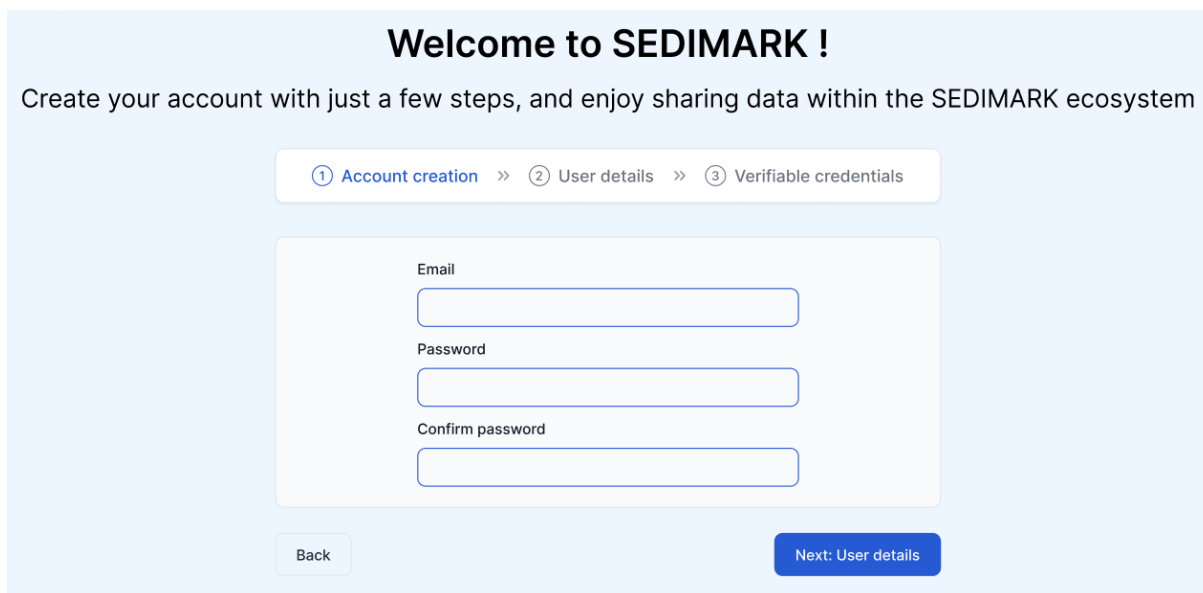
Upon signing in, participants will be asked to connect their wallet using the MetaMask browser extension [2].

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	17 of 60				
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0	Status:	Final

3.3 Participant registration

New users wishing to join the SEDIMARK ecosystem can do so in three simple steps:

1. Creating an account.
2. Providing details to create their decentralized identifier (DID).
3. Receiving their DID and verifiable credentials.



Welcome to SEDIMARK !

Create your account with just a few steps, and enjoy sharing data within the SEDIMARK ecosystem

① Account creation >> ② User details >> ③ Verifiable credentials

Email

Password

Confirm password

Back Next: User details

Figure 5 - Marketplace registration form: account creation

The creation of an account only requires an email and a password. Users will also be requested to install the MetaMask browser extension in order to connect their wallet.

Once done, the next step asks users to provide a few more details for two purposes: the DID creation and the account customisation. Since a DID is created for all new users, these data are mandatory and subsequently marked by an asterisk (*). The inputs shown in the figure below are directly inspired from the ones required by Gaia-X to create a DID [3], but may not fully correspond to the needs of SEDIMARK: this form is expected to change during the Marketplace implementation. This also holds for the account customisation section. In the latter, most, if not all, inputs are optional. They are only intended to enable users to improve their image in the Marketplace, as it will appear when they interact with other participants.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	18 of 60				
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0	Status:	Final

Welcome to SEDIMARK !

Please provide some details so we can create your DID.

① Account creation >> ② User details >> ③ Verifiable credentials

Organization details

Legal name*

Legal Registration Number*:

EORI
 EUID
 lei code
 tax ID
 vat ID

Parent organization

Sub organization

Headquarters address*

Legal address*

I agree to the [terms & conditions*](#)

Customize your account


First name

Last name

Upload avatar picture

Choose file

No file chosen



Back

Next: Verifiable credentials

Figure 6 - Marketplace registration form: user details

Once the user has completed this step, her/his DID gets created and his verifiable credentials as a participant issued and stored in his/her wallet. For the sake of transparency, users can review them and copy or download them. Upon hitting the *Finish* button, users get redirected to the dashboard page.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	19 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

Welcome to SEDIMARK !

Your onboarding is complete! Check your DID and Verifiable Credentials.

① Account creation >> ② User details >> ③ Verifiable credentials

DID

```
{
  "@context": [
    "https://www.w3.org/ns/did/v1",
    "https://w3id.org/security/suites/ed25519-2020/v1"
  ],
  "id": "did:example:123456789abcdefghi",
  "authentication": [
    {
      "id":
        "did:example:123456789abcdefghi#keys-1",
```

Legal Person Credentials



Participant Credentials



Back

Finish

Figure 7 - Marketplace registration form: getting verifiable credentials

Participants will be able to modify their personal data or password at any time using the menu items accessible upon clicking their avatar in the navigation bar.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	20 of 60				
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0	Status:	Final



4 Catalogue browsing

This chapter introduces the Catalogue of Offerings can be browsed in the Marketplace. Any user, being an authenticated Participant or a simple visitor, can access the Catalogue, either by performing a quick search in the home page, or by hitting the *Catalogue* button in the navigation bar. The content of the Catalogue however may vary depending on the user: for instance, visitors may not see restricted access Offerings.

Figure 8 - Marketplace Catalogue browsing page

As shown in Figure 8, the Catalogue browsing page consists of a simple list of Offerings, whose display format can be picked between tiles or one Offering per row using the top left button bar. As for the onboarding, we strive to keep the interface as simple as possible, aiming at

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	21 of 60				
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0	Status:	Final

providing a similar user experience as the one of other online Marketplaces such as eBay or Amazon.

The search bar is the same as the one in the home page: simply requiring some words to be matched with the Offering title, description or tags/keywords, and a dropdown to choose whether to search only for datasets, services or providers. The side bar enables users to further filter or sort the results of the search, by default sorted by relevance. Users can shrink the results list by entering price or creation date ranges, as well as unchecking locations, providers or tags they are not interesting in.

Each Offering in the result list is concisely described by its title, a short description phrase and a set of facts associated with icons, such as whether the offering is a dataset or a service and who published it, with a light emphasis on its price and creation date. A more detailed description can be accessed by expanding the Offering entry.

On top of the result list, an Offering is highlighted (in blue on Figure 8), outcoming from the Recommender system, based on the user's browsing habits and preferences. She/he can like or dislike the suggestion, therefore providing feedback to the Recommending system. He/she can also request another suggestion or simply discard them altogether.

Upon selecting an Offering from the results, the user is redirected towards a page detailing it, an example of which is shown on Figure 9. A separate card on the left side, following the user as she/he scrolls through the description of the Offering, indicates the price of the Offering (which can be free) as well as its expiry date and the terms and conditions associated with it. The user can also click on the Provider's name or icon to see his profile page or contact him/her. The verifiable credentials attesting of the creation of the Offering can also be checked, copied and downloaded, although by default its tab will be collapsed to avoid cluttering the view.

Finally, the user can like or dislike the Offering, an information which can be used to further nail down future offering recommendations. Like existing Marketplaces, a row of similar Offerings will also be suggested to the user in this page.

Document name:	D4.5 Data sharing platform and incentive – First version			Page:	22 of 60		
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0	Status:	Final



← Back to search

IMDb Top Rated English Movies

👍 23 🗣️ 1

Dataset | 📍 London, UK | 📅 Published 2 weeks ago | 🔄 Updated 3 days ago



Top rated movie in the UK, between 2010 and 2020

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ridiculus mus mauris vitae ultricies leo integer malesuada nunc. Venenatis tellus in metus vulputate eu scelerisque felis. Condimentum vitae sapien pellentesque habitant morbi tristique.

2€

🕒 Accessible for three months

[Terms & Conditions](#)

[Buy](#)

Database entry description

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ridiculus mus mauris vitae ultricies leo integer malesuada nunc. Venenatis tellus in metus vulputate eu scelerisque felis. Condimentum vitae sapien pellentesque habitant morbi tristique.

Cinema CSV

Offering credentials

```
{
  "@context": [
    "https://www.w3.org/ns/did/v1",
    "https://w3id.org/security/suites/ed25519-2020/v1"
  ],
  "id": "did:example:123456789abcdefghi",
  "authentication": [
    {
      "id": "did:example:123456789abcdefghi#keys-1",
```

Provided by



IMDb Ltd

✉ example@imdb.com

☎ +33 8 36 65 65 65

Downloaded

8

times

You may also like ...

Eviden HPC

👍 23 🗣️ 1 ✕

128 core 1 TB memory cluster, across 4 servers, to host your apps. Pay only when apps are active.

€ 2 euros

📍 Paris, FR

📅 2023-11-09

🔔 Service

👤 Eviden SA

Eviden HPC

👍 23 🗣️ 1 ✕

128 core 1 TB memory cluster, across 4 servers, to host your apps. Pay only when apps are active.

€ 2 euros

📍 Paris, FR

📅 2023-11-09

🔔 Service

👤 Eviden SA

Eviden HPC

128 core 1 TB memory cluster, across 4 servers, to host your apps. Pay only when apps are active.

🔔 Service



Figure 9 - Example of an Offering description page

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	23 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

5 Offering provision and consumption

This chapter describes how SEDIMARK Participants can create, consume and manage their Offerings in the Marketplace.

5.1 New Offering registration

An authenticated Participant can publish a new Offering in the Catalogue using the *Publish* button in the navigation bar. She/he is then redirected to a wizard consisting of three steps:

1. Defining the Asset: to indicate whether the Offering should reuse an existing Asset or not, and provide a high level description of its content.
2. Setting the access to the Asset: more precisely, to parametrize the query that will allow the SEDIMARK components to fetch the Asset.
3. Associating a price and policies to the Offering.
4. Reviewing and submitting the Offering.

For clarity, all screenshots displayed in this section focus only on these steps. Repetitive elements of the frontend, such as the navigation bar and footers, are removed. The overall pages can be freely checked in the Figma files.

5.1.1 Asset definition

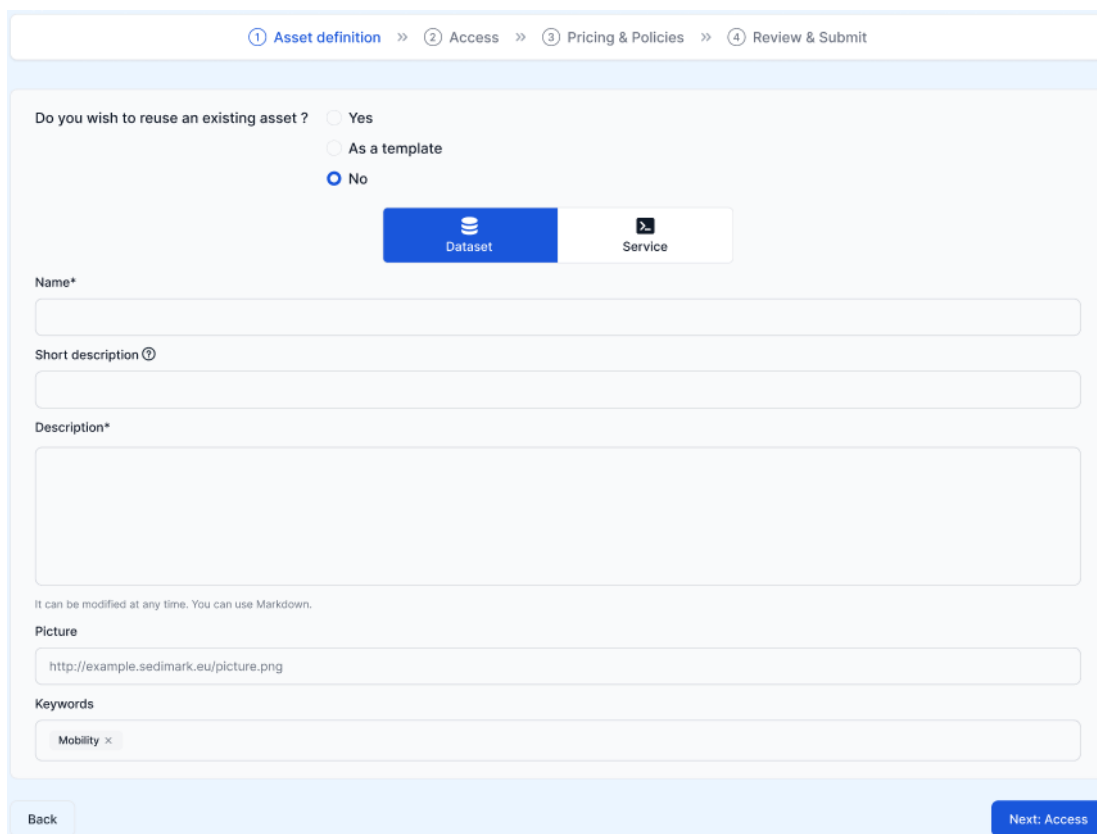


Figure 10 - Marketplace Offering publication first step: asset definition (from scratch)

The first step to the publication of an Offering is to define the Asset the Offering is about. The user is offered the possibility to reuse an existing Asset, either directly (see Figure 11) or as a template, or create a new one from scratch (see Figure 10). Reusing an Asset directly is useful

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	24 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

when renewing an Offering which has expired, while reusing as template simply provide a convenience by prefilling the Asset definition and access steps with the information from the selected existing Asset. If the user chose to reuse the Asset directly, this information will not be modifiable, and the user can move directly to step 3 (pricing and policies).

