Ref. Ares(2025)6572201 - 12/08/2025

DATA SPACE 4 MOBILITY MODELS INTEGRATION

Project Acronym:	**23-ES-TG-DS4MOBMODELS**
Project ID :	**101176007**

Deliverable number:	**D1.3**
Deliverable name:	**Models and Data Management Plan**
Work Package number:	**1.PROJECT MANAGEMENT AND COORDINATION**
Due date :	**Month 6**
Delivery date :	**12/08/2025**

Lead Beneficiary :	**EYSA**
Main author :	**David Murias**

Co-funded by the European Union

DS4MOB MODELS

DOCUMENT INFORMATION

The DS4MOBMODELS project (Grant Agreement nº: 101176007) project is co-funded by the EU. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the EU or the CINEA. Neither the EU nor the CINEA can be held responsible for them.

REVISION HISTORY

Version	Date	Lead Beneficiary
V 1.0	12/08/2025	EYSA

REFERENCES

D2.2 Data Space Governance

D5.2 Technical specifications of the cloud infrastructure

GLOSSARY OF ACRONYMS AND ABBREVIATIONS

API	Application Programming Interface
CRISP-DM	Cross-Industry Standard Process For Data Mining
DIKW	Data, Information, Knowledge And Wisdom
DLP	Data Loss Prevention
DS4MOBMODELS	Data Space for Mobility Models Integration
EU	European Union
FAIR	Findable, Accessible, Interoperable And Reusable
GDPR	General Data Protection Regulation
GPS	Global Positioning System
GTFS	General Transit Feed Specification
HTTP	Hypertext Transfer Protocol
IDS	Intrusion Detection Systems
IRM	Information Rights Management
JSON	Javascript Object Notation
MFA	Multi-Factor Authentication
SQL	Structured Query Language
SSO	Single Sign-On
SSL	Secure Sockets Layer
TEN-T	Trans-European Transport Network
WAF	Web Application Firewalls
WP	Work Package
XSS	Cross-Site Scripting
XXE	Xml External Entities

TABLE OF CONTENT

Document information 2
Revision history 2
References 3
Glossary of acronyms and abbreviations 3
Table of content 4
Table of figures 6
Executive summary 1
Context 2
Objective of the deliverable 3
1. Methodology 4
1.1 Business problem understanding 5
1.2 Data Understanding 5
1.3 Data Preparation 5
1.4 Modelling 6
1.4.1 Public transport 6
1.4.2 Freight transport 7
1.4.3 Traffic 7
1.4.4 Transport emission evaluation 7
1.5 Evaluation 8
1.6 Deployment 8
2. Data summary 9
2.1 DIKW pyramid 9
2.2 Data types 11
2.3 Data usage 12
2.4 data utility 13
2.5 Data states 13
3 FAIR data 13
3.1 FAIRification process 14
3.2 Findable 15
3.3 Accessible 16
3.4 Interoperable 17
3.5 Reusable 18
4 Other research outputs 18

5 Allocation of resources 19
6 Data security, privacy and quality.....19
6.1 Data Security.....20
6.1.1 Cloud security.....21
6.1.2 OWASP security.....22
6.1.3 Sharepoint security.....25
6.2 Data Privacy.....26
6.2.1 Data governance and privacy benefits.....26
6.2.2 Data governance and privacy best practices.....26
6.3 Data Quality 27
6.3.1 Importance of data quality 27
6.3.2 Data quality dimensions.....28
7 Roles and responsibility.....28
8 Ethics 29
8.1 Ethic in personal data 30
9 References.....30

TABLE OF FIGURES

Figure 1: CRISP-DM Lifecycle 4

Figure 2: DIKW Pyramid 9

Figure 3: FAIRification process 14

EXECUTIVE SUMMARY

This work package (WP) oversees the overall management of models and data within the project, addressing all aspects related to the organization, administration, and governance of data and models.

This includes the development, implementation, and continuous updating of the **Data Management Plan (DMP)** ensuring compliance with relevant regulations and standards, for instance **FAIR principles (findable, accessible, interoperable and reusable)** as well as **GDPR (General Data Protection Regulation)**.

The present document aims to process and monitor the lifecycle of data and models from the **methodology** to collection and processing to sharing, preservation, and reuse and managing access rights, data security, and quality assurance.

The general objective of this WP is to guarantee that all data and model-related activities are carried out efficiently, transparently, and in alignment with project objectives. It ensures that datasets and models are made available and submitted on time, within the established budget, and with a high degree of quality, while identifying and mitigating risks associated with data and model management.

This WP also covers coordination activities related to data and model governance, including the organization of consortium meetings focused on data and models, status monitoring of data-related technical progress, tracking of achievements including data model milestones, deliverables, emerging issues and risks, and planning of next steps and targets for data and model management.

CONTEXT

The **DS4MOBMODELS** project aims to **define and create an advanced Mobility Data Space** that integrates transport, traffic, and travel data across various regions, including the Spanish state, the Nouvelle-Aquitaine-Euskadi-Navarre Euroregion, and the Central Portugal region.

The **primary goal** is to enhance the **availability**, **quality**, and **interoperability** of mobility data to address key national, cross-border, regional, and local use cases related to sustainable urban mobility, traffic management, and real-time operations within the Trans-European Transport Network (TEN-T).

The project will develop a data platform that will support the integration of various mobility models and data science tools to facilitate efficient, safe, and sustainable mobility management. The initiative will also establish robust data governance in line with European and national regulations, ensuring secure and interoperable access to the data for all the stakeholders in a non-discriminatory way.

The DS4MOBMODELS project **aligns with the broader objectives of the European Data Strategy**, the Digital Europe Programme, and national digital and mobility strategies in Spain, France and Portugal, contributing to the creation of a more integrated and resilient European mobility infrastructure.

OBJECTIVE OF THE DELIVERABLE

Within it, this document, Deliverable *D1.3 – Models and Data Management Plan* of the DS4MOBMODELS project (Grant Agreement No. 101176007), sets out a **comprehensive strategy for the management of all datasets** generated, collected, and processed throughout the project's lifecycle. It details the models, data management processes, and lifecycle stages, and defines if, how and under what terms data will be shared or made openly accessible. Measures for the curation and long-term preservation of data are described, in full alignment with the FAIR principles.

Another aspect of this strategy is the adoption of the **CRISP-DM** (Cross-Industry Standard Process for Data Mining) methodology. CRISP-DM provides a robust and structured framework for the entire data mining process, encompassing the following key phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. By integrating the CRISP-DM methodology into our data management approach, we ensure that each stage of data modelling and analysis is systematic, transparent, and reproducible, thereby enhancing the quality and reliability of our outputs.

The document further addresses the purpose of the data, adherence to FAIR principles, management of other research outputs, and considerations regarding security, quality, and ethics. It also specifies the roles and responsibilities related to the Models and Data Management Plan.

In addition, this deliverable establishes guidelines for quality assurance in the development of the project's modules, including the continuous monitoring of potential risks and the implementation of mitigating actions as needed, ensuring that the highest standards of data quality are achieved and maintained.

As the project evolves, the Models and Data Management Plan will be systematically reviewed and updated to reflect new challenges, regulatory requirements, and emerging best practices, including ongoing alignment with the CRISP-DM methodology. This approach guarantees ongoing compliance, project relevance, and the successful achievement of the project's objectives.

1. METHODOLOGY

The data management approach within the DS4MOBMODELS Data Space leverages the **CRISP-DM (Cross-Industry Standard Process for Data Mining)** methodology as a foundational framework. It's a widely adopted framework that outlines the steps involved in a data analytics or data science project.

As well, as the Data Space CRISP-DM was conceived in 1996 and became a European Union project under the ESPRIT funding initiative in 1997. The presence of the word data mining is pure history and nowadays, it is possible to interchange it for Machine Learning as well. For this reason further the present deliverable refers the Data Mining simply as data.

CRISP-DM provides a **structured and flexible process model for carrying out data-driven projects**, ensuring systematic progress from the initial understanding of business objectives through data preparation, modelling, evaluation, and eventual deployment of results. By integrating CRISP-DM into the Data Space, the project ensures that all data-related activities ranging from acquisition and curation to advanced analytics and sharing are managed in a consistent, transparent, and reproducible manner.

This methodology supports the alignment of technical processes with project goals and regulatory standards, fostering both innovation and compliance throughout the data lifecycle.

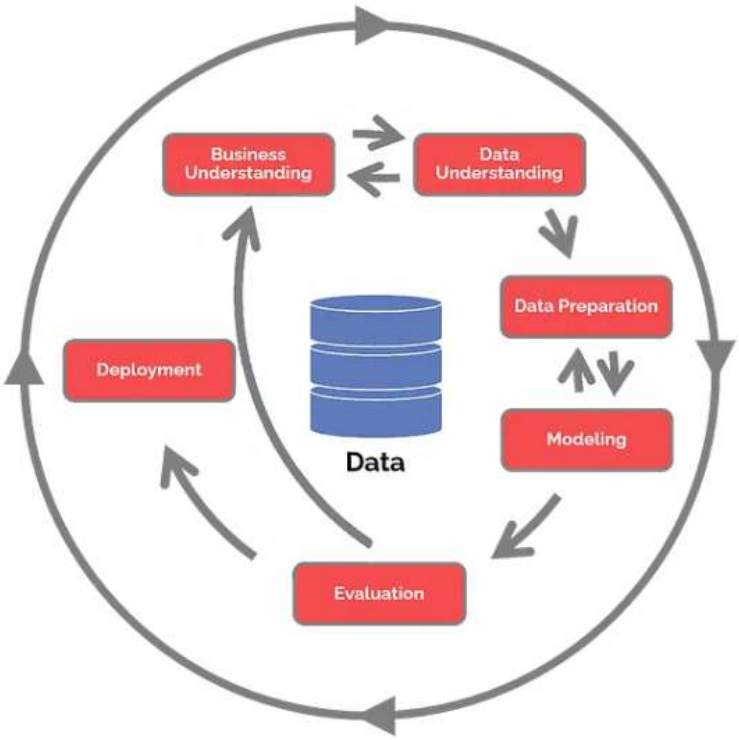


Figure 1: CRISP-DM Lifecycle

The image illustrate above visualize the CRISP-DM process model. CRISP-DM process model for data analytics provides an overview of the lifecycle of the data analytics projects. It contains the phases of a project, their respective tasks, and the relationships between them. By simply looking at the description of each task, It is not possible to identify the relationship between each task, as it depends upon the specific problem and interest of the users.

The life cycle of a data's project consists of six phases. The sequence of the phases is not rigid, moving back and forth between different stages is always required when applying the process models in real-world

projects. The outcome or result of each phase determines which phase, or particular task of a phase, has to be performed next, as each task is interdependent in many cases. For example, if in the project is required a predictive model gives less accuracy than expected, then it is possible maybe go back to analyse the data for another perspective in order to ensure data understanding and data preparation steps fit with the objectives of the model. Thereafter developers redevelop the predictive model to improve its accuracy.

The outer circle in the CRISP-DM process model symbolizes the cyclical nature of data mining itself. Data mining does not end once a solution is deployed. The lessons learned during the process and from the deployed model solution can trigger new, often more focused business questions or new objectives. The rest of the chapter explain briefly about each phase in the CRISP-DM process model.

1.1 BUSINESS PROBLEM UNDERSTANDING

This initial phase focuses on understanding the goals, objectives, and requirements of the public transport system from both planning and operational perspectives. In the context of the Data Space, this involves **engaging with transport authorities**, **operators**, and **stakeholders** to identify key challenges, opportunities for service improvement, and regulatory considerations.

The insights gathered are then translated into a data mining problem definition and a preliminary plan, outlining a specific set of tasks and desired outcomes that leverage mobility data to support evidence-based decision-making and achieve strategic objectives in public transport management.

For greater clarity and a more comprehensive understanding, this deliverable encourages readers to consult Deliverable *D2.2 Data Space Governance*, which describes the stakeholders and the use cases defined within the Mobility Data Space project.

1.2 DATA UNDERSTANDING

The data understanding phase starts with initial data collection and proceeds with activities such as feature description, primary data analysis, and exploratory data analysis that enable to become familiar with the data, identify the data quality problems such as missing values, inconsistent data entries, and/or identify compelling subsets to form a hypothesis regarding confidential information. Data understanding has been described later in these pages in the chapter 6.3.2.

1.3 DATA PREPARATION

The data preparation phase encompasses all the processes required to transform raw, heterogeneous data into a structured and high-quality dataset ready for modelling.

This phase is iterative and flexible, often involving repeated cycles of refinement rather than following a strict linear sequence. Key activities include consolidating and organizing data tables, selecting relevant features, engineering new features to enhance predictive power, and applying necessary transformations to ensure consistency and usability.

Additionally, this phase involves thorough data cleaning to address issues such as missing values, outliers, and inconsistencies, thereby ensuring that the dataset is optimized for the chosen modelling tools and techniques. Effective data preparation is essential for achieving robust and reliable analytical results.

1.4 MODELLING

The modelling phase is a critical step within the Data Space, where advanced analytical techniques are selected and applied to address the diverse range of mobility-related use cases defined in the project. Depending on the specific problem statement, various modelling approaches are considered including statistical, machine learning, simulation, and optimization methods each tailored to the operational or planning context of the use case.

Throughout this phase, model parameters are iteratively calibrated and evaluated to achieve optimal performance, with careful attention to the data requirements of each technique. It is common for multiple modelling techniques to be compared for the same problem, ensuring robust and reliable insights. Moreover, the iterative nature of modelling often necessitates revisiting the data preparation phase to meet the specific input requirements of selected models.

Below, it has been presented an overview of the analytics and modelling approaches likely to be employed across the main use cases identified within the Mobility Data Space:

1.4.1 PUBLIC TRANSPORT

Use case: Optimization of bus routes and frequencies
Analytics: Network optimization, demand analysis, simulation
Models: Mixed-integer programming, genetic algorithms, agent-based simulation

Use case: Bus fleet route analysis and optimization
Analytics: Fleet scheduling, route matching, operational efficiency
Models: Vehicle scheduling algorithms, linear programming

Use case: Intermodal node demand analysis
Analytics: Demand forecasting, spatial analysis
Models: Time series forecasting (ARIMA, Prophet), regression analysis, spatial clustering

Use case: Real-time public transport user information
Analytics: Real-time data integration, predictive passenger flows
Models: Streaming analytics, short-term demand prediction, anomaly detection

Use case: Cross-border passenger transit analysis
Analytics: Flow analysis, cross-border mobility patterns
Models: Origin-destination matrix estimation, network flow models

Use case: Urban transport peak demand analysis
Analytics: Peak load identification, temporal demand modelling
Models: Time series clustering, forecasting models

1.4.2 FREIGHT TRANSPORT

Use case: Port-based freight flow analysis
Analytics: Flow estimation, operational efficiency
Models: Network analysis, simulation models

Use case: Airport-origin freight flow analysis
Analytics: Cargo movement tracking, capacity analysis
Models: Descriptive analytics, capacity forecasting

Use case: Logistics platform location in corridors
Analytics: Location optimization, corridor analysis
Models: Facility location models, spatial optimization

Use case: Cross-border freight transit analysis
Analytics: Transit time analysis, cross-border bottlenecks
Models: Statistical delay models, regression analysis

1.4.3 TRAFFIC

Use case: Main road segment capacity analysis
Analytics: Capacity estimation, traffic flow analysis
Models: Macroscopic traffic models, queuing theory

Use case: Predictive bottleneck resolution
Analytics: Bottleneck detection, congestion prediction
Models: Predictive machine learning (random forests, gradient boosting), simulation

Use case: Highway and main road congestion monitoring
Analytics: Real-time congestion tracking, anomaly detection
Models: Streaming analytics, time series anomaly detection

Use case: Special operations congestion prediction
Analytics: Scenario modelling, event-based congestion prediction
Models: Scenario simulation, event-driven forecasting

Use case: Real-time network traffic mapping
Analytics: Network-wide data integration, visualization
Models: Spatial-temporal mapping, heatmap generation

1.4.4 TRANSPORT EMISSION EVALUATION

Use case: Transport emissions evaluation
Analytics: Emissions estimation, sustainability assessment
Models: Emissions calculation models, life cycle assessment

Use case: Informed shared mobility
Analytics: Shared mobility usage patterns, optimization
Models: Clustering, recommendation systems

Use case: Freight emissions analysis
Analytics: Freight emissions calculation, route optimization
Models: Emissions factor models, route optimization algorithms

USE CASE: Electric vehicle charging infrastructure
Analytics: Charging demand forecasting, infrastructure planning
Models: Spatial optimization, time series forecasting

USE CASE: Urban mobility tracking by mode and distance
Analytics: Modal split analysis, mobility patterns
Models: Classification models, clustering

Use case: Low Emission Zone monitoring
Analytics: Compliance tracking, spatial impact analysis
Models: Geospatial analytics, rule-based monitoring

Use case: Active mobility monitoring
Analytics: Walking/cycling pattern analysis, infrastructure usage
Models: Time series analysis, clustering

Given the wide variety of analytics and modelling needs, flexibility and iteration are central to this phase. Models are continuously validated and refined based on performance metrics and operational feedback. As requirements evolve, new modelling techniques may be explored, and data may need to be re-prepared to accommodate additional features or to better meet model input specifications. This iterative, use-case-driven approach ensures that the Mobility Data Space delivers actionable, data-driven insights for both planning and operational excellence.

1.5 EVALUATION

The evaluation phase is one of the most crucial phases of any data science project lifecycle. At this stage in the project, the DS4MOBMODELS have built machine-learning models. Models might be performing well on training data, but it is necessary to test and evaluate it on unseen data for the model to achieve its objectives. Appropriate evaluation metrics are measured and well-tested at this stage. At the end of this phase, a decision on using the data mining results should be reached.

1.6 DEPLOYMENT

The creation of the model is generally not the end point within the data space for public transport. Even when a model's primary purpose is to analyse mobility data and understanding transport patterns, the insights gained must be effectively communicated and made accessible to a range of end users including transport planners, operational staff, policy makers, and even passengers or the general public. In the context of the data space, deployment may involve integrating predictive or optimization models into day-to-day transport management systems, such as real-time passenger information displays, dynamic fleet allocation, congestion alerts, or demand-responsive scheduling tools.

Depending on the specific needs, the deployment phase may be as straightforward as providing interactive dashboards for monitoring key indicators, or as complex as embedding automated data-driven processes within city-wide transport operations.