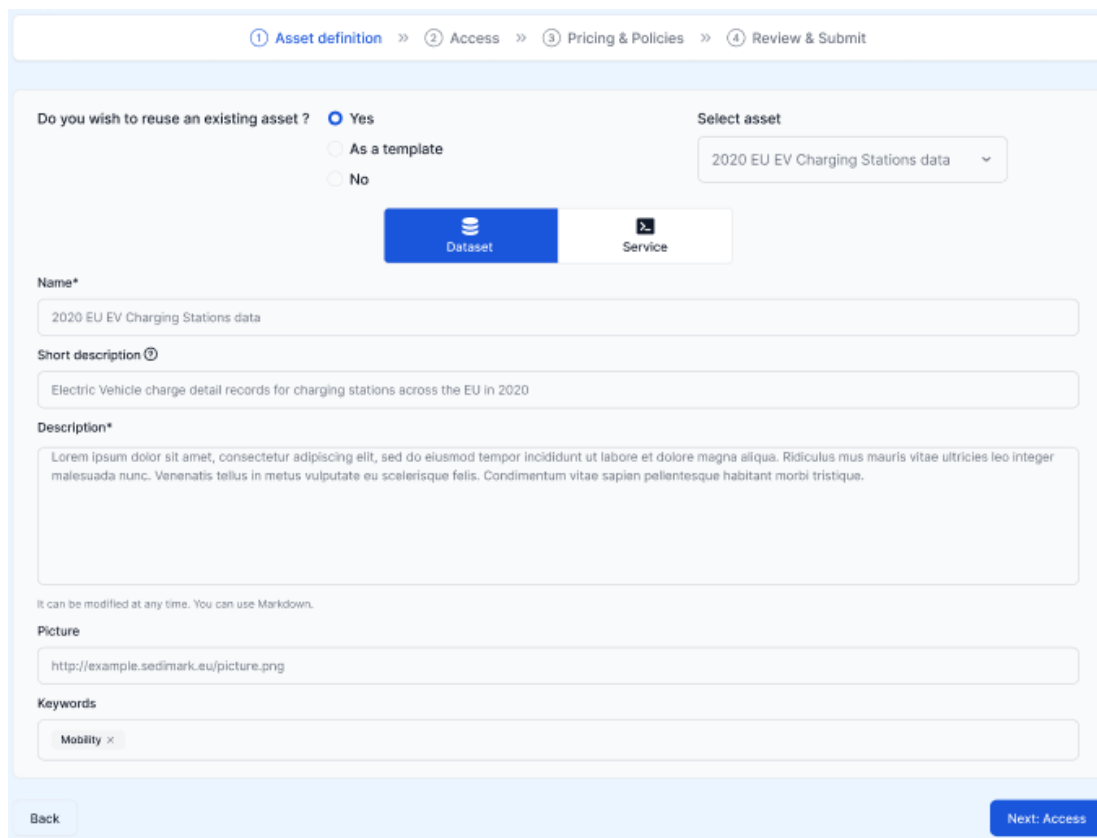


Figure 11 - Marketplace Offering publication first step: Asset definition (reusing an Asset)

When creating a new Asset from scratch, this first definition step only requires a high-level description of the Asset, a title and some tags/keywords. These will primarily be used to populate the Offering description prospective consumers will access from the Catalogue.

For its first version, SEDIMARK will only support one type of Offering: datasets. The possibility of creating a *service* Offering will be added in a subsequent iteration.

5.1.2 Asset access

After the high-level description of the Asset, the user needs to indicate how to actually access it by filling the form shown in Figure 12. The first version of SEDIMARK will only support providing URL to access the Asset resources: the other entries (GraphQL, MQTT) are shown as placeholders and might not reflect in the final version of the Marketplace to be delivered at the end of the project.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	25 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

① Asset definition >> ② Access >> ③ Pricing & Policies >> ④ Review & Submit

Access Type

URL
GraphQL
MQTT

GET
http://example.sedimark.eu/api/v1/data/

Headers

-

+

Query parameters

Required

+

Rights & Usage

License

Terms & conditions

The dataset contains personally identifiable information?

Back
Next: Pricing

Figure 12 - Marketplace Offering publication second step: asset access

The first part of this form corresponds to a graphical tool to create an HTTP request. The user is only required to provide an URL and a method to access it (GET, POST). He/she can optionally provide headers. The possibility to expose query parameters for the Consumer to enable her/him to refine the request, is an additional experimental feature under consideration.

In the second part of this form, users can associate a license to their datasets, as well as a link pointing towards some terms and conditions, which will be accessible from the online Offering page. Moreover, in order to comply with GDPR, if the dataset contains personally identifiable information, the provider can indicate who is the data controller, the purposes of the data collection, as well as the personal data protection regime and some contacts reachable for more information and content withdrawal.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	26 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

5.1.3 Pricing & policies

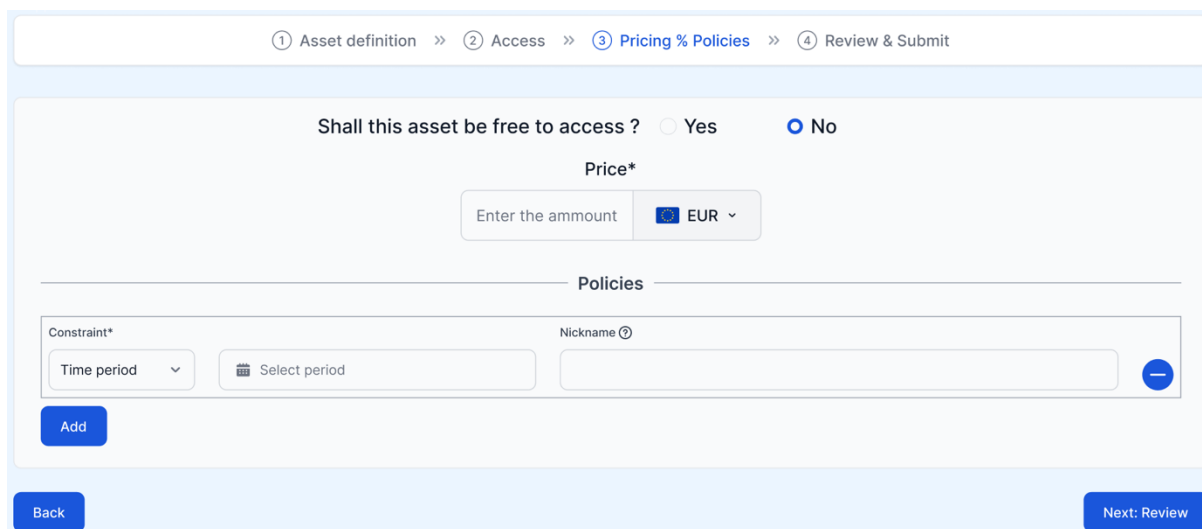


Figure 13 - Marketplace Offering publication third step: pricing and policies

Now that the user has described its Asset and how to access it, it is time to add the last layer that will turn this Asset into an Offering other Participant can consume, by setting the policies ruling its usage.

At first, in an attempt to promote open access, the user is asked whether or not the Offering should be free of charge. If not, a price input will be shown, where any amount can be displayed, in either fiat or crypto currencies.

The next part of this form is dedicated to the setting of policies to refine the data usage rights. The initial version of SEDIMARK will allow only a time period policy, i.e. enabling user to select a period of time during which their dataset should be accessible in the Catalogue. As the SEDIMARK project goes on, a wider panel of policies will be offered to users, following the Open Digital Rights Language (ODRL) policy information model [4], allowing them, for instance, to restrict data usage to specific locations.

Once done with setting the pricing and validity period of her/his dataset, the user can move on the last step: reviewing the Offering and submitting it for publication in the Catalogue. Since this step simply shows the Offering as it should appear online, alongside a submit button, it is not shown here. Such a preview of a Catalogue offering can be checked by the reader in section 4.

5.2 Offering management dashboard

The *Dashboard* button located in the navigation bar enables authenticated users to access, at any time, to platform to manage their consumed and provided Offerings. This dashboard is composed of 4 tabs, accessible from a side bar:

- **Overview**: for quick facts about the participant's usage of the platform.
- **Assets**: for the participant to manage its provided Assets.
- **Policies**: to enable users to create policy templates to be used when creating new offerings. This won't be implemented in the first version of the platform.
- **Contracts**: where all consumed and provided Offerings can be monitored.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	27 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

The side bar also features three menu icons, to respectively collapse/expand it, access the documentation of the Marketplace and the settings of dashboard. This feature, left for future versions of SEDIMARK, will enable users to customise their dashboards.

5.2.1 Overview

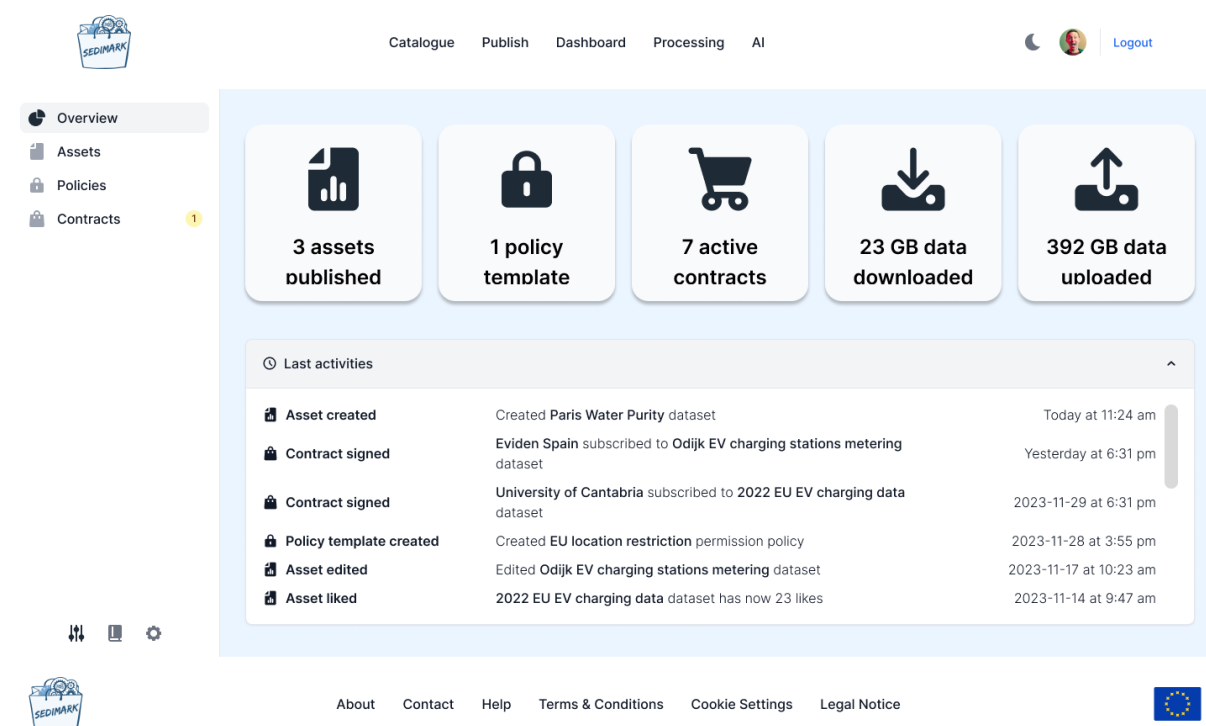


Figure 14 - Marketplace Offering management dashboard: overview

As shown in the figure above, this interface is split in two parts: a row of quantitative facts about the Participant usage, and a log of the last activities in the Marketplace relevant to the user. The latter may contain entries reporting actions taken by the user itself, such as creating or editing Assets, or by her/his partners, for instance when the latter purchase an Offering provided by the user or access the users' Assets.