The ultimate goal is to ensure that the value generated by advanced analytics translates into actionable decisions, improved efficiency, and enhanced mobility services for all stakeholders who could be interest in a specific use case.

2. DATA SUMMARY

So far, the main concept of the DS4MOBMODELS is about data. The stakeholder involved in this project actively generated large amount of data every day. These data not only arise from their business activity but also come from external sources that provide context and meaning to this activity.

Although probably, in terms of quantity, it is not always necessarily dealing with new data but despite the fact that many of them are intrinsic to the business itself, their value, exploitation, and profitability have not been fully realized until recent years.

The ultimate goal of the DS4MOBMODELS is to inference to the data that has been generated the concept of **wisdom**. In order to understand what it means, it is necessary introduce the DIKW pyramid.

2.1 DIKW PYRAMID

Very often, the concepts of data and information are used interchangeably. We talk about data management and information management as if they were the same thing, perhaps giving the first a more operational and recursive nuance and context, and the second a more analytical one. At the same time, most would agree that knowledge is a concept that requires more elaboration and has a higher level of abstraction. If we also include wisdom within the set, then the conceptual complexity increases even further.

The relationship among these four concepts has been deeply studied, both from the perspective of information sciences and epistemology, even though there is no clear consensus among the different schools of thought. A common way to represent this relationship is through the so-called DIKW pyramid. The DIKW stands for: Data, Information, Knowledge and Wisdom.

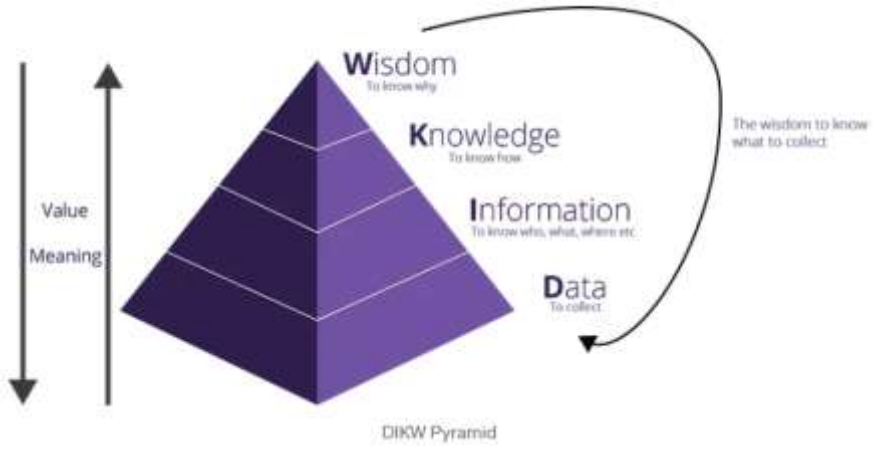


Figure 2: DIKW Pyramid

At the base of the pyramid lies on the **data**. Data is raw, unprocessed, and often unstructured. It consists of discrete facts or observations that lack context or meaning on their own. These data can come in many forms, such as integer, text, float among others and it is typically collected from various sources like sensors or user interactions.

In the context of the DS4MOBMODELS, raw data might include the number of passengers boarding a bus at each stop, GPS coordinates of vehicles throughout the day for example.

While this data is valuable, it is essentially meaningless in isolation it's just a collection of numbers and records.

The next level of the pyramid is **information**, which is data that has been processed, organized, and structured to provide context. Information answers basic questions like "who," "what," "where," and "when." It gives meaning to data by identifying patterns, trends, or relationships.

At the **knowledge** level, we move beyond simply organizing data and start interpreting it. Knowledge is about understanding relationships, causality, and implications. It answers the "how" and "why" questions, enabling us to draw meaningful conclusions and make predictions.

At the top of the pyramid is **wisdom**, the ultimate goal of the DIKW framework. Wisdom is the ability to apply knowledge in a meaningful and ethical way to make decisions, solve problems, and achieve goals. It involves critical thinking, experience, and judgment.

Wisdom is not just about knowing what to do. It's about understanding *why* a particular course of action is the best choice and considering the broader implications of that decision. It also requires balancing competing priorities, such as efficiency, fairness, and sustainability.

In addition to establishing a hierarchy, the model behind this pyramid provides a more or less agreed-upon definition of each of these four concepts, in such a way that each one builds upon the previous step.

From the point of view of information technologies more ascending the pyramid, there are concepts that are less programmable and less likely to be manipulated by algorithms and they are increasingly striving to contradict this. On this ascent, the DS4MOBMODELS also incorporate roles into this value chain. From a business perspective, business users those closer to decision-making are responsible for generating knowledge, entering at this level of the pyramid and leading the rest of the ascent. Up to this point, it is the engineers who are responsible for capturing the data and processing it to facilitate the generation of knowledge. The importance of each of these roles in this value chain is evident.

On a layer of governance, responsible for providing a series of common and unified services ranging from traceability to access control, an infrastructure is built for data management that enables the processing and transformation of data into information.

Although it is possible to characterize the different operations that make up this value chain, it is not always easy to establish when data ceases to be data and becomes information. There are blurred areas, especially when it comes to organization and distribution, where the differences are not clear. For this reason, it is common to associate information with systems responsible for access or consumption by end users, while data remains confined to those that are internal, under the responsibility of engineers and the technology department.

In addition to this approach, it is necessary to identify and position the elements we are dealing with. The DS4MOBMODELS has not lose sight of the fact that systems only know how to operate with data, and do so in an agnostic and decontextualized manner. In this sense, DS4MOBMODELS understands data as the basic storage units on which computers operate through processes and applications. This operational vision is compatible and reconcilable with our conceptual model, especially from the moment DS4MOBMODELS understands information as a set of processed and organized data. Therefore, when DS4MOBMODELS generally talks about data and also encompassing information as a concept, always adding the relevant nuances.

Regardless of what the DIKW pyramid means in the general theory of information and knowledge, for this project it is useful to define the context in which we will operate when addressing DS4MOBMODELS systems.

It is possible to define DS4MOBMODELS as the set of operations, techniques, and technologies oriented toward the processing of large and varied volumes of data, with the aim of generating valid information about which to develop knowledge and support business decision-making. That is, we focus on the base of the pyramid with the objective of enabling and facilitating the third level. Consequently, DS4MOBMODELS systems are those hardware and software components responsible for supporting these operations, this concept is covered in the chapter 5.

2.2 DATA TYPES

Within the DS4MOBMODELS project, a wide variety of data types will be managed and integrated to support advanced mobility modelling and analytics. As described in others deliverables dedicated to the Use Case involved in this project and the companies that generate them, it can be affirmed that these data include structured data such as public transport schedules, and vehicle telemetry, traffic counts as well as semi-structured data like mobility service usage, sensor outputs, and event records.

Additionally, the platform incorporates geospatial data like routes, zones, and network maps. Both public and private data sources will be considered, encompassing information generated by infrastructure, vehicles, mobile devices, and user interactions. By harmonizing these diverse data types across regions and stakeholders, the project ensures comprehensive coverage and interoperability, enabling robust support for real-time operations, long-term planning, and sustainable mobility initiatives within the European mobility ecosystem.

- **Public Transport Network Data** (e.g., GTFS): Includes static and dynamic schedules, stops, routes, timetables, service calendars, and network topology for multiple transport modes (bus, tram, metro, train, etc.).
- **Ticketing Data**: Validation data from physical fleet devices (onboard validators, ticket machines, gates). Validation and transaction data from web and mobile applications, including user purchases and digital ticket validations.
- **Traffic Data**: Real-time and historical road traffic data, such as traffic volumes, speeds, congestion levels, incidents, and sensor readings from highways and main roads.
- **Passenger Flow Data**: Boarding and alighting counts, origin-destination matrices, crowding levels, and passenger load statistics.
- **Freight and Logistics Data**: Cargo manifests, shipment tracking, port and airport freight movements, logistics hub operations, and corridor utilization data.
- **Vehicle and Fleet Data**: Fleet inventory, vehicle locations (GPS), assignments, maintenance records, and operational statuses.
- **Environmental and Emissions Data**: Measurements of air quality, greenhouse gas emissions, low emission zones monitoring, and electric vehicle charging events.
- **Urban Mobility and Active Transport Data**: Data on walking, cycling, shared mobility (bikes, scooters), and tracking of trips by mode and distance.
- **Infrastructure Data**: Information about charging stations, road segments, transit hubs, logistics

platforms, and other critical infrastructure.

- **User and Stakeholder Feedback**: Survey data, complaints, satisfaction ratings, and feedback from end-users and stakeholders.
- **Spatial and Geographic Data**: GIS layers, maps, geofences, and spatial boundaries for analytic and visualization purposes.

2.3 DATA USAGE

Within the context of the data space for public transport, re-use of existing data is a central aspect of the project's strategy. The project systematically identifies and leverages high-quality, relevant datasets that have been previously collected by transport authorities, operators, and open-data repositories. Examples of such data include:

- **Public Transport Network Data** (e.g., GTFS): Widely available from municipal or national transport agencies and open-data portals, these datasets provide up-to-date information on routes, schedules, stops, and network structures across different modes of transport. Their re-use enables rapid integration of multimodal network information for analysis, modelling, and service improvements.
- **Ticketing Data**: Where feasible, historical and ongoing ticket validation data collected by transport operators are re-used to analyse mobility patterns, demand forecasting, and service optimization. This may come from both on-board validation systems and digital channels (web/app).
- **Traffic Data**: Existing road traffic datasets, including real-time feeds and historical records, are sourced from traffic monitoring systems, city dashboards, and third-party providers for congestion analysis, bottleneck prediction, and multi-modal integration.
- **Environmental and Emissions Data**: Where available, air quality and emissions datasets from governmental monitoring stations or research initiatives are incorporated to support sustainability use cases.

Re-using existing data accelerates the project timeline, reduces duplication of effort, and ensures consistency with established data standards. It also enables the project to build upon proven, validated datasets, thereby improving the reliability of analytics and models.

Another aspect to take into account besides re-using data is creating new ones. Feature engineering as part of the data creation and preparation process. Feature engineering involves deriving new variables or attributes from raw data to enhance the performance of analytical models.

The process not only adds value to newly collected data but also maximizes the utility of both existing and generated datasets, ensuring that the resulting features are well-suited to address the specific analytical challenges and objectives of the project.

Any external datasets (e.g., GTFS feeds, open traffic data) are only re-used if their terms of use are compatible with the project's objectives. Careful attention is paid to copyright, licensing, and intellectual property rights.

Official open-data sources are prioritized, and, where necessary, explicit permission or data-sharing agreements are established with data owners. When data is purchased, contractual terms are reviewed to ensure compliance with copyright and data protection regulations.

2.4 DATA UTILITY

The data generated and managed by this project holds significant utility beyond the immediate scope of the consortium. Potential beneficiaries include:

- **Public Sector Authorities:** Regional and national transport agencies, city planners, and policy-makers could leverage the data to inform evidence-based decisions, optimize public transport networks, and support sustainable urban mobility initiatives.
- **Research and Academic Institutions:** Universities and research centers focused on mobility, urban planning, data science, and smart cities may use the datasets for further analysis, comparative studies, and the development of new methodologies and models.
- **Private Sector Companies:** Transport operators, mobility service providers, and technology companies can utilize the data to develop or enhance digital mobility solutions, integrate services, and improve operational efficiency.
- **Civic Organizations and NGOs:** Advocacy groups and non-profit organizations working on transport accessibility, environmental sustainability, and inclusive mobility can use the data to support their missions and propose data-driven policy recommendations.
- **General Public and Everyday Users:** Citizens, commuters, and community groups may benefit from increased transparency and access to mobility data, empowering them to make informed travel choices and participate more actively in co-designing local transport solutions.
- **European and International Initiatives:** Organizations involved in transnational projects (e.g., TEN-T, EU open data initiatives) could integrate the data to enable cross-border analysis and foster harmonized mobility practices.

2.5 DATA STATES

A fourth dimension of data that is more related with security but a preview needed to be understood. Data security will be covered in chapter 6.1.

There are three basic states of data within any organization:

At Rest: This involves data that is stored with high latency (low-changing state). Usually, this is any kind of historical data that's stored in any system, no matter if it's an application, a cloud database, or even cloud- or locally-stored files (Excel, on-prem databases, among others)

In Motion: This is data that travels from one place to another. Again, with a broad understanding, it is possible to think about any data going from an application to a database.

In Use: Data in use is data that is being edited, accessed, and processed by users or applications. Examples include opened documents, real-time analytics programs, and data processed by CPUs or stored in memory (RAM).

3 FAIR DATA

It has been described previously that a data space can be considered as federated digital ecosystem where organizations and stakeholders can share, access, and exchange data in a secure, controlled, and interoperable manner.

In the context of mobility and transport, a data space enables the integration of diverse datasets from various regions and providers, fostering collaboration and innovation. Central to the effectiveness of a data space are the FAIR principles, Findability, Accessibility, Interoperability, and Reusability which ensure that data is

properly described, easy to locate, accessible under clear conditions, compatible across different systems, and reusable for future applications.

Adhering to the FAIR principles not only maximizes the value of data assets but also supports transparency, efficiency, and trust among stakeholders. This is particularly important for large-scale, cross-border mobility initiatives where high-quality, interoperable data is essential for optimizing traffic management, supporting sustainable mobility strategies, and enabling advanced analytics and real-time decision-making.

Before to dive into the FAIR principles, it's necessary to introduce the process which guide the DS4MOBMODELS to model the four principles.

3.1 FAIRIFICATION PROCESS

The FAIR Data Principles apply to metadata, data, and supporting infrastructure (e.g., search engines). Most of the requirements for findability and accessibility can be achieved at the metadata level. Interoperability and reuse require more efforts at the data level. The scheme below depicts the FAIRification process, focusing on data, but also indicating the required work for metadata:

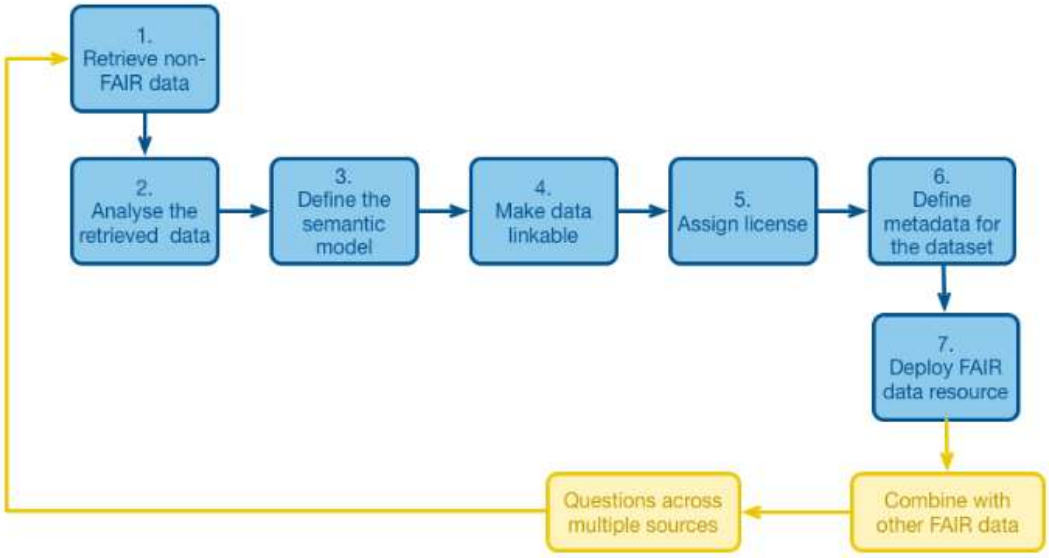


Figure 3: FAIRification process

The FAIRification process consists of the following steps:

- **Retrieve non-FAIR data**: gain access to the data to be FAIRified.
- **Analyse the retrieved data**: inspect the content of the data: Which concepts are represented? What is the structure of the data? What are the relations between the data elements? Different data distributions require different methods for identification and analysis. For instance, most of the data within the DS4MOBMODELS is in a relational database. The relational schema provides information about the dataset structure, the types involved (the field names), cardinality, etc.
- **Define the semantic model**: define a *semantic model* for the dataset, which describes the meaning of entities and relations in the dataset accurately, unambiguously, and in a computer-actionable way. Depending on the dataset, defining a proper semantic model may require a significant effort, even for experienced data modellers.

  A good semantic model should represent a consensus view in a particular domain, for a particular

purpose. Therefore, it is good practice to search for existing models. Semantic models often contain multiple terms from existing ontologies and vocabularies. A vocabulary is a computer-readable file that captures terms, their URIs, and descriptions. An ontology can be roughly described as a vocabulary with hierarchies, meaningful relations among concepts, and their constraints. These conceptual models allow us to classify our data models and data items using the provided terms, concepts, and conceptual structures.

- **Make data linkable**: The non-FAIR data can be transformed into linkable data by applying the semantic model defined in step 3. Currently, this is done using Semantic Web and Linked Data technologies. This step promotes interoperability and reuse, facilitating the integration of the data with other types of data and systems. However, the user should evaluate the feasibility of this step for the given data. It is a sensible thing to do for many types of data (e.g., structured data), but it may not be relevant for other types.
- **Assign license**: Although license information is part of the metadata, we have incorporated the license assignment as a separate step in the FAIRification process to highlight its importance. The absence of an explicit license may prevent others to reuse data, even if the data is intended to be open access.
- **Define metadata for the dataset**: As explained by many of the FAIR principles, proper and rich metadata support all aspects of FAIR.
- **Deploy FAIR data resource**: deploy or publish the FAIRified data, together with relevant metadata and a license, so that the metadata can be indexed by search engines and the data can be accessed, even if authentication and authorisation are required.

In the following subchapters can discover the FAIR principles.

3.2 FINDABLE

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

- **F1. (Meta)data are assigned a globally unique and persistent identifier:** Principle F1 is arguably the most important because it will be hard to achieve other aspects of FAIR without globally unique and persistent identifiers. Hence, compliance with F1 will already take you a long way towards publishing FAIR data.

  Globally unique and persistent identifiers remove ambiguity in the meaning of your published data by assigning a unique identifier to every element of metadata and every concept/measurement in your dataset. In the context of the DS4MOBMODELS, identifiers consist in a alpha-numeric string which will be related by its domain, making it effectively unique. Many data repositories will automatically generate globally unique and persistent identifiers to deposited datasets. Identifiers can help other people understand exactly what you mean, and they allow computers to interpret your data in a meaningful way. In addition, identifiers will help others to properly cite your work when reusing your data.