At this stage of the project, this interface is kept very simple. It is subject to significant changes in the future and will be improved during feedback collection from early users.

5.2.2 Assets

This part of the Offering management dashboard enables users to review and edit their provided Assets. The latter are displayed as a list, in which each entry is succinctly described by its title, access URL, type, creation and last edit dates, and finally the number of Contracts it is involved in, highlighting how many of these are still active (i.e. not expired). The user can edit an Offering using the pen button or be redirected to the list of Contracts involving an Asset using the forward arrow button.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	28 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final



The screenshot displays the 'Assets' management dashboard. At the top, there are navigation links: Catalogue, Publish, Dashboard, Processing, AI, and a user profile with a 'Logout' button. The main content area features a search bar, a pagination control (1, 2, 3), and a list of five asset offerings. Each offering card includes the asset name, URL, type (Dataset or Service), creation date, last update date, and contract status (e.g., '3 contracts, 3 active'). Edit and refresh icons are present for each card. The left sidebar contains a navigation menu with 'Overview', 'Assets', 'Policies', and 'Contracts'. Below this is a 'Sort' section with radio buttons for 'Creation date', 'Last edit date', 'Name', and 'Contracts'. A 'Filters' section has checkboxes for 'Datasets' and 'Services', and two 'Select period' date pickers. At the bottom of the sidebar, there is a 'Hide inactive' toggle. The footer includes the SEDIMARK logo, navigation links (About, Contact, Help, Terms & Conditions, Cookie Settings, Legal Notice), and the European Union flag.

Figure 15 - Marketplace Offering management dashboard: Assets

The Asset list can be filtered and sorted using the additional menu actions in the side bar. For instance, users can, within two clicks, see only the list of their currently active dataset Offerings. For users who created a lot of Assets, a pagination tool as well as a search bar are provided to further facilitate browsing.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	29 of 60				
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0	Status:	Final

5.2.3 Contracts

The screenshot displays the 'Contracts' management dashboard. At the top, a yellow banner alerts that a contract with 'Example Consulting Ltd' on '2020 EU EV Charging Stations data' expires in 11 days. Below this, there are buttons for 'Consumed' and 'Provided'. The main area lists several contracts:

- Odijk EV Charging Stations Metering**: URL: http://example.sedimark.eu/data.json, Dataset, 2023-12-03, expires 2024-06-03, Free, 2 requests. Events include 'Transfer completed', 'Transfer started', 'Transfer cancelled', and 'Contract signed'.
- Example Consulting Ltd**: Contact info: example@imdb.com, +33 8 36 65 65 65.
- Eviden HPC**: URL: http://example.sedimark.eu/api/v1/, Service, 2023-11-09, expires 2024-11-09, 2 euros, 5 requests.
- 2020 EU EV Charging Stations data**: URL: http://example.sedimark.eu/data.json, Dataset, 2023-09-15, expires 2023-12-14, Free, 12 requests. Status: Expires soon.
- 2020 Germany Covid-19**: URL: http://example.sedimark.eu/data.json, Dataset, 2022-11-12, expires 2023-11-12, Free, 24 requests. Status: Expired.

The sidebar on the left includes navigation (Overview, Assets, Policies, Contracts), sorting options (Creation date, Expiry date, Price, Requests), filters (Datasets, Services), and grouping options (Assets, Partners).

Figure 16 - Marketplace Offering management dashboard: Contracts (provided)

The Contracts management dashboard is very similar to the asset one described previously. It also consists of a list of Contracts, but the latter is split in two parts, to separate consumed contracts from provided ones. Moreover, since contracts are associated with many events during their lifetime (agreement of both parts, access to the assets, expiry), an alert banner has been placed on top of the list to emphasize the most important ones, such as a contract expiring soon (see Figure 16) or a successful data transfer following the consumption of an Offering (see Figure 17).

To facilitate the retrieval of a Contract, every entry in the list is described by a heading composed of the Asset name and the other SEDIMARK Participant involved in the transaction.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	30 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

A non-expanded item also displays the asset type, the creation and expiry date of the Contracts, as well as the associated price. Other relevant information, such as the number of requests to Access the asset, may be displayed as well.

The screenshot shows the SEDIMARK Marketplace Offering management dashboard. The top navigation bar includes 'Catalogue', 'Publish', 'Dashboard', 'Processing', and 'AI'. A sidebar on the left contains 'Overview', 'Assets', 'Policies', and 'Contracts' (with a count of 1). Below the sidebar are sorting options (Creation date, Expiry date, Price, Requests) and filters (Datatasets, Services). The main content area features a green success message: 'Successfully transferred data from Breast cancer data (2022 update)'. Below this is a 'Consumed' button and a search bar. The main list displays four contracts:

- Breast cancer data (2022 update)**: Dataset, URL: http://example.careandco.com/data.json, Partner: Care & Co SARL, Created: 2023-12-03, Expires: 2024-06-03, Price: 10 euros, Requests: 1. Includes a 'Start transfer' button and a log of events: Transfer completed (Today at 11:24 am), Transfer started (Yesterday at 6:31 pm), Transfer cancelled (2023-11-29 at 6:31 pm), Transfer started (2023-11-29 at 6:28 pm), Purchased completed (2023-11-14 at 9:47 am), and Contract signed (2023-11-17 at 10:23 am). Contact info for Care & Co SARL is also shown.
- Eviden HPC**: Service, URL: http://example.sedimark.eu/api/v1, Partner: Bruxelloise des logiciels SRL, Created: 2023-11-09, Expires: 2024-11-09, Price: 2 euros, Requests: 5.
- 2020 EU EV Charging Stations data**: Dataset, URL: http://example.sedimark.eu/data.json, Partner: Example Consulting Ltd, Created: 2023-09-15, Expires: 2023-12-14, Price: Free, Requests: 12.
- 2020 Germany Covid-19**: Dataset, URL: http://example.sedimark.eu/data.json, Partner: Company GmbH, Created: 2022-11-12, Expires: 2023-11-12, Price: Free, Requests: 24.

The footer contains 'About', 'Contact', 'Help', 'Terms & Conditions', 'Cookie Settings', 'Legal Notice', and the European Union flag.

Figure 17 - Marketplace Offering management dashboard: Contracts (consumed)

Each Contract can be expanded to reveal additional information such as a log of the events around it (purchase completion, data transfer status, ...) and the partner's contacts. In the case of consumed Contracts (Figure 17), a button bar is added to start/cancel the data transfer.

As for the Asset management dashboard, convenient actions to sort and filter Contracts are available in the dashboard. Since it is expected that the number of Contracts of a given user

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	31 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

can be large, another feature to enable them to group contracts by partners or by Assets have been added.

Document name:	D4.5 Data sharing platform and incentive – First version			Page:	32 of 60		
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0	Status:	Final

6 Data processing dashboard

The UI representations of the pipeline orchestration will be designed to allow the user to create and maintain a clear and organized workflow for data processing, from initial data loading to final data export and analysis. The workflows will evolve during the development of SEDIMARK and beyond the project, as more complex requirements will be formulated by the stakeholders and general users of the Marketplace. In this chapter we will introduce a simple workflow described in what follows:

1. A custom Python script is responsible for creating necessary prerequisite for subsequent data processing task.
2. Data loader is the block that according to its name reads data from a given source and creates Python Pandas data frames to be used later in the pipeline. The sources can be either CSV, JSON files or other services like a broker service or streaming services.
3. Data transformation is a process that performs various alterations of the loaded data. This includes data curation, null values transformations, outlier detection, data imputation etc.
4. Data exporting represents a step of the pipeline that reads the transformed data and stores it into the destination format: CSV, JSON, JSON-LD, etc. This can be stored via an on premise storages or distributed via REST calls.

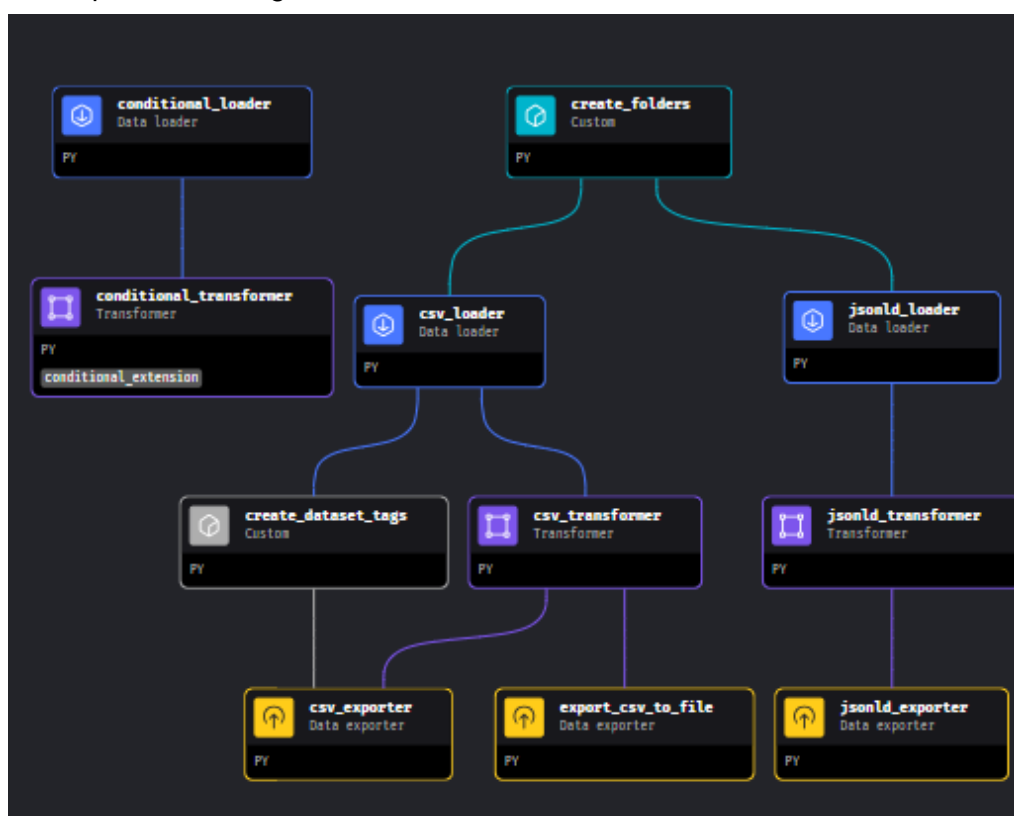


Figure 18 - Data processing pipeline

Figure 18 presents a sample of such pipelines as described in more detail in deliverable SEDIMARK_D3.1.

The code running in each block is easy to configure, so for developers the task of setting and handling these blocks of codes like the one presented in Figure 19 is a manageable endeavour.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	33 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

```

PY ■ DATA LOADER csv_loader
else:
    print("Request to /clean failed with status code:", res
    print("Response content:", response.content)

with open(csv_filepath, "rb") as csv_file:
    # Prepare the POST request with the file as payload
    files = {'file': ('data.csv', csv_file)}

    # Make the POST request to the "/clean" endpoint
    response = requests.post(f"{server_url}/clean/v2", files=fi

    # Check the response
    if response.status_code == 200:
        data = response.json()
        print("Message:", data["message"])
        cleaned_df_req = data.get("cleaned_df")
        if cleaned_df_req:
            cleaned_df = pd.read_json(cleaned_df_req)
            return cleaned_df
        else:
            print("No cleaned_df data found in the response.")
    else:
        print("Request to /clean/v2 failed with status code:",
        print("Response content:", response.content)

@pytest
def test_output(output, *args) → None:
    """
    Template code for testing the output of the block.
    """
    assert output is not None, 'The output is undefined'

```

Figure 19 - One block of the processing code

For a regular user, without programming knowledge or with Dev/Ops skills, the creation or management of such pipelines that assure the data flows of the SEDIMARK platform, is an impossible task. To create a bridge between the user's requirements regarding the pipeline handling and the code behind that manages the flows, we propose a web interface that mimics the data processing pipelines in Figure 18, that allows the easy configurations of each block.