- **F2. Data are described with rich metadata (defined by R1 below):** In creating FAIR digital resources, metadata can (and should) be generous and extensive, including descriptive information about the context, quality and condition, or characteristics of the data. Rich metadata allow a computer to automatically accomplish routine and tedious sorting and prioritising tasks that currently demand a lot of attention from researchers. The rationale behind this principle is that someone should be able

to find data based on the information provided by their metadata, even without the data's identifier. As such, compliance with F2 helps people to locate your data, and increase re-use and citations.

- **F3. Metadata clearly and explicitly include the identifier of the data they describe:** This is a simple and obvious principle, but of critical importance to FAIR. The metadata and the dataset they describe are usually separate files. The association between a metadata file and the dataset should be made explicit by mentioning a dataset's globally unique and persistent identifier in the metadata. As stated in F1, many repositories will generate globally unique and persistent identifiers for deposited datasets that can be used for this purpose.
- **F4. (Meta)data are registered or indexed in a searchable resource:** Identifiers and rich metadata descriptions alone will not ensure *findability*. Perfectly good data resources may go unused simply because no one knows they exist. If the availability of a digital resource such as a dataset, service or repository is not known, then nobody and no machine can discover it. There are many ways in which digital resources can be made discoverable, including indexing.

3.3 ACCESSIBLE

Once one of the stakeholder DS4MOBMODELS finds the required data, it is needed to know how they can be accessed, possibly including authentication and authorisation.

- **A1. (Meta)data are retrievable by their identifier using a standardised communications protocol:** (Meta)data are retrievable by their identifier using a standardised communications protocol within the Data Space. The Data Space have a connector module which enables secure, standardised, and interoperable access to data resources by acting as a gateway between data providers and consumers.

  The connector module ensures that both metadata and datasets can be discovered and accessed reliably through well-defined APIs or protocols (such as HTTPS or RESTful services). This guarantees that, once data or metadata is identified through a persistent and unique identifier, it can be retrieved efficiently and consistently within the Data Space ecosystem, regardless of the underlying platform or technology.

  - **A1.1 The protocol is open, free, and universally implementable:** In order to maximize data interoperability and reusage, the mobility data space will use communication protocol used for data and metadata exchange is open (publicly available specifications), free of charge, and universally implementable. This ensures that all participants in the data space, regardless of their technical platform or geographical location, can retrieve at least the metadata without restrictions. Such an approach lowers barriers to entry, supports broad participation, and aligns with the principles of fairness and transparency. Consequently, this criterion will directly influence the selection of connectors, repositories, or platforms within the data space where data will be shared and accessed.
  - **A1.2 The protocol allows for an authentication and authorisation procedure, where necessary:** The *A* in FAIR does not necessarily mean open or free. Rather, it implies that one should provide the exact conditions under which the data are accessible. Hence, even heavily protected and private data can be FAIR. Ideally, accessibility is specified in such a way that a machine can automatically understand the requirements, and then either automatically execute the requirements or alert the user to the requirements. It often makes sense to request users to create a user account for a repository. This allows to authenticate the owner (or contributor) of each dataset, and to potentially set user-specific rights. Hence, this

criterion will also affect your choice of the repository where you will share your data.

- **A2. Metadata are accessible, even when the data are no longer available:** Datasets tend to degrade or disappear over time because there is a cost to maintaining an online presence for data resources. When this happens, links become invalid and users waste time hunting for data that might no longer be there. Storing the metadata generally is much easier and cheaper. Hence, principle A2 states that metadata should persist even when the data are no longer sustained. A2 is related to the registration and indexing issues described in F4.

3.4 INTEROPERABLE

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

- **I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation**: Humans should be able to exchange and interpret each other's. But this also applies to computers, meaning that data that should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings. The main goal of this principle is to provide a common understanding of digital objects by means of a language for knowledge representation to be used to represent these objects.

  The principle I1 defines some properties that these languages should have. The chosen language should have a formal specification, i.e., the language's syntax and grammar are defined in a precise way. Another requirement is that the knowledge representation language specifications should be shared and accessible so others can read the specifications and learn the language. Finally, in other to support interoperability, the language should be designed to be used in more than one scenario.

- **I2. (Meta)data use vocabularies that follow FAIR principles:** When we are describing data or metadata, we often use vocabularies that provide the terms or concepts that are adequate to represent their content. However, if we use vocabularies in our data or metadata, we should make sure that they are also FAIR in their own right so that others, humans or machines, can find, access, interoperate and reuse them. The controlled vocabulary used to describe datasets needs to be documented and resolvable using globally unique and persistent identifiers. This documentation needs to be easily findable and accessible by anyone who uses the dataset.

  Communities should define the required FAIRness level of the vocabularies used in their midst. Minimally, it is reasonable to expect that the vocabulary and its terms/concepts have globally unique and persistent identifiers (F1) that can be resolved using a standardised communication protocol (A1) and is described with a formal, accessible, shared and broadly applicable language for knowledge representation (I1).

- **I3. (Meta)data include qualified references to other (meta)data:** A qualified reference is a cross-reference that explains its intent. For example, *X is regulator of Y* is a much more qualified reference than *X is associated with Y*, or *X see also Y*. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data, balanced against the time/energy involved in making a good data model. To be more concrete, you should specify if one dataset builds on another data set, if additional datasets are needed to complete the data, or if complementary information is stored in a different dataset. In particular, the scientific links between the datasets need to be described. Furthermore, all datasets need to be properly cited (i.e., including their globally unique and persistent identifiers).

3.5 REUSABLE

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

- **R1. (Meta)data are richly described with a plurality of accurate and relevant attributes:** It will be much easier to find and reuse data if there are many labels are attached to the data. Principle R1 is related to F2, but R1 focuses on the ability of a user (machine or human) to decide if the data is actually USEFUL in a particular context. To make this decision, the data publisher should provide not just metadata that allows discovery, but also metadata that richly describes the context under which the data was generated. This may include the experimental protocols, the manufacturer and brand of the machine or sensor that created the data, the species used, the drug regime, etc. Moreover, R1 states that the data publisher should not attempt to predict the data consumer's identity and needs. We chose the term 'plurality' to indicate that the metadata author should be as generous as possible in providing metadata, even including information that may seem irrelevant.
- **R1.1. (Meta)data are released with a clear and accessible data usage license:** Under I, we covered elements of *technical* interoperability. R1.1 is about *legal* interoperability. What usage rights do you attach to your data? This should be described clearly. Ambiguity could severely limit the reuse of your data by organisations that struggle to comply with licensing restrictions. Clarity of licensing status will become more important with automated searches involving more licensing considerations. The conditions under which the data can be used should be clear to machines and humans.
- **R1.2. (Meta)data are associated with detailed provenance:** For others to reuse your data, they should know where the data came from, who to cite and/or how you wish to be acknowledged. Include a description of the workflow that led to your data: Who generated or collected it? How has it been processed? Has it been published before? Does it contain data from someone else that you may have transformed or completed? Ideally, this workflow is described in a machine-readable format.
- **R1.3. (Meta)data meet domain-relevant community standards:** It is easier to reuse data sets if they are similar: same type of data, data organised in a standardised way, well-established and sustainable file formats, documentation (metadata) following a common template and using common vocabulary. If community standards or best practices for data archiving and sharing exist, they should be followed. For instance, many communities have minimal information standards (e.g., MIAME, MIAPE). FAIR data should at least meet those standards. Other community standards may be less formal, but nevertheless, publishing (meta)data in a manner that increases its use(ability) for the community is the primary objective of FAIRness. In some situations, a submitter may have valid and specified reasons to divert from the standard good practice for the type of data to be submitted. This should be addressed in the metadata. Note that quality issues are not addressed by the FAIR principles. The data's reliability lies in the eye of the beholder and depends on the intended application.

4 OTHER RESEARCH OUTPUTS

In addition to the management of data, the DS4MOBMODELS project recognizes the importance of planning for the management and sharing of other research outputs generated or re-used throughout the project's lifecycle. These outputs may include digital assets such as software (e.g., simulation tools, data processing scripts, analytics dashboards), workflows and protocols (e.g., data integration pipelines, feature engineering methods, model validation procedures), and analytical models (e.g., predictive algorithms, optimization models, simulation frameworks). While the primary focus is on digital outputs, any physical

outputs will also be managed according to best practices for preservation and sharing that will be establish during the lifetime of the present project.

To ensure full alignment with the FAIR (Findable, Accessible, Interoperable, Reusable) principles, the project applies similar management strategies to these research outputs as it does for data:

- **Findability**: All digital outputs will be thoroughly documented with comprehensive metadata and assigned persistent identifiers. As required by the project, these outputs will be registered and uploaded to the European Portal, ensuring they are accessible to all users as needed.
- **Accessibility**: Wherever possible, software, workflows, and models will be made openly available under appropriate open-source or open-access licenses. Access protocols, usage instructions, and any necessary dependencies will be clearly described, ensuring that others can access and use these outputs without barriers. The accessibility rules will be established across stakeholders.
- **Interoperability**: Research outputs will be developed using widely adopted standards, formats, and interfaces (e.g., open APIs, standard data models, interoperable software libraries) to facilitate integration with other tools, datasets, and platforms within the mobility and transport domain.
- **Reusability**: Comprehensive documentation, versioning, and licensing information will accompany each output, enabling others to understand, adapt, and re-use them in new contexts. Where appropriate, workflows and protocols will be described in machine-readable formats (e.g., using workflow languages or protocol registries) to support automation and reproducibility.

The project will regularly review and update its management plan for all research outputs, ensuring that new developments, evolving best practices, and stakeholder needs are addressed. By extending the FAIR principles to software, models, and other non-data assets, the project maximizes the scientific and societal value of its research, fostering transparency, innovation, and collaboration both within and beyond the public transport sector.

5 ALLOCATION OF RESOURCES

- What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?
- How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)
- Who will be responsible for data management in your project?
- How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

6 DATA SECURITY, PRIVACY AND QUALITY

To Ensure data security, privacy and quality many public transport authorities and operators have introduced automated systems to support interoperable and efficient data exchange across mobility services. Stakeholders such as transport operators, mobility service providers, and authorities input relevant information into these systems, enabling further processing and secure sharing of data within the Data Space ecosystem.

The submitted data is subsequently used for a wide range of purposes, including service optimization, demand forecasting, passenger information, regulatory compliance, safety monitoring, and strategic planning.

In order to be successful, the DS4MOBMODELS project has to ensure the security, privacy, and quality of data.

- **Data security** involves protecting data from unauthorized access and potential breaches throughout its lifecycle, including creation, transmission, storage, and sharing within the Data Space.
- **Data privacy** considerations ensure that any shared data complies with relevant legal and ethical standards, such as data minimization and anonymization, and respects the rights of individuals whose data may be processed.
- **Data Quality** encompasses attributes such as accuracy, completeness, consistency, and reliability. For example, accurate and up-to-date service schedules, real-time vehicle locations, and passenger counts are necessary for effective operations and decision-making.

If data exchanged within the transport Data Space is not accurate, secure, or privacy-compliant, its reuse can have far-reaching negative consequences. Inaccurate or compromised data can lead to incorrect service planning, inefficiencies, reduced safety, privacy violations, and a loss of trust among stakeholders. The repercussions are amplified when poor-quality or insecure data is reused by partner organizations or public authorities, further undermining the reliability of the transport ecosystem.

Therefore, when establishing a Data Space for public transport, the partners involved within the project need to prioritize robust approaches for data security, privacy, and quality. This can be achieved through both technical measures and organizational measures.

Technical mechanisms help maintain the security and integrity of data fields and enforce privacy requirements, while organizational practices support ongoing data quality and compliance.

Below are several suggestions on implementing measures to ensure data security, privacy, and quality in automated public transport Data Spaces:

- Employ strong encryption and authentication for all data exchanges.
- Apply strict access control and monitoring to prevent unauthorized data access or manipulation.
- Enforce data validation rules to promote accuracy and consistency.
- Implement privacy-by-design principles, including data minimization and anonymization where appropriate.
- Establish regular data quality assessments and audits.
- Provide ongoing training and awareness for staff and stakeholders regarding data security, privacy, and quality best practices.

6.1 DATA SECURITY

As organizations increasingly leverage cloud technologies to enable seamless data sharing and collaboration, the concept of the Data Space has emerged as a critical framework for secure and trustworthy data exchange across organizational and sectoral boundaries.

The analysis of how the DS4MOBMODELS stores the data, the cloud technology came in minds and the technical aspect of it will be reflected in the *D5.2 Technical specifications of the cloud infrastructure*. Doing so, participants share, access, and use data in a federated environment providing scalability, flexibility, and connectivity.

However, the use of cloud platform introduces new challenges and considerations for data security. Sensitive data moving through the Data Space must be adequately protected against unauthorized access, data breaches, and misuse. Ensuring data security in the cloud requires a comprehensive approach that encompasses technical safeguards such as encryption, identity and access management, secure APIs, and continuous monitoring as well as clear governance policies and compliance with relevant regulations.

The objective is to foster a trusted environment where all participants in the Data Space can confidently share and utilize data, knowing that security, privacy, and data sovereignty are maintained at every stage of the data lifecycle. Addressing data security in the cloud is therefore fundamental to the success of Data Spaces, enabling innovation and collaboration while protecting the interests of all stakeholders.

Cloud data security refers to the technologies, services, policies, and processes that protect sensitive data and other digital assets—within, across, and outside of clouds. It helps ensure that assets remain safe from security threats, human error, and internal threats like data loss, leakage, misuse from breaches, corruption, theft, and unauthorized access. Sensitive data can include public and non-public information such as names, birthdates, government information, IP address, intellectual property, and biometric information.

6.1.1 CLOUD SECURITY

Cloud-based technologies enable collaboration across many workspaces and geographic regions but can prove most difficult to shield from cyberattacks, ransomware, and data leaks.

People may confuse cloud data security for data security or cloud security but cloud data security is not just about the data you have. It also encompasses data not bound by the constraints of your hardware. This includes, as they have been described in chapter 2.5:

- **Data in use**: Securing data used within an application.
- **Data in motion**: Transmitting data safely as it moves within a network through encryption or additional security measures.
- **Data at rest**: Protecting data stored in any network location.

DS4MOBMODELS needs to reduce the risk of threats as much as possible while protecting the stakeholders data, managing its usage, and always staying operational. DS4MOBMODELS implements a cloud-native application protection platform (CNAPP) helps optimize your data security efforts by correlating intelligence and distilling that data information into a single view.

Data security managed within the project accomplish the following functions:



Encryption: Encryption is a technique that scrambles data so that only authorized users can view it. This technology ensures that only those with permission can access that data while deterring attackers from leaking, selling, or using that data for consecutive attacks.



Recovery: In case of any system failure, corruption, or breach, you'll need a plan to recover any critical data. Your team needs a backup in a different format, like a hard drive or local network.



Data obscuring: By masking letters with proxy characters, you can add another layer to your cloud data protection. Even if attackers can access your data, the information will be unreadable until an authorized user returns it to its original state.



Authentication: Before you're able to gain access to data, implementing authentication technology like passwords, swipe cards, and tokens can help safeguard the system. An identity and access management (IAM) system can quickly verify a person's identity and grant the permissions needed to complete a task.



Data erasure: This method ensures data is appropriately discarded so hackers cannot obtain and use that data for malicious purposes. Another option could be data wiping, which erases the unused data, but those files might still be vulnerable to threats. Erasing data instead of data wiping might work better as that data is removed from all storage devices and can't be recovered.



Private cloud: A private cloud is a computing service offered through the internet or private network to select users in a single business. Ensuring that your data is inaccessible to third parties, private clouds offer a high level of security and privacy through firewalls and internal hosting.

6.1.2 OWASP SECURITY

In addition to these functionality, DS4MOBMODELS is taking into account the OWASP Top 10. The OWASP Top 10 is a report, or **awareness document** that outlines security concerns surrounding web application security. It is updated regularly to ensure it consistently presents the 10 most critical risks facing organizations. OWASP recommends that all companies incorporate the document's findings into their corporate processes to ensure they minimize and mitigate the latest security risks.

The OWASP Vulnerability Report is based on a consensus of security experts from around the world. It ranks risks based on the frequency of security flaws, the severity of the vulnerability, and its potential impact. This provides developers and security professionals with information on the most prominent risks and enables them to minimize the potential for risks in their organizations' security practices.

By the time, the present deliverable was write, the latest OWASP report lists the top 10 vulnerabilities as follows:

- Injection
- Authentication interrupted
- Exposure of confidential data
- XML External Entities (XXE)
- Access control broken
- Incorrect security settings
- Cross-site scripting (XSS)
- Insecure deserialization
- Use of components with known vulnerabilities
- Insufficient logging and monitoring

6.1.2.1 Injection

Injection attacks occur when untrusted data is injected into web applications via form input or other types of data submissions. A common type of injection attack is a Structured Query Language (SQLi) injection, which occurs when cybercriminals inject SQL database code into an online form used for plain text.

These types of attacks can be prevented by sanitizing and validating data submitted by users. Data validation ensures that suspicious data is rejected, and data sanitization helps organizations clean up data that appears suspicious. Database administrators can also establish controls that minimize the amount of information that injection attacks can expose.

6.1.2.2 Authentication interrupted

Authentication vulnerabilities can allow attackers to gain access to user accounts, including administrator accounts, which they could use to compromise and take full control of corporate systems.

Websites commonly suffer from broken authentication, which typically occurs as a result of issues in the application's authentication mechanism. This includes poor session management, which attackers can exploit using brute-force techniques to guess or confirm user accounts and login credentials.

The OWASP Top 10 provides a list of broken authentication vulnerabilities, including web applications that:

- Allow attacks such as credential stuffing
- Allow weak or default passwords
- Employing inefficient processes for lost user credentials and passwords
- Missing or ineffective multi-factor authentication (MFA)
- Expose session IDs in the Uniform Resource Locator (URL), do not rotate session IDs, and do not properly invalidate session IDs and authentication tokens after a period of inactivity

These vulnerabilities are typically caused by insecure software, which is often the result of inexperienced developers writing them, a lack of security testing, and rushed software releases.

Broken authentication vulnerabilities can be mitigated by implementing MFA methods, which offer greater certainty that a user is who they claim to be and prevent automated and brute-force attacks. These vulnerabilities can also be prevented by ensuring developers apply best practices to website security and are given adequate time to properly test code before applications are put into production.