In this case the loader will be loaded and stored back in the Stellio broker provided by EGM. The following processing of the data features multiple steps, among which cleaning the null values and detecting outliers, which consists of annotating the outlying values in a newly created column (*true* if the row represents an outlier or *false* otherwise). An exhaustive and more detailed list of the data processing steps is provided in deliverable SEDIMARK_D3.3, especially in its chapter 3.

The pipeline configuration interface is designed to graphically present the mage.ai pipelines in an easy-to-interpret manner. Each block (Data Loader, Transformation, Exporting) is represented as an interactive element.

The configuration of each block is possible by a mouse click which opens the config panel. Each block will show a different form, for example the Data Transformation will present options

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	34 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

for outlier detection like thresholds, Data Loader will present options of how to select data sources (CSV, JSON, streaming services etc).

The Pipeline execution and monitoring will be possible by accessing the pipeline flow control which is located at the top of the GUI.

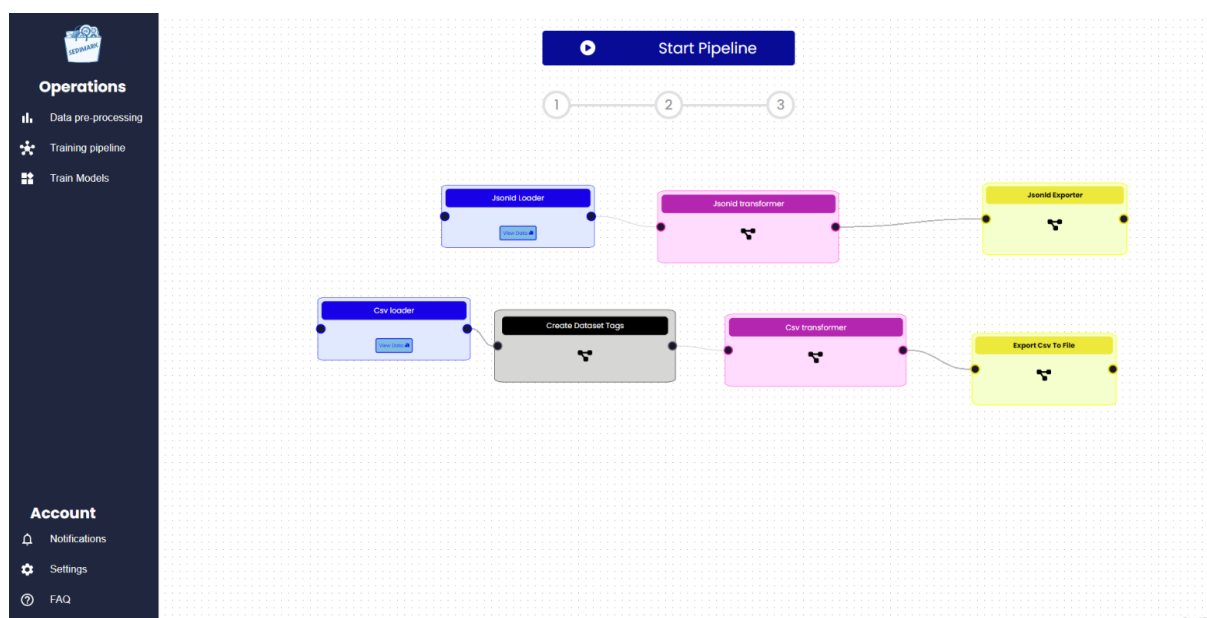


Figure 20 - Data processing pipelines UI representation

In Figure 20 we have depicted a data processing flow, the initial view. As the platform will evolve so the controls and handling of the controls will do so.

In this current version the flow is read as an already defined pipeline in mage.ai.

The user has multiple benefits from this interface:

- To see in a simplified manner the flow of its data.
- To modify the variables of the existing blocks which are then saved and used in the backend of the mage.ai.
- To run the pipeline which has been just configured, in order to see the results.

Document name:	D4.5 Data sharing platform and incentive – First version			Page:	35 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0
				Status:	Final



Figure 21 - Data processing variable selection form

Figure 21 is a sample of such configuration of variables. For example, here the user has a view of all existing pipelines in the mage.ai. It includes a search bar that allows users to filter the pipeline list based on the pipeline name. The checkboxes are for selecting multiple pipelines to perform bulk actions (like starting, stopping, or deleting), ensuring that the UI provides clear options to execute these actions.

Further in Figure 21, we will see how variables can be configured so that the behaviour/execution of the pipelines will be modelled by the values stored in this variable.

This interface has the purpose to allow the user to view the data flows created in code behind and to tweak the parameters of the processing block to achieve the best performances.

In future we will add to the interface the capabilities to:

- Create new pipelines from scratch.
- Drag and drop the blocks from a toolbox control.
- Connect the blocks seemingly.
- Map the pipeline as a new mage.ai pipeline with all the variables configured as such.
- Schedule the pipeline according to the user's wish.
- Save/Load pipeline to later reuse or modify.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	36 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

7 AI/ML dashboard

In the landscape of an evolving platform like SEDIMARK, comprehensive ML pipelines have to be designed in order to meet the user needs like extracting meaning from large datasets, or generating potential outcomes based on the historical data.

These pipelines represent a cohesive system conceived to integrated feature extraction, model training, prediction, and post-analysis into a singular workflow such as the one presented in Figure 22.

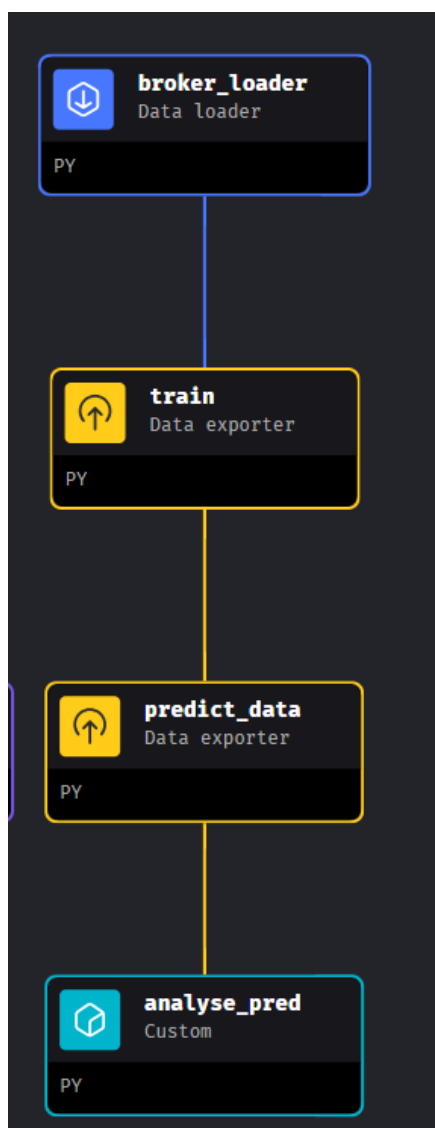


Figure 22 - The AI processing pipeline

The pipeline created to demonstrate this workflow begins with the initial setup which means defining configurations stored in an *io_config.yaml* file which in this case is to authenticate and log activities within MLFlow framework. This file contains critical information like MLFlow credentials, SQW access keys or endpoint URLs.

Then, the features which will fuel the training module will be extracted from the data, and a model is “fit” upon that data. In this case a Regression model is employed, but generally any

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	37 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

custom ML model can be created, according to the specific requirements of the use-case. This process is logged and monitored by MLFlow ensuring that every experimental run is thus recorded for future reference and reproducibility.

Post-training, the *export_train* or *predict_data*, functionalities, will load the trained model using the MLFlow system. Considering that this is a regression model, in this case it will use the model to predict future data points: it extrapolates the temporal sequence to generate values for the forthcoming days preparing the dataset for upcoming inferences. Considering that these are temperature values coming from the Stellio broker provided by EGM.

```

PY ■ DATA EXPORTER train
mlflow.set_tracking_uri("http://62.72.21.79:5000")

mlflow.sklearn.autolog()

if 'data_exporter' not in globals():
    from mage_ai.data_preparation.decorators import data_exporter

def train_linear_regression(data,model_name="temperature_linear_reg

    data['UnixTime'] = data['Time'].astype(int) // 10**9 # Convert

    # X and Y features
    X = data[['UnixTime']].values

    y = data[column_name].values

    # Split the data into training (90%) and test (10%)
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_

    # train linear regression model
    model = LinearRegression()
    with mlflow.start_run(experiment_id=mlflow.get_experiment_by_na
        model.fit(X_train, y_train)
    return model_name, X_test,y_test,run

@data_exporter
def export_train(data, *args, **kwargs) → None:

    model_name=kwargs.get('model_name')
    print(f"model name is {model_name}")

    column_name=kwargs.get('column_name')
    print(f"column name is {column_name}")
    time_column=kwargs.get('time_column')

    if model_name is None:
        model_name="temperature_test"

```

Figure 23 - The train block code

The final block of this AI orchestrated pipeline is the *analyze_prediction* function. This module will take over the predictions, transforming them in visual narratives via the matplotlib library. It extends the test dataset to use the model's foresight and illustrate a graphical representation of actual versus predicted values, offering a window into model's performance, which permits the user to validate the trustworthiness of the chosen model.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	38 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

Such pipelines can be of course customized in order to meet the user's needs. The customization consists in the data sources the user chooses the type of inference he/she will use, the model's parameters, the datasets the validation will be applied to, or the type of visualizations picked to gain insights about the model's outcome or its performance.

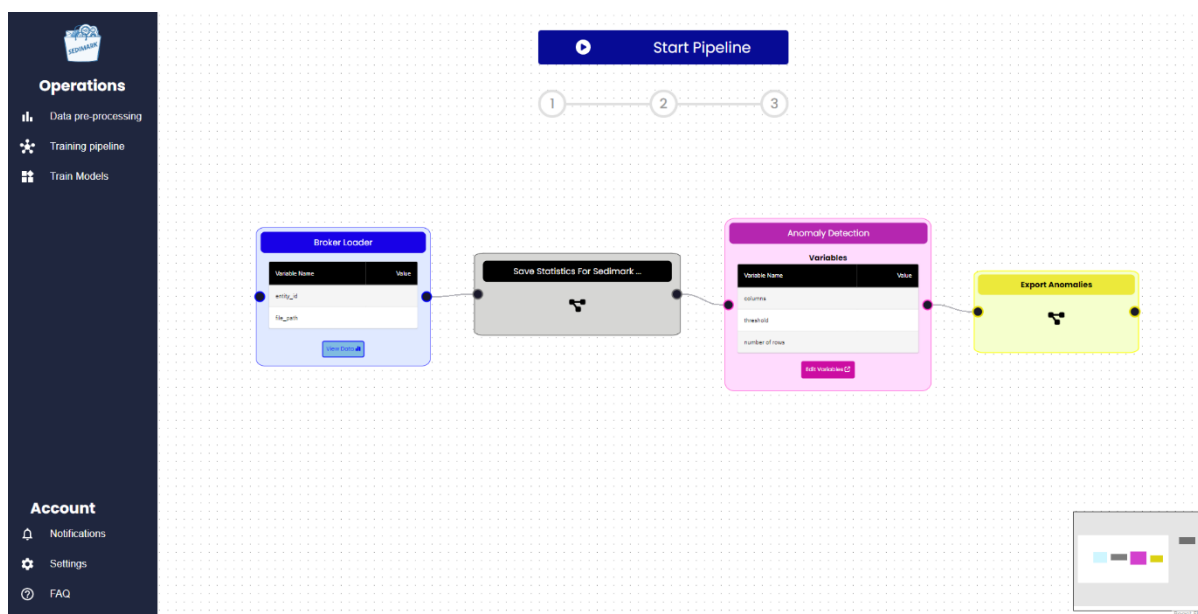


Figure 24 - The UI current implementation of the AI pipeline

In Figure 24, we present an AI orchestration the works in the context of SEDIMARK. We will briefly describe each block in what follows.

The Broker Loader block retrieves the necessary variables and data points (e.g., temperature readings or other sensor data) for further processing.

Following the data retrieval, statistical analyses or summary statistics are computed and saved. This can include measures like mean, median, or standard deviation, which help to understand the data distribution and characteristics before further processing.

This Anomaly Detection block identifies data points that deviate significantly from the norm, which can be indicative of errors, outliers, or novel insights. Variables such as columns to be analysed, the threshold for defining an anomaly, and the number of rows (data points) to predict are configured here.

Once anomalies are detected, they are exported for further analysis or for corrective actions.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	39 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

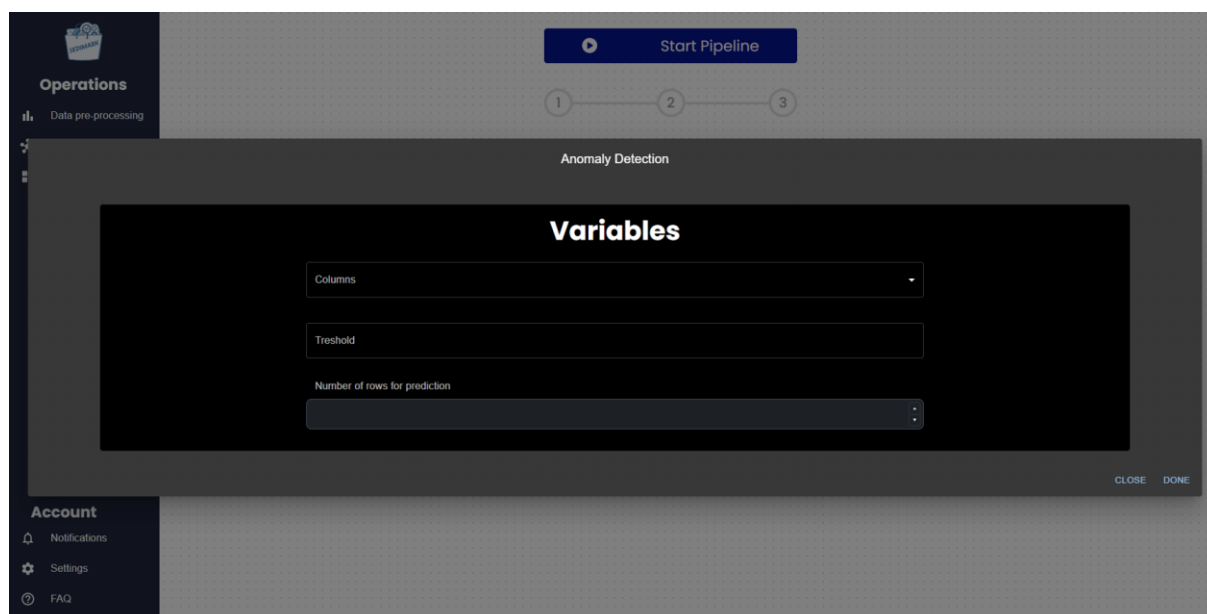


Figure 25 - The configuration parameters for the Anomaly detection algorithm

Figure 25 presents a Form used to configure variables of various blocks. These variables can be AI model's parameters, or filtering of columns that are fitted into the model, or simply any data processing variable like thresholds or properties that are used within the mage.ai pipeline blocks.

In the SEDIMARK platform, the data processing and the AI orchestration pipelines depicted in Figure 22 and Figure 24 serve as a dynamic tool that streamlines the process from data acquisition to predictive analytics. This allows users to harness machine learning capabilities to uncover deep insights and forecast trends with precision.

The strengths of such UI interactions are:

- flexibility in the sense that the users can customize the pipelines specific to their use-case.
- integration with tools like MLFlow or anomaly detection libraries, without having the user to go in the depths of configuring such tools.
- the ability to transform predictions or classifications or other ML methods to visual narratives that could build user's confidence in the predictive outcomes.
- The design prioritizes the user's needs, offering a balance between automated processes and customizable options.

Document name:	D4.5 Data sharing platform and incentive – First version			Page:	40 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0
				Status:	Final

8 Recommender

8.1 Overview

SEDIMARK aims to provide personalisation services to users of the marketplace, so that they are able to easily and quickly find assets to purchase, regardless of the nature of these assets, namely regardless of if these are data, AI models or services. Personalisation comes into place with providing results that are tailored to the user needs and preferences, so that it is more likely that they will have a better experience and they will be more satisfied with their interactions with the system. Within SEDIMARK, the personalised services are mostly related with providing recommendations to participants (mostly to consumers, but without neglecting providers in some scenarios that will be described below) when they look for an Asset to purchase. In this chapter we describe the first draft of the implementation of a recommender system within SEDIMARK, discussing briefly the existing literature, how recommendations are being provided in the first implementation and what will be the steps for the future versions of the platform.

8.2 Overview of Recommender Systems

Broadly speaking, the goal of a Recommender System (RS) is to suggest “relevant” or “good” items to the user. To design a recommender system, one needs to address the following three main points [5]:

- what do we define as “items” in the RS, for example, the RS design for Netflix users would define movies as items, Spotify’s items are songs or artists.
- how do we define “good” or “relevant” recommendation, and this is closely related to the next point.
- how do we evaluate the performance of the RS

To better understand all stakeholders involved in the RS, consider Figure 26 (adapted from [5]).

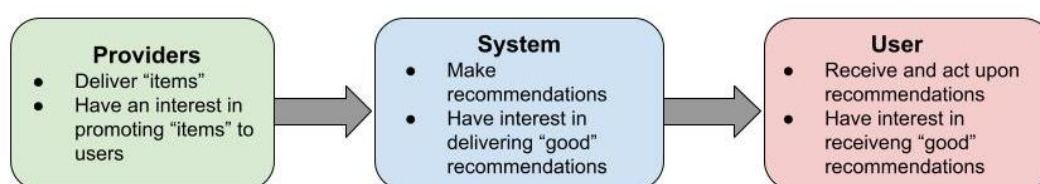


Figure 26 - Stakeholders in RS (adapted from [5])

In SEDIMARK a provider offers a dataset, ML models or services or all of them, the system is the recommendation module as part of the SEDIMARK toolbox and the user is the consumer using the Offerings.

RS can be divided into the following categories [6]:

1. **Collaborative filtering:** these methods are based on interactions between items and users as recorded by the system. For example, in Netflix, each user has a history of seen movies, therefore for these types of RS the entire history of viewed movies for each user is used as an input into the RS. The main goal of collaborative filtering is to find similar users and recommend items based on the items liked by similar users, such methods are referred to in the literature as neighbourhood-based methods. Often the

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	41 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

user-item interactions are stored in the form of a sparse matrix. Methods that fall in this category include (i) latent space methods such as matrix factorisation techniques, and (ii) deep learning techniques such as multilayered perceptrons or convolutional neural networks.

2. **Content-based:** the goal of content-based methods is to find similar items to the items that the current user likes. In this case, each item is described with a set of attributes and the goal of the content-based algorithm is to group similar items based on their attributes. Using the Netflix example above, if the user likes comedy movies, the system will find other comedy movies and recommend those to the user.
3. **Community-based:** the recommendations generated by community-based methods rely on the preferences of the user’s “friends”. Apart from user-item interactions, these methods also need social networks to describe the friendships between users.
4. **Demographic:** the assumption of the methods falling into this category is that people from different demographic areas should have different recommendations. An example could be recommending articles based on the language of the user’s country.
5. **Knowledge-based:** these methods rely on domain knowledge. One example of such a method is the case-based approach [7], where the problem description defines the user’s needs and the solutions are the recommendations.
6. **Hybrid:** these methods rely on a combination of two or more techniques described above. The goal is to use multiple methods to overcome problems faced by each particular method. For example, a hybrid method combines collaborative filtering and a community-based method and uses the community-based approach to overcome the “cold start” problem, which arises when a new user joins the system and has not consumed any particular item yet.

Overall, the goal of RS is to rank the items based on the user’s preference, with the items appearing at the top of the list being the most “relevant” items to the current user. The current version of the SEDIMARK recommendation module includes several content-based methods. These are described in more detail next.

8.3 Design of the recommender module

8.3.1 Internal Structure of the Recommender module

The Recommender module is a functional component that is part of the “SEDIMARK specific components” of the Marketplace Enabler as discussed in deliverable SEDIMARK_D2.2 [1]. The goal is to provide a complementary service to the users of the SEDIMARK platform so that they can receive personalised results when performing queries for discovering Offerings and Assets through the Marketplace. The internal structure of the Recommender module and the interfaces and interactions with external components are depicted in the figure below:

Document name:	D4.5 Data sharing platform and incentive – First version			Page:	42 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0
				Status:	Final

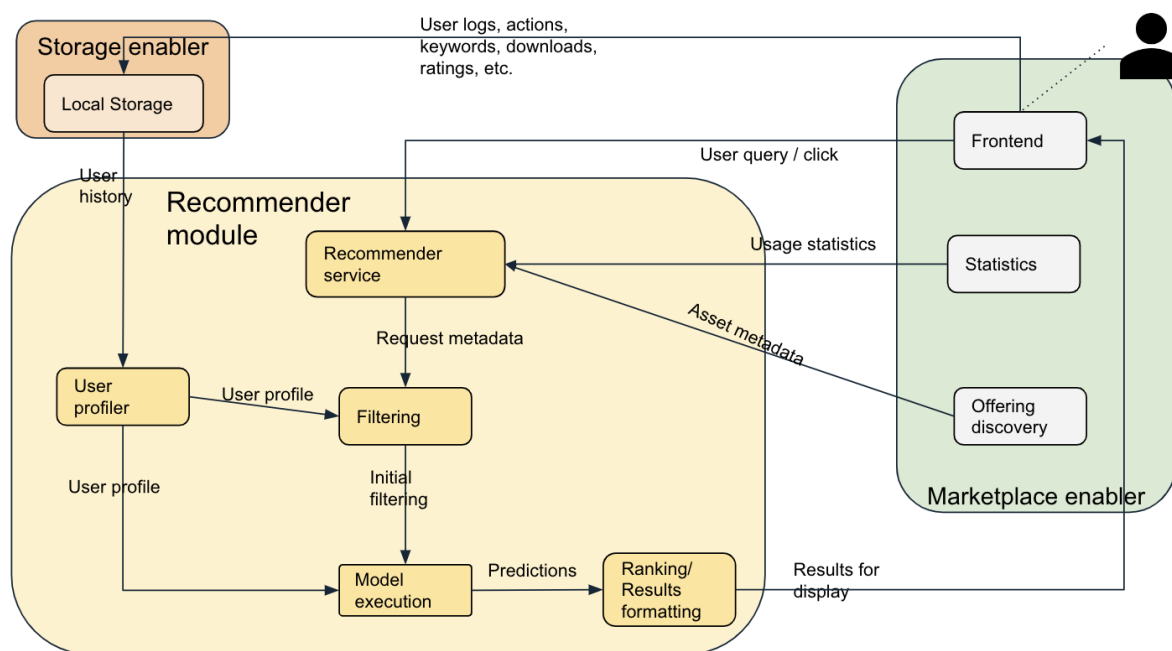


Figure 27 - Internal structure of the recommender module and its interactions with external modules.