Other tactics include checking for weak passwords, ensuring users protect their accounts with strong, unique passwords, and using secure session managers.

6.1.2.3 Exposure of confidential data

The exposure or leakage of sensitive data is one of the most common forms of cyberattack. Sensitive data, such as credit card information, medical details, Social Security numbers, and user passwords, can be exposed if a web application doesn't effectively protect it. Attackers who are able to access and steal this information can use it as part of broader attacks or sell it to third parties.

Protecting sensitive data is increasingly important given the strict rules and penalties of data and privacy regulations, such as the European Union's General Data Protection Regulation (GDPR). To do so, organizations must be able to protect data at rest and data in transit between servers and web browsers.

A website's data can be protected using a Secure Sockets Layer (SSL) certificate, which establishes an encrypted link between a web browser and a server. It also protects the integrity of the data while in transit between a server or firewall and the web browser. Exposure of sensitive data can also be prevented by encrypting data through secure encryption processes, protecting stored passwords with secure hash functions, and ensuring that strong and up-to-date algorithms, keys, and protocols are in place.

6.1.2.4 XML External Entities (XXE)

XXE attacks target web applications that parse Extensible Markup Language (XML). They occur when a weakly configured XML parser processes XML input that contains a reference to an external entity, such as a hard drive. XML parsers are often vulnerable to XXE by default, meaning developers must manually eliminate the vulnerability.

The OWASP Top 10 states that XXE attacks typically target vulnerable XML processors, vulnerable code, dependencies, and integrations.

XXE attacks can be prevented by ensuring that web applications accept less complex forms of data (such as JavaScript Object Notation (JSON) web tokens), patching XML parsers, or disabling the use of external

entities. Organizations can also defend against XXE attacks by implementing application programming interface (API) security gateways, virtual patching, and Web Application Firewalls (WAFs).

6.1.2.5 Broken access control

Access control refers to the specific data, websites, databases, networks, or resources that users are allowed to visit or access. Broken access controls give users access to resources beyond what they require. This allows attackers to bypass access restrictions, gain unauthorized access to sensitive systems and data, and potentially gain access to administrator and privileged user accounts.

The risk of breached access controls can be reduced by implementing the concept of least privileged access, regularly auditing servers and websites, enforcing MFA, and removing inactive users and unnecessary services from servers. Organizations can also protect access controls by using authorization tokens when users log in to a web application and invalidating them after they log out. Other recommendations include logging and reporting access failures and using rate limiting to minimize the damage caused by automated attacks.

6.1.2.6 Incorrect security settings

Security misconfigurations are considered the most common vulnerability in the OWASP Top 10. They are most often caused by organizations using default website or content management system (CMS) configurations, which can inadvertently expose application vulnerabilities. Common misconfigurations also include failure to patch software flaws, unused web pages, unsecured directories and files, default sharing permissions on cloud storage services, and unused or unnecessary services.

Security misconfiguration can occur across the entire application stack: web and application servers, databases, network services, custom code, frameworks, pre-installed virtual machines, and containers.

Security misconfigurations can be avoided by changing the webmaster or CMS's default settings, removing unused code features, and controlling user feedback and the visibility of user information. Developers should also remove unnecessary documentation, features, frameworks, and samples, segment the application architecture, and automate the effectiveness of web environment configurations and adjustments.

6.1.2.7 Cross-site scripting (XSS)

A cross-site scripting vulnerability occurs when web applications allow users to submit custom code to public URLs or websites. XSS attacks occur when cybercriminals inject malicious scripts into a website, allowing them to modify the website’s display. The attacker then relies on victims visiting the page from a browser to execute their code, which they typically accomplish through social engineering or embedding malicious links in phishing emails . Exploiting an XSS vulnerability can give an attacker full control over browsers and allow them to inject malicious JavaScript code into websites.

Organizations can prevent XSS vulnerabilities by using a WAF to mitigate and block attacks , while developers can reduce the chances of XSS attacks by separating untrusted data from active browsers. This includes using frameworks that prevent XSS by design, implementing data sanitization and validation, preventing untrusted Hypertext Transfer Protocol (HTTP) request data, and implementing a Content Security Policy (CSP).

6.1.2.8 Insecure deserialization

In terms of data storage and computer science, serialization means converting objects or data structures into byte strings. Deserialization means converting those byte strings into objects. Insecure deserialization involves attackers manipulating the data before it is deserialized.

OWASP's protection advisory regarding insecure deserialization focuses on supercookies containing serialized user information. If attackers can successfully deserialize an object, they can assign themselves an administrator role, serialize the data, and compromise entire web applications.

This can be prevented by prohibiting serialized objects and prohibiting the deserialization of data from untrusted sources. OWASP also recommends monitoring deserialization activity, implementing integrity checks on any serialized objects to prevent data tampering, isolating deserialized code in low-privilege environments, ensuring all deserialization exceptions and failures are logged, and restricting and monitoring network connectivity from containers and servers that deserialize data.

6.1.2.9 Use of components with known vulnerabilities

Software components such as frameworks and libraries are often used in web applications to provide specific functionality, such as icon sharing and A/B testing. However, these components can often lead to vulnerabilities that, unknown to developers, provide a security hole for an attacker to launch a cyberattack.

This is often due to developers not keeping applications up-to-date, legacy code not working on new updates, and web administrators being concerned about updates breaking their websites or lacking the expertise to apply updates. However, attackers are constantly on the lookout for potential vulnerabilities that developers have not detected, commonly known as zero-day attacks, that they can exploit.

Organizations can prevent this through virtual patching, which protects outdated websites from exploits through the use of firewalls, intrusion detection systems (IDS), and a WAF. Vulnerabilities can also be prevented by maintaining a component inventory and removing any unused or unmaintained components, only using components from trusted sources, and ensuring all components are patched and updated at all times.

6.1.2.10 Insufficient registration and monitoring

Many web applications don't do enough to detect data breaches , allowing attackers to not only gain unauthorized access to their systems but also allow them to remain in place for months and years. Organizations should log and monitor their applications for unusual or malicious behavior to prevent their websites from being compromised.

Companies should also maintain audit trails that allow them to track any suspicious changes, record anomalous activity, and trace unauthorized access or account compromises.

6.1.3 SHAREPOINT SECURITY

Security in DS4MOBMODELS project is not just about data. In addition to the many prevention rules establish within it, security is present also the document repository available in the project to the stakeholder. This repository is the Sharepoint.

SharePoint offers robust built-in security features such as authentication, role-based access control, encryption, and audit logs, true end-to-end security requires measures that extend beyond the platform. These additional layers of security help mitigate risks related to user behavior, integration with other systems, data flows, and compliance requirements. Key aspects include:

- **Network Security**:
  - **Firewalls and Network Segmentation**: Ensure SharePoint servers are protected by network firewalls and segmented from less-secure network zones.
  - **VPN/Zero Trust Access**: Require VPN or zero trust access models for remote users, limiting exposure to only necessary endpoints.
- **Identity and Access Management:**
  - **Multi-Factor Authentication (MFA):** Enforce MFA for all users, especially administrators and external collaborators.
  - **Single Sign-On (SSO) Integration**: Integrate with secure enterprise identity providers for

centralized user management and de-provisioning.

- **Least Privilege Principle**: Regularly review permissions to ensure users only have access to what they need.

- **Data Protection**:
  - **Data Loss Prevention (DLP):** Implement DLP policies to detect and prevent sensitive data from being shared or leaked outside authorized channels.
  - **Document Encryption**: Use additional encryption (beyond SharePoint's default) for highly sensitive files, both at rest and in transit.
  - **Information Rights Management (IRM):** Apply IRM to restrict what users can do with documents (e.g., prevent downloading, printing, or forwarding).
- **Monitoring and Incident Response**:
  - **Continuous Monitoring**: Use SIEM (Security Information and Event Management) tools to monitor for unusual activity, failed login attempts, or suspicious file access.
  - **Alerts and Automated Responses**: Configure alerts for potential data breaches or policy violations, and establish automated workflows for rapid response.

6.2 DATA PRIVACY

Data governance refers to the overall management of the availability, usability, integrity, and security of the data used in an organization. It involves defining who owns the data, establishing data quality standards, and implementing data access controls.

Privacy, on the other hand, refers to the protection of personal information, ensuring that it is collected, processed, and used only for the purposes for which it is intended.

By combining the concepts of data governance and privacy, DS4MOBMODELS can create a robust framework that ensures the success of their data analytics initiatives, while also safeguarding sensitive stakeholders information from being compromised or leaked to the public. This powerful combination enables organizations to maintain compliance with regulatory requirements, gain customer trust, and avoid reputational damage caused by data breaches.

6.2.1 DATA GOVERNANCE AND PRIVACY BENEFITS

Implementing data governance and privacy has numerous benefits for organizations.