The Recommender module comprises five sub-modules, each of which is responsible for a separate task of the recommendation service. More detailed information is given below:

- **Recommender service:** this component plays the role of the wrapper service that manages the recommendation functionalities. It is responsible for interacting with the external components and especially the Frontend, in order to (i) receive the recommendation query, (ii) analyse it, (iii) extract the required information, (iv) convert the query into a specific format to be used internally in the service and then (v) start the process for executing the recommendation pipeline to get the final results and return them to the Frontend to be displayed to the end user.
- **User profile:** this component is the main component enabling the personalisation of the recommendations, computing the user profiles based on user's past interactions with the Marketplace and user demographics.
- **Filtering:** this is the component that does an initial filtering of the items/assets to be recommended to the user, so that a reduced list of candidate items will be the input to the Recommender model, aiming to reduce the complexity of the computations and speed up the model execution.
- **Model execution:** this is the actual component that uses input about the user profile and the candidate items, as well as external information (statistics) and executes the trained recommendation model to make predictions about how well the candidate items match the user profile.
- **Ranking/Results formatting:** this component handles the final step of the recommendation module that takes the results of the recommender model and ranks the candidate items (based on several criteria), creating the final ordered list of items to be recommended to the user in the correct format to be used by the Frontend to display the list to the user.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	43 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

As it can be seen in Figure 27, the Recommender module interacts mainly with components of the Marketplace enabler and of the Storage enabler. Below is a description of the interactions and the information that is exchanged between the respective components:

- **Recommender Service ← Frontend**: SEDIMARK assumes that the user of the platform (either a provider or a consumer) will navigate the Frontend, discovering new offerings and assets for purchase. Any **queries** that the user makes to the system for asset/offering discovery will be forwarded to the Recommender Service, so that it uses the context of the query in order to provide personalised recommendations to the user. Additionally, considering that recommendations are also provided when users click on interesting items on the Frontend or when they purchase items, any user “**clicks**” on the Frontend are also being forwarded to the Recommendation Service for launching a new recommendation process.
- **Recommender Service ← Statistics**: The Recommender systems normally use extra contextual information about recent trends and popularity of items for ranking of items in order to help users explore more the available lists of items. To do so, the Recommender Service needs to periodically get information from the Statistics module regarding the usage of the items (i.e. how many likes/dislikes, how many times downloaded or clicked) and the recency of the actions on the items (i.e. how many clicks/downloads the last day, week, month).
- **User Profiler ← Local storage**: One important task of the Recommender Service is to be able to identify the user preferences over the items, so that it can recommend items that have higher chance of getting a positive reaction by the user. In order to do so, the User Profiler module needs to get demographic information about the user (i.e. gender, age, occupation, location) and the history of the user interactions with the Marketplace, i.e. past clicks, purchases, likes, etc. The Local Storage is assumed to keep the logs of these interactions of the user with the Frontend in a way that the User Profiler can easily look for and get access to.
- **Recommender Service ← Offering Discovery**: The Recommender Service needs to know the full catalogue of available items and contextual information about the items so that it can identify the most preferable for the user. In this respect, the Recommender Service queries the Offering Discovery module and receives information about the Assets that are available from the providers, as well as any available information that is extracted from the Offering descriptions for these Assets.
- **Recommender Service → Filtering**: The Filtering module gets the query information and the Asset list from the Recommender Service in order to perform a first filtering of the items to reduce the candidate list.
- **Filtering ← User Profile**: The Filtering module also receives the user profile information to help take the filtering decisions, to avoid filter out preferable items or to filter out items that the user has disliked or purchased in the past.
- **Model Execution ← User Profile**: The Model Execution module needs to receive all the information about the user profile to use it as input for the model to be executed.
- **Model Execution ← Filtering**: The Model Execution module needs to get from the Filtering module the (filtered) candidate list of items to be recommended, along with their contextual information.
- **Ranking/Results Formatting ← Model Execution**: After the Model Execution module runs the inference of the model it outputs predictions about candidate items which are forwarded to the Ranking module in order to produce the final ranked list of items in the correct format to be sent to the Frontend. The Ranking module can also include methods

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	44 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

for adding diversity or novelty to the ranked list, to improve the quality of recommendations.

- **Ranking/Results Formatting** → **Frontend**: After the final ranked list of recommendations is computed it is then converted to the proper format and sent to the Frontend to be displayed to the user.

8.3.2 Recommendation data flow

The data flow for the recommendation service was presented in SEDIMARK_D2.2 in Section 7.8 [1]. Here in Figure 28, we present an improved version of the data flow for the sake of completeness. When a user interacts with the Frontend of the marketplace, the Recommendation module receives a recommendation request, which includes the metadata of the request (i.e. the search query, the item clicked, etc.). Then, the Recommendation module sends the information to the User Profiling to get the profile and preferences of the user. The User Profile computes this information periodically, getting the historical data and the logs from the Logging module. The Recommendation module also interacts with the Offering Discovery and the Statistics module to get the candidate list of items to be recommended, the metadata of the items and the usage statistics. Then, having all the required information, the Recommendation module filters out the unneeded candidate items, runs the Recommender model (using the Model Inference module), ranks the predictions and formats the results in the proper way to be sent to the Frontend for displaying them to the user.

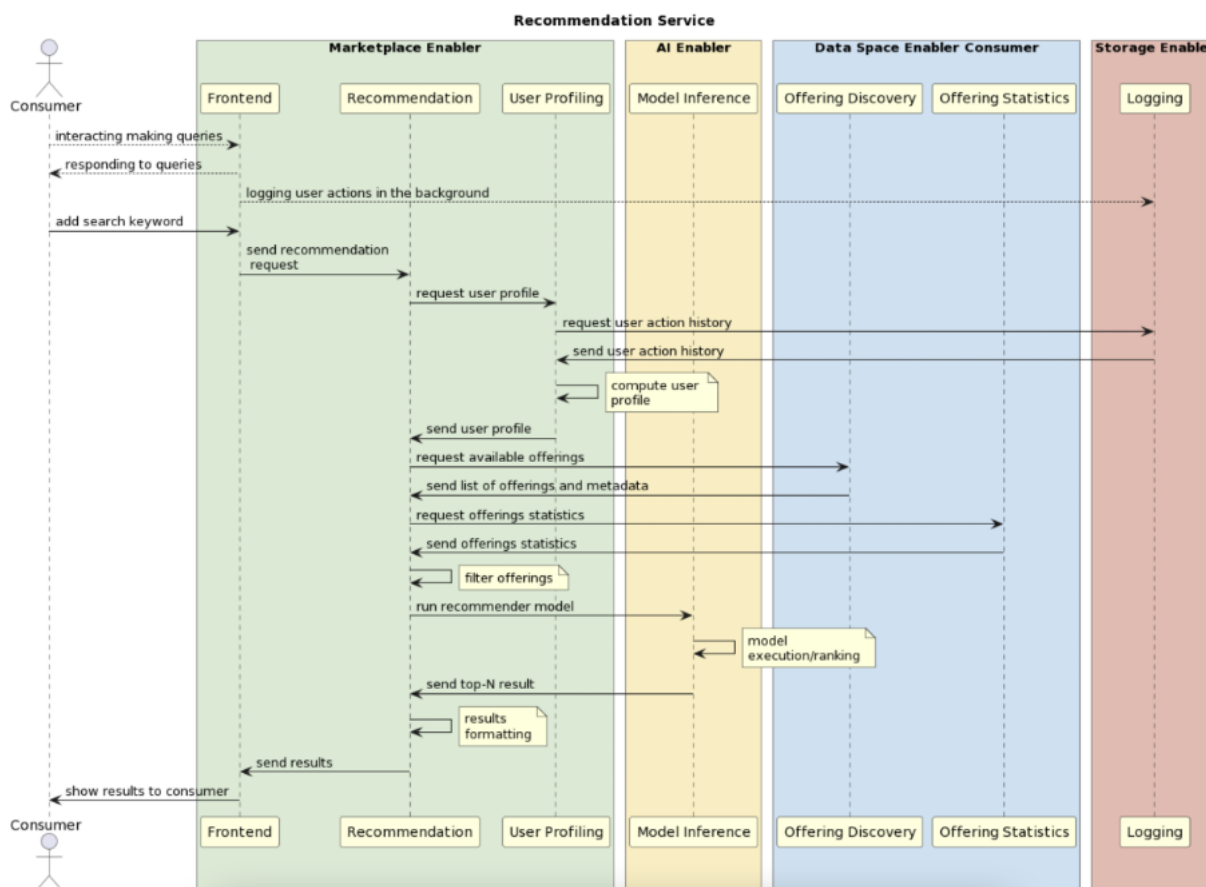


Figure 28 - Recommendation service data flow example

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	45 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

8.3.3 User and item profiling

One key requirement for building recommender systems is to have representative features to describe the users and the items that are going as input to the Recommender model. Significant research has been devoted into extracted features from datasets. Within SEDIMARK, the decentralised nature of the Recommender System assumes that all the user information will be stored locally on the user's Local Storage, and this will be used in order to extract demographic information about the user, as well as historical data with regard to the user interaction with the Marketplace.

The user profile information is split into two parts:

- **Demographic information:** This can be either added directly by the user to the system or can be gathered by the system via a questionnaire when the user first logs into the system:
 - **Age/gender** is normally used as part of the user profile in Recommender systems, however this information seems mostly irrelevant for recommending datasets and models, but might be useful when recommending services.
 - **Occupation** is considered an important factor, since depending on the domain of occupation related assets can be recommended.
 - **Domains of interest** is also an interesting feature for this type of recommendations. For example, researchers can select communications or biology as domains of interest to get more recommendations for Assets from these domains.
 - User **location** can also be used to recommend services and datasets that are in areas close to the user. For example, in case of weather monitoring, a user might be more interested into getting a weather dataset from their city or country compared to a city on the other side of the world.
- **Historical data:** this is captured by the Frontend and stored on the Local storage, from where the Recommender module gets the information. Within SEDIMARK, historical data for the user can be the following:
 - User **activity** on the Frontend, i.e. clicks, purchases, etc.
 - Previous user **search queries** to identify trends in what the user is looking for and compute their preferences through that information.
 - User **likes/dislikes** on specific Assets of the Marketplace, which will help the Recommender system have direct information about the interests and the preferences of the user.
 - User **reaction** to recommendations, by clicking on a recommended item or disliking the recommendation. This will help the recommender system improve its results.

The item profile information is used to characterise the assets that are candidates for recommendation by the recommender system. This information will be extracted from the Offering descriptions that the providers will define and will be received by the Recommender system as part of the interaction with the Offering Discovery module. The example description of the Assets is inspired by the DescribeML language, which is a tool used to describe ML datasets [8]. Considering that in SEDIMARK there are three main types of assets, different item profile information is aimed to be used by the system:

- **Dataset profile information:** the information regarding datasets that might be used as item features by the Recommender system is the following:

Document name:	D4.5 Data sharing platform and incentive – First version			Page:	46 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0
				Status:	Final

- **Domain of interest**, i.e. water, energy, health, transport, environment, etc.
- **Type** of dataset, i.e. streaming or fixed dataset.
- Dataset **category**, i.e. measurements, tabular, user data, etc.
- Dataset **features**, i.e. what are the column names in the dataset table.
- **Location**, in a predefined format.
- **Purpose**, defining what was the purpose for gathering and sharing the dataset.
- **Target** usage, i.e. classification, regression, recommendations, etc.
- **Statistics** regarding the usage of the dataset.
- **Size**, in a predefined format.
- **Flags**, i.e. if the dataset is being available as part of distributed model training within SEDIMARK.
- Extra **keywords** that can be processed and converted to features.
- **AI models profile information**: for the profile of AI models in addition to most of the above-mentioned dataset profile information we assume that the following information will be useful for the Recommender:
 - **Dataset** used for training the model.
 - **Domain of interest**.
 - **Target activity**, i.e. classification, regression, rating prediction, etc.
 - **Type** of model, i.e. simple, DNN, LSTM, etc.
 - Model **description**, in the predefined format of SEDIMARK.
 - **Training framework** used, i.e. Tensorflow, Keras, PyTorch.
- **Service profile information**: it is assumed that the service profile information required will be similar to the dataset information.

8.4 Implementation

8.4.1 Overview

Within SEDIMARK, we define two main scenarios for providing recommendations to consumers to use the Frontend of the Marketplace:

- query-based recommendations: this takes place when the user is performing a query aiming to discover some offering or asset.
- item-based recommendations: this takes place when a user interacts with an item (click, purchase, rate, etc.)

More details on the models that are implemented for each of the scenarios are given below.

8.4.2 Query based recommendation

In the query-based recommendation scenario the goal is to provide to users recommendations that are related to the discovery query that the user makes for finding interesting Assets and Offerings. To do so, the process is split in two parts:

- identifying items that are related with the user query.
- personalise the list of candidate items as defined by the previous step.

Document name:	D4.5 Data sharing platform and incentive – First version			Page:	47 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0
				Status:	Final

For the first step, the main idea is to identify the related *features* in the dataset of Assets and then process these features in the appropriate way to use it in the Recommendation model. Considering the item features described above, there can be two three main types of features:

- numerical, which are inputted as they are in the model.
- categorical, which are usually one hot encoded before being used by the model.
- plain text, which are usually processed using mechanisms from natural language processing (NLP).

In the current implementation, the main goal is to identify “*keyword*” features that can be extracted both by the metadata of the assets and the user query, and use a model to “match” the keyword features, in order to find the assets that match best to the specific query of the user. Then, the top-N assets that have a higher similarity score with the user query are the ones that are presented to the user as recommendations (or are the candidate ones for the “personalisation” step). In this first implementation, there is no “personalisation” step, since there are yet no user data available.

[BERT-based implementation](#)

The first step of the process is to use the Assets’ Self-Descriptions in order to extract keywords using a language model. The options currently are to use:

- “KeyBERT” [9], which is a language model specifically made for keyword extraction, based on the Bidirectional Encoder Representations from Transformers (BERT) family of language models developed by Google [10]. The benefit of using KeyBERT is that it is an easy to use and minimal keyword extraction technique to extract keyphrases.
- sciBERT [11], which is a BERT model trained on scientific text and might be more useful in terms of extracting keywords from datasets and models in the SEDIMARK scenario.

In our example dataset, the metadata that are used as input to the keyword extractor models are the dataset description, additional relevant information added by the dataset owner and information about the fields of the datasets. This process will create a list of keywords for each one of the assets.

The next step of the process is to extract the keywords from the user query using the same exact BERT model as for the asset keywords.

Then, what follows is to create a similarity matching between the keywords of the Assets and the keywords of the query. This is currently done using “*sentence transformers*” [12], either directly or through the “*sentence similarity*” [13] Python library. The goal of using the sentence transformers is to have a more meaningful way to compute the similarity of the keywords compared to standard methods that compare strings like Hammington distance [14] or Levenshtein distance [15], which do not take into account the semantic similarity of the strings. Sentence transformers are based on BERT and are used to encode the keywords into word embeddings so that they can be easily compared.

For comparison of the keyword embeddings, we use the cosine similarity as the metric and use the scores in order to rank the assets and produce the top-N ranked list to display to the user.

[LSI-based implementation](#)

The second implementation of the process uses Latent Semantic Indexing (LSI) [16] as implemented by the gensim Python library [17]. LSI uses singular value decomposition (SVD)

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	48 of 60				
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0	Status:	Final

to identify patterns in relationships between terms and concepts in texts. In our implementation, we use LSI to create a model of the abstracts of the assets, to use it to extract text embeddings. After building that model, we extract the embeddings for the user query, using the LSI model we built before. Then, we compute the similarity between the two sets of embeddings to create the top-N rank list. This method using LSI doesn't require or using a language model and can be much faster and more lightweight, without needing to download a BERT model compared to the previous described method.

8.4.3 Item-based recommendation

In this scenario, the goal is to exploit the information about user actions on the Frontend i.e. to provide similar items to the ones that the user either purchased or clicked. This functionality resembles the "items similar to your purchase" recommendations of online websites.

In our current implementation, the first step is similar to the query-based recommendation, to extract the keywords for the items (assets) to be recommended. This can be done in any of the ways described above, i.e. using KeyBERT. The next step is to create term frequency - inverse document frequency (tf-idf) embeddings of the keywords [18], so that they are easy to compare. Tf-idf is widely used in information retrieval to measure the importance of a word in a document. After that, the cosine similarity matrix of the items of the dataset is computed, by calculating the cosine similarity between the embeddings of pairs of items.

Then, the recommendation list for similar items to a target item is computed by sorting the column of the similarity matrix that corresponds to the index of the target item.

8.5 Results

For the testing of the recommendation models, since currently there is no related data within the project, we used the datasets from [19]. This includes a list of research datasets with a number of accompanying features, i.e. description, related papers, modalities, tasks, number of papers using it, keywords, etc. For showcasing the recommendations, an initial simple user interface on Jupyter Notebook was developed, with a simple input box, where the user can input their query for an Asset. There is no limit on the length of the input or how the input should be formatted. The processing part of the recommendation system will convert the query to embeddings with one of the methods discussed in the previous sections.

SEDIMARK DEMO

Query for asset:

As also discussed above, two types of recommendations are provided as output of the query. First, there is a list of recommendations regarding datasets that best fit the query of the user. Also, for each recommended dataset, a list of "related datasets" is also presented. This was developed to emulate the action when a user "clicks" or "purchases" a dataset.

Document name:	D4.5 Data sharing platform and incentive – First version			Page:	49 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0
				Status:	Final

8.5.1 Query based recommendation examples

Below we provide some results for example queries:

Query: “reviews” → looking for dataset that include user reviews

Results:

Acronym	Keywords
Coffereview Dataset	coffee reviews review bean grading
Yelp Review Polarity	polarity yelp positive reviews review
Cookie	amazon conversational dataset recommendation agent
Amazon Product Data	reviews amazon dataset ratings metadata
BeerAdvocate	beeradvocate beer ratings reviews review
Amazon Fine Foods	finefoods reviews foods amazon ratings
Amazon Review	amazon reviews review sentiment dvds
Casino Reviews	casino dataset reviews google sentiment
Commonsense-Dialogues	commonsense dialogues crowdsourced social contexts
Multi-Domain Sentiment Dataset v2.0	ratings sentiment reviews amazon domains

Query: “speech recognition”

Recommendations:

Acronym	Keywords
VoxForge	voxforge transcribed speech dataset recognition
MAVS	audio smartphone smartphones speaker recognition
CI-AVSR	cantonese audio avsr dataset recognition
CN-Celeb-AV	celeb dataset audio cn recognition
Fontenay Dataset	transcribed handwritten texts fontenay recognition
MIntRec	intent multimodal mintrec recognition audio
ARVSU	utterances addressee arvsu recognition utterance
ADVANCE	multimodal aerial recognition dataset audio
KOHTD	handwritten handwriting recognition papers dataset
MSI	eeg emotion hierarchical recognition induction

Query: “audio processing”

Recommendations:

Acronym	Keywords
VoxForge	voxforge transcribed speech dataset recognition
MAVS	audio smartphone smartphones speaker recognition
CI-AVSR	cantonese audio avsr dataset recognition
CN-Celeb-AV	celeb dataset audio cn recognition
Fontenay Dataset	transcribed handwritten texts fontenay recognition
MIntRec	intent multimodal mintrec recognition audio
ARVSU	utterances addressee arvsu recognition utterance
ADVANCE	multimodal aerial recognition dataset audio
KOHTD	handwritten handwriting recognition papers dataset
MSI	eeg emotion hierarchical recognition induction

8.5.2 Item based recommendation examples

Regarding item-based recommendations, we recommend 5 similar datasets to the one the user has selected. Example results are shown below:

Dataset: “Coffereview Dataset”

Recommendations:

Query for asset:

Acronym	Keywords
DBRD	dbrd sentiment reviews dataset review
BeerAdvocate	beeradvocate beer ratings reviews review
NSMC	korean reviews nsmc review naver
AmazonQA	amazonqa amazon reviews questions review
Amazon Review	amazon reviews review sentiment dvds

Dataset: “CIFAR-10”

Recommendations:

Query for asset:

Acronym	Keywords
20Newsgroup (10 tasks)	dataset classes pycontinual classification class
VOC-MLT	voc voc2007 classes tailed class
CIFAR-100	classes images class label labelers
Tobacco-3482	dataset images tobacco classes 3482
Stickers	stickers sticker images image alpha

8.6 Future work

As specified in previous sections, over time SEDIMARK will collect historical data for each user such as for example the user’s past purchases, preferences for certain items or services, and so on. As the system builds enough data, we aim to include more sophisticated RS models based on collaborative filtering as specified in Section 5.2. As can be seen in the previous section, current results are encouraging, but there is still quite some room until they are “production-ready”. To improve the quality of our recommendations even further we aim to use hybrid techniques, specifically in cases of new users joining the system and to avoid the cold start problem.

Collaborative filtering techniques rely on user-item interactions and one major concern with such data is user privacy. To this end, we aim to develop decentralised approaches, where users do not need to share their raw private data with other users or global servers. But instead each user builds the model privately only sharing the model updates. Although research has shown that such models could be reverse-engineered or are susceptible to a wide range of attacks [20] [21]. We aim to address such issues as follows:

- User privacy: collaborative filtering models generally consist of two types of parameters, one related to users and the other related to items. We aim to keep the user’s model parameters private and share the parameters of the items during the model training process [22]. To improve the user privacy further we also aim to add differential privacy to the shared parameters [23].
- Back door poisoning attacks: FL systems are susceptible to back door attacks [20] [21], where a malicious user can alternate the final model in order to get some gain. For example, in the case of a recommender system the aim of a malicious user is to promote certain item to increase their revenue. We aim to employ techniques in order to prevent such attacks such as for example [24], where the authors propose a detection method consisting of four main parts: (i) reverse engineering, (ii) global reverse trigger generation, (iii) outlier detection and (iv) model repair.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	52 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

9 Open Data enabler

The Open Data enabler is a component of the SEDIMARK ecosystem whose goal is to promote data sharing in the Marketplace by providing free and open datasets or services. A high level description can be found in deliverable SEDIMARK_D2.2 section 6.12 [1]. This chapter focus on the architecture of this component and the implementation of its first version, as well as its expected evolution across and beyond the lifetime of the project.

9.1 Architecture

Because it aims at populating the Offering Catalogue with datasets or services any Participant can access for free, the Open Data enabler itself is actually a Participant in the SEDIMARK ecosystem, acting solely as a Provider. As such, the Open Data enabler do not need any components to be installed on the Participants premises to work: it will be hosted on Atos premises. It can therefore be continuously improved during the lifetime of SEDIMARK, for instance by adding Offerings and updating them, with no impact on its Participants: the changes will only be reflected in the Catalogue. Besides, having a Participant dedicated to the development of one of SEDIMARK's component is an advantage for its consortium, offering an additional platform to test them.

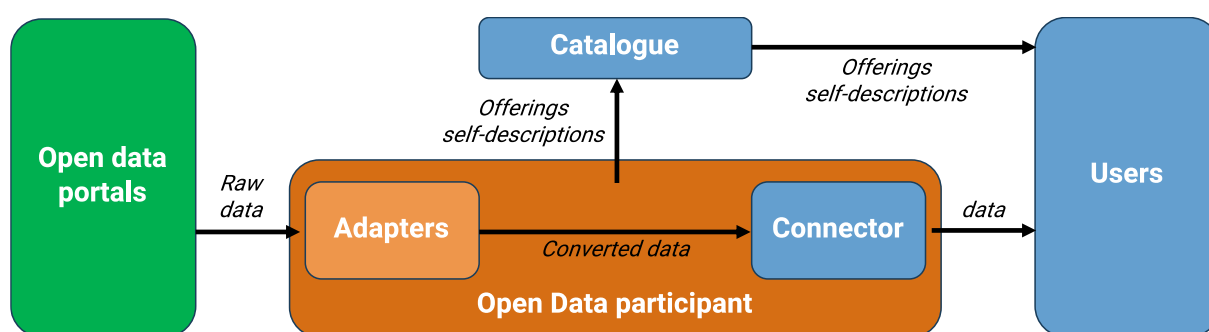


Figure 29 - Open Data enabler architecture

Figure 29 depicts the working principle of the Open Data enabler. As a Participant, the Open Data enabler is equipped only with the SEDIMARK essential components, especially to interact with the federated Catalogue, in order to post Offerings corresponding to the open datasets or portal it exposes, as well as a connector to ensure data transfers to the consumers' premises. The Open Data enabler only needs a minimal subset of SEDIMARK components: AI and data processing toolboxes, as well as the Marketplace frontend, won't be necessary for it to run.

To fetch data from open data portals, it relies on a set of *adapters*: their role is to ensure any request to the open dataset or API can be forwarded to the Connector, so Consumers can access them. Consequently, an adapter can be implemented either as an extension of the Connector (by the extension mechanisms provided by the Eclipse Data Space Connector for instance [25]) or as a separate software component exposing an endpoint for the Connector to scrape. The latter case is expected to be less likely as data spaces adoption enlarges and Connector functionalities get extended, an effort which SEDIMARK contributes to.

9.2 Kaggle data offering

For the first version of SEDIMARK, only dataset Offerings will be supported. Therefore, to release the first working version of the Open Data enabler, we will use datasets available in

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	53 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

Kaggle [26]. A python client [27] can be used to access the datasets, so the adapter will consist of a shallow REST API microservice using FastAPI [28], scraping the data from Kaggle using its client, and exposing for consumption by the connector. This approach will allow the SEDIMARK consortium to test the first version of the platform with a variety of dataset offerings. Care will be taken to select datasets whose license allow resharing in any form, as well as potential commercial use (such as CC-BY [29] or Open Data commons [30]).

In later iterations of the platform, SEDIMARK will strive to expose the full Kaggle API, as well as other data portals such as CKAN [31], via service Offerings.

Document name:	D4.5 Data sharing platform and incentive – First version			Page:	54 of 60		
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0	Status:	Final

10 Data sharing incentives

This chapter describes how SEDIMARK intends to foster data sharing in its marketplace through two approaches. On the one hand, use cases operators will act as participants in the SEDIMARK ecosystem to showcase its functionalities: section 10.1 gives such an example with the city of Helsinki. On the other hand, section 10.2 presents some initial ideas of features to be implemented in the marketplace frontend to enable users to emphasize the quality of their datasets.