- **Improves data quality**: Accurate, complete, and reliable data leads to better decision-making and improved business outcomes
- **Reduces the risk of data breaches and non-compliance with data protection regulations:** The consequences of a data breach can be severe, including financial losses, damage to reputation, and legal liabilities. By implementing proper data governance and privacy measures, organizations can reduce the risk of these incidents occurring.
- **Increases trust in data analytics**: Customers, partners, and other stakeholders are more likely to trust organizations with strong data governance and privacy policies in place, leading to increased loyalty, improved relationships, and better business outcomes

6.2.2 DATA GOVERNANCE AND PRIVACY BEST PRACTICES

Effective data governance is essential to ensuring that data is managed in a secure, compliant, and effective manner. Here are some best practices for data governance:

- **Establish Clear Policies and Procedures**: Data governance policies and procedures should be clearly defined, communicated, and implemented. This includes data quality standards, data classification rules, data access controls, and data retention policies.
- **Define Data Ownership**: It is essential to clearly define data ownership, including who is responsible for the accuracy, completeness, and security of the data.
- **Establish Data Quality Standards**: Data quality standards should be established to ensure that data is accurate, consistent, and relevant. This includes establishing data validation rules and ensuring that data is properly formatted and normalized.
- **Implement Data Access Controls**: Data access controls should be implemented to ensure that data is only accessed by authorized users. This includes implementing role-based access controls, data encryption, and multi-factor authentication.
- **Monitor and Audit Data Usage**: Regular monitoring and auditing of data usage is critical to ensure that data is being used appropriately and that data governance policies and procedures are being followed.

6.3 DATA QUALITY

Data quality measures how well a dataset meets criteria for accuracy, completeness, validity, consistency, uniqueness, timeliness and fitness for purpose, and it is critical to all data governance initiatives. To respond to the objectives of the DS4MOBMODELS project data quality standards have to be meet in order to make data-driven decisions.

6.3.1 IMPORTANCE OF DATA QUALITY

In addition, Data quality is critically important in a Data Space context for several reasons:

- **Trust and Collaboration:** Data Spaces are built on the principle of trusted data sharing among multiple organizations or stakeholders. High-quality data ensures that all participants can rely on the data exchanged, fostering trust and effective collaboration.
- **Accurate Decision-Making:** Organizations use shared data to inform operational, strategic, and regulatory decisions. Inaccurate, incomplete, or inconsistent data can lead to poor decisions, inefficiencies, or even safety risks—especially in sectors like healthcare, mobility, or public services.
- **Interoperability:** Data Spaces typically involve the integration of data from diverse sources and systems. High data quality—characterized by consistency, accuracy, and standardized formats—facilitates seamless interoperability and integration across platforms.
- **Compliance and Accountability:** Many Data Spaces operate in regulated environments where data is subject to legal or contractual requirements. Ensuring data quality helps participants meet compliance obligations and maintain accountability for how data is used and shared.
- **Efficiency and Automation:** Processes such as automated analytics, AI/ML applications, and real-time data exchanges depend on high-quality data. Poor data quality can lead to processing errors, increased costs, and reduced value from digital transformation initiatives.
- **Reputation and User Confidence:** Data quality issues can erode confidence in the Data Space ecosystem, leading to reluctance among participants to share or utilize data. Maintaining high data quality helps protect the reputation of the Data Space and the organizations involved.

6.3.2 DATA QUALITY DIMENSIONS

In the context of the project the data are evaluated based on a number of dimensions, which can differ based on the source of information. These dimensions are used to categorize data quality metrics:

- **Completeness**: This represents the amount of data that is usable or complete. If there is a high percentage of missing values, it may lead to a biased or misleading analysis if the data is not representative of a typical data sample.
- **Uniqueness**: This accounts for the amount of duplicate data in a dataset. For example, when reviewing customer data, you should expect that each customer has a unique customer ID.
- **Validity**: This dimension measures how much data matches the required format for any business rules. Formatting usually includes metadata, such as valid data types, ranges, patterns, and more.
- **Timeliness:** This dimension refers to the readiness of the data within an expected time frame. For example, customers expect to receive an order number immediately after they have made a purchase, and that data needs to be generated in real-time.
- **Accuracy**: This dimension refers to the correctness of the data values based on the agreed upon "source of truth." Since there can be multiple sources which report on the same metric, it's important to designate a primary data source; other data sources can be used to confirm the accuracy of the primary one. For example, tools can check to see that each data source is trending in the same direction to bolster confidence in data accuracy.
- **Consistency**: This dimension evaluates data records from two different datasets. As mentioned earlier, multiple sources can be identified to report on a single metric. Using different sources to check for consistent data trends and behaviour allows organizations to trust the any actionable insights from their analyses. This logic can also be applied around relationships between data. For example, the number of employees in a department should not exceed the total number of employees in a company.
- **Fitness for purpose**: Finally, fitness of purpose helps to ensure that the data asset meets a business need. This dimension can be difficult to evaluate, particularly with new, emerging datasets. These metrics help teams conduct data quality assessments across their organizations to evaluate how informative and useful data is for a given purpose.

7 ROLES AND RESPONSIBILITY

Along the document, it has be illustrates actors who are involved in the project as the companies who generate data and share them in the Data Space, companies which are involved in the infrastructure, models among others and third parties. For a rigours definition of the roles the D2.2 Data Space Governance must be read. This chapter focus on the role and responsibility that one of those actors must respect given their role.

- **Project Coordinator**: The project coordinator holds overall responsibility for data management across the entire project. This includes ensuring that the Data Management Plan (DMP) is implemented, updated as necessary, and adhered to by all partners. The coordinator oversees the secure storage, sharing, and long-term preservation of all project data, and ensures that all data handling is compliant with relevant legislation (such as GDPR) and project policies. The coordinator is also accountable for the availability of data shared via the project's digital platforms (e.g., website, European Portal), and for organizing regular backups and the protection of proprietary or sensitive

data.

- **Work Package (WP) Leaders**: Each WP leader is responsible for the quality control and integrity of data generated or processed within their respective work package. This includes establishing data validation procedures, ensuring that data is accurate, well-documented, and formatted according to project standards, and supporting the project coordinator in implementing the DMP. WP leaders are also responsible for ensuring that any data they manage is stored appropriately and made available for sharing in line with FAIR principles, unless legal or contractual restrictions apply.
- **Project Partners**: All project partners are required to comply with the policies set out in the DMP. Partners are responsible for the proper creation, management, and secure storage of any datasets or digital outputs they generate or process. They must ensure that data is handled in accordance with applicable legal, ethical, and security requirements, and that appropriate metadata and documentation are provided to facilitate data sharing and re-use.
- **IT and Security Teams**: Where relevant, IT and security teams will support the technical aspects of data security, including access controls, encryption, monitoring, and disaster recovery measures, in coordination with the project coordinator and WP leaders.

8 ETHICS

All data-related activities within the project are designed and **conducted in strict compliance** with **ethical standards** and the applicable national, EU, and international legislation, including the **General Data Protection Regulation** (GDPR), the **Charter of Fundamental Rights of the European Union**, and the **European Convention on Human Rights**. Informed consent will be obtained from all individuals participating in any study or data collection process, and explicit authorization will be requested for the publication of data, even in cases where the data has been anonymized to protect individual identities. Special attention is paid to safeguarding the privacy, dignity, and image rights of minors or any members of vulnerable populations, ensuring that all specific legal and ethical requirements are fully respected.

The project does not collect, process, or distribute personal data except in cases where it is absolutely necessary for research objectives and only after appropriate ethical approval has been secured from the relevant institutional review boards or ethics committees. In these rare instances, strict protocols are enforced to anonymize or pseudonymize data, thereby minimizing any potential risk of re-identification and ensuring compliance with data protection regulations. For market survey data or any information involving human participants, all datasets will be anonymized prior to analysis and dissemination, and no sensitive or identifying information will be made publicly available.

Data storage and archiving are managed using secure, access-controlled systems to prevent unauthorized access, loss, or misuse of information. Data retention policies are clearly defined, and data will only be stored for as long as necessary to fulfill the research objectives or legal requirements. Upon project completion, data that is no longer needed will be securely deleted or archived in accordance with best practices and legal obligations.

The research activities carried out in the framework of this project are primarily focused on basic research and do not involve the use of embryonic stem cells, research on animals or primates, or the direct involvement of patients or children unless specifically described and ethically approved. In light of the project's scope, the risk of ethical issues arising during data collection, processing, or archiving is considered to be extremely limited. Nevertheless, the project maintains a proactive and transparent approach to ethics, with continuous oversight and review mechanisms in place to address any unforeseen ethical challenges promptly and effectively.

8.1 ETHIC IN PERSONAL DATA

In the context of this project, strict ethical standards are applied whenever personal data collection is considered. Informed consent will always be requested from all individuals participating in studies or surveys, ensuring that participants fully understand the nature and purpose of the data collection, as well as how their data will be used, shared, and preserved over the long term. Explicit authorization will be sought not only for the collection and analysis of data but also for its publication and potential sharing with third parties, even if the data is anonymised.

All consent forms and questionnaires will clearly outline the options available to participants, including the right to withdraw consent at any stage without penalty. Where possible, the project aims to minimize the collection of personal data, and in many cases, no personal or sensitive data will be collected or distributed. If it becomes necessary to process personal data, all procedures will be fully compliant with the General Data Protection Regulation (GDPR) and relevant national and EU legislation, and will be subject to oversight by the appropriate ethics committees or data protection officers

9 REFERENCES

- European Data Strategy
- Digital Europe Programme
- Horizon Europe Data Management Plan Template.
- Research Data Management Plan Template. UCA Library.
- Guidelines on FAIR Data Management in Horizon 2020. Version 3.0.
- DS4MOBMODELS deliverable_D1.2 Project Information Repository_v2.0
- General Data Protection Regulation