10.1 City of Helsinki as a participant in the marketplace

SEDIMARK's core features will be continuously evaluated through its 4 use cases, with a special attention to include:

- Connecting remote data platforms seamlessly for enhanced collaboration and insights.
- Enabling efficient and privacy-preserving data sharing while ensuring data security and confidentiality.
- Offering diverse, high-quality, certified data and services to meet various industry needs and standards.
- Supporting the EU's Common Data Spaces initiative for fostering innovation and digital transformation in the EU.

All pilot sites will contribute to enrich the marketplace, acting as participants in the SEDIMARK ecosystem, in order to provide high quality public offerings showcasing its catalogue to prospective users.

The City of Helsinki will be one of such pioneer participants, demonstrating how the SEDIMARK marketplace can be used by stakeholders in city mobility to foster data sharing, activating engagement and collaboration, making use of Forum Virium's expertise guiding projects to co-create smart city innovations that enhance urban residents' quality of life while minimizing environmental impact. It will at first provide dataset offerings based on its mobility data platform [32]. As the SEDIMARK project goes on, it will not only expose its data APIs as service offerings, but also keep collaborating with various stakeholders to create new projects and pilots, aiming at improving the usability and usefulness of data, as well as developing tools for a mobility digital twin to enable the testing and development of new smart mobility solutions in real urban environments.

All information resources are managed in accordance with the FAIR model, which defines the principles for the fair and efficient utilization of data in an organization [33], to ensure the high quality and interoperability of the provided data.

10.2 Marketplace frontend features to foster data sharing

The first version of the marketplace will be focused on enabling dataset offerings provision and consumption, as well as the access to the SEDIMARK toolboxes for data processing and AI. The first version of the recommender system presented in chapter 8 will be a first key step in fostering data sharing, allowing users to be quickly informed of new or existing datasets matching their interests, without having to perform complex searches in the catalogue to discover them. Yet, many other features are considered to further incentivise data sharing in the second version of the marketplace.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	55 of 60				
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0	Status:	Final

At first, enabling participants to rate the offerings they consumed will make sure some datasets are rewarded for their quality and applicability. A first approach will consist of keeping this rating simple and intuitive, inspired from common online marketplaces such as Amazon or Uber, where users can rate the product or service, they bought over a five star scale.

Another incentive relies on the work done in WP3 (Distributed data quality management and interoperability), where quantitative metrics are being established to provide insights on the quality of datasets available or to be put in the SEDIMARK catalogue. Users will have the possibility to use the SEDIMARK toolbox to assess the quality of their datasets and use its various components to improve it, should need be. Upon creating the corresponding data assets, they will be able to add such metrics in the online representation of their offerings. Additionally, the marketplace will draw inspiration from other data sharing platform such as Kaggle, which provide a usability score derived from various properties of the dataset (presence of descriptions and tags, provision of a license, update frequency ...), to display other informative criterion satisfied by the offerings.

Document name:	D4.5 Data sharing platform and incentive – First version			Page:	56 of 60		
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0	Status:	Final



11 Conclusions

This report recaps the initial approach to the architecture of the SEDIMARK Marketplace and data sharing platform and hence describes its main objective, as well as introduces its constitutive modules and anticipates their functionalities within the overall solution.

SEDIMARK looks to build a decentralised marketplace as one of its major goals, to encourage users from diverse backgrounds to exploit and benefit from its various functionalities. In such decentralised data and services sharing platform different techniques will help to clean, protect, discover and even enrich the offerings, thus paving the way for diverse businesses and research scenarios to make use of them and get in return a significant profit.

To achieve such objective, the current deliverable presented the modules about to be developed to shape the SEDIMARK Marketplace. The exercise starts with a recap of the basic actions that will be covered by such platform as well as with the presentation of a brief analysis on the users expected to appear in there.

How those users will be welcomed into the system is a matter to be dealt with from the onboarding and authentication side. Hence, the corresponding description of the home page of the marketplace and the process for participants to log in their accounts ensues. This includes the step-by-step guide to explain how new uses can perform their registration and be welcomed into SEDIMARK.

Up next, the presentation continues with the introduction into the Catalogue of Offerings that will be presented in the Marketplace and where any Participant, both registered and plain visitors, will have the chance to browse and find the Offerings that better suit their needs. The actions performed by Participants in the Catalogue will have an impact on the Recommendation system, as hinted in the text. This Recommender deserves a section of its own and thus the report delivers a detailed explanation on how it works, which are its main design guidelines, and how data flows within and from it. Eventually, a depiction of how SEDIMARK implements the Recommender appears, alongside a summary of the initial results obtained in the test processes conducted to validate the Recommendation models.

To understand the aforementioned Offerings, the report includes a complete go through what they are and how they may be registered and then published in the Catalogue. This includes the option to describe them as well as fixing their pricing and policies. In order to perform these operations, authenticated users will have the chance to access a specific dashboard that helps to manage their Offerings, both provided and consumed. The overview of the tabs this tool comprises contributes to obtain a holistic view of the type of actions to perform over the assets and policies, as well as get a glimpse of the contracts in place.

Willing to submit the most complete experience possible, the Marketplace also sets up secondary dashboards that will lead to data processing orchestration and to build up AI/ML pipelines for model training, both federated and distributed.

SEDIMARK's vision also involves dealing with diverse data sources. To make it a reality the Open Data enabler plays a relevant role since it is charge of promoting data sharing in the marketplace through free and open datasets or services. This report explains its architectural description and introduces its initial iteration, where datasets coming from Kaggle feed the module.

Finally, the document reflects on an initial approach on the way SEDIMARK will offer certain incentives to promote data sharing. This is an action tightly related to the project use cases

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	57 of 60				
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0	Status:	Final



and thus the input from their respective owners results truly valuable. Then, the report presents a collection of ideas and opportunities which will in turn feed further iterations of this activity.

All in all, this document includes the description of the first version of the SEDIMARK data sharing platform. Given the project structure what readers find here is not set in stone but subjected to evolution along the course of the project. Therefore, a new version of this exercise will be presented in SEDIMARK_D4.6 a couple of months before SEDIMARK comes to an end, in July 2025, where updates on what this report presents and a final version of the SEDIMARK Marketplace will be discussed.

Document name:	D4.5 Data sharing platform and incentive – First version			Page:	58 of 60		
Reference:	SEDIMARK_D4.5	Dissemination:	PU	Version:	1.0	Status:	Final

12 References

- [1] “D2.2 SEDIMARK Architecture and Interfaces. First version,” p. 83.
- [2] “Download MetaMask: The Premier Blockchain Wallet App and Browser Extension.” Accessed: Dec. 15, 2023. [Online]. Available: <https://metamask.io/download/>
- [3] “Participant - Gaia-X Trust Framework - main version (fb420580).” Accessed: Dec. 15, 2023. [Online]. Available: <https://gaia-x.gitlab.io/policy-rules-committee/trust-framework/participant/>
- [4] “ODRL Information Model 2.2.” Accessed: Dec. 15, 2023. [Online]. Available: <https://www.w3.org/TR/odrl-model/#policy>
- [5] S. Milano, M. Taddeo, and L. Floridi, “Recommender systems and their ethical challenges,” *Ai & Society*, vol. 35, pp. 957–967, 2020.
- [6] Y. Koren, S. Rendle, and R. Bell, “Advances in Collaborative Filtering,” in *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds., New York, NY: Springer US, 2022, pp. 91–142. doi: 10.1007/978-1-0716-2197-4_3.
- [7] D. Bridge, M. H. Göker, L. McGinty, and B. Smyth, “Case-based recommender systems,” *The Knowledge Engineering Review*, vol. 20, no. 3, pp. 315–320, 2005.
- [8] J. Giner-Miguel, A. Gómez, and J. Cabot, “DescribeML: A dataset description tool for machine learning image 1,” *Science of Computer Programming*, vol. 231, p. 103030, 2024, doi: <https://doi.org/10.1016/j.scico.2023.103030>.
- [9] M. Grootendorst, “KeyBERT: Minimal keyword extraction with BERT. Zenodo; 2020.”
- [10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” 2019.
- [11] I. Beltagy, K. Lo, and A. Cohan, “SciBERT: A Pretrained Language Model for Scientific Text.” 2019.
- [12] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.” 2019.
- [13] Susheel, “Sentence Similarity.” Nov. 16, 2023. Accessed: Dec. 15, 2023. [Online]. Available: https://github.com/Susheel-1999/Sentence_Similarity
- [14] G. T. Reddy *et al.*, “An Ensemble based Machine Learning model for Diabetic Retinopathy Classification,” in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 2020, pp. 1–6. doi: 10.1109/ic-ETITE47903.2020.235.
- [15] D. K. Po, “Similarity based information retrieval using Levenshtein distance algorithm,” *Int. J. Adv. Sci. Res. Eng.*, vol. 6, no. 04, pp. 06–10, 2020.
- [16] C. H. Papadimitriou, H. Tamaki, P. Raghavan, and S. Vempala, “Latent semantic indexing: A probabilistic analysis,” in *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, 1998, pp. 159–168.
- [17] “Gensim: topic modelling for humans.” Accessed: Dec. 15, 2023. [Online]. Available: <https://radimrehurek.com/gensim/models/lsimodel.html>
- [18] J. Ramos and others, “Using tf-idf to determine word relevance in document queries,” in *Proceedings of the first instructional conference on machine learning*, Citeseer, 2003, pp. 29–48.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	59 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final

- [19] V. Viswanathan, L. Gao, T. Wu, P. Liu, and G. Neubig, “DataFinder: Scientific Dataset Recommendation from Natural Language Descriptions.” 2023.
- [20] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, “How To Backdoor Federated Learning,” in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, S. Chiappa and R. Calandra, Eds., in Proceedings of Machine Learning Research, vol. 108. PMLR, Aug. 2020, pp. 2938–2948. [Online]. Available: <https://proceedings.mlr.press/v108/bagdasaryan20a.html>
- [21] C. Xie, K. Huang, P.-Y. Chen, and B. Li, “DBA: Distributed Backdoor Attacks against Federated Learning,” in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:213447399>
- [22] E. Duriakova *et al.*, “PDMFRec: A Decentralised Matrix Factorisation with Tunable User-Centric Privacy,” in *Proceedings of the 13th ACM Conference on Recommender Systems*, in RecSys ’19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 457–461. doi: 10.1145/3298689.3347035.
- [23] K. Wei *et al.*, “Federated Learning With Differential Privacy: Algorithms and Performance Analysis,” *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020, doi: 10.1109/TIFS.2020.2988575.
- [24] C. Zhao, Y. Wen, S. Li, F. Liu, and D. Meng, “FederatedReverse: A Detection and Defense Method Against Backdoor Attacks in Federated Learning,” in *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, in IH&MMSec ’21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 51–62. doi: 10.1145/3437880.3460403.
- [25] “EDC Connector.” Eclipse Dataspace Components, Dec. 06, 2023. Accessed: Dec. 08, 2023. [Online]. Available: <https://github.com/eclipse-edc/Connector>
- [26] “Kaggle: Your Machine Learning and Data Science Community.” Accessed: Dec. 08, 2023. [Online]. Available: <https://www.kaggle.com/>
- [27] “Kaggle API.” Kaggle, Dec. 08, 2023. Accessed: Dec. 08, 2023. [Online]. Available: <https://github.com/Kaggle/kaggle-api>
- [28] “FastAPI.” Accessed: Dec. 08, 2023. [Online]. Available: <https://fastapi.tiangolo.com/>
- [29] “About CC Licenses,” Creative Commons. Accessed: Dec. 08, 2023. [Online]. Available: <https://creativecommons.org/share-your-work/cclicenses/>
- [30] “Database Contents License (DbCL) v1.0 — Open Data Commons: legal tools for open data.” Accessed: Dec. 08, 2023. [Online]. Available: <https://opendatacommons.org/licenses/dbcl/1-0/>
- [31] “CKAN - The open source data management system,” ckan.org. Accessed: Dec. 08, 2023. [Online]. Available: <http://ckan.org/>
- [32] “Data Catalog,” Mobility Lab Helsinki. Accessed: Dec. 21, 2023. [Online]. Available: <https://mobilitylab.hel.fi/data/>
- [33] M. D. Wilkinson *et al.*, “The FAIR Guiding Principles for scientific data management and stewardship,” *Scientific Data*, vol. 3, no. 1, p. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.

Document name:	D4.5 Data sharing platform and incentive – First version	Page:	60 of 60
Reference:	SEDIMARK_D4.5	Dissemination:	PU
		Version:	1.0
		Status:	Final