Ref. Ares(2026)8293480 - 31/08/2026

FEDORA

INTEGRATED CROSS-SECTORAL DATA SPACE

Project deliverable D3.1

HORIZON Research and Innovation Action | 101203465 |
Call HORIZON-CL5-2023-D6-01

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

DELIVERABLE ADMINISTRATIVE INFORMATION

TITLE OF THE DELIVERABLE	D3.1 - Integrated Cross-Sectoral Data Space
WORK PACKAGE	WP3
TYPE OF DELIVERABLE	R — Document, report
DISSEMINATION LEVEL	PU - Public
STATUS – VERSION, DATE	Final – V3.0, 27/08/2026
DELIVERABLE LEADER	ITS-VIE
CONTRACTUAL DATE OF DELIVERY	31/08/2026
SUBMISSION DATE	31/08/2026
KEYWORDS	Data Space, Catalogue, Harmonisation, Enrichment, Synthetisation

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

LIST OF CONTRIBUTORS

Name	Organisation
Hakim Bajim, Ida Hönigmann, Stephan Krause	ITS Vienna Region, Verkehrsverbund Ost-Region (VOR) GmbH (ITS-VIE)
Renee Obregon Gonzalez, Jonathan Randriarison, Marion Hennechart, Thierry Chevallier	Akkodis I&S SAS (AKKIS)
Eider Irigoyen, Julen Macia	Fundacion Centro De Tecnologias De Interaccion Visual Y Comunicaciones Vicomtech (VICOM)

QUALITY CONTROL

	Name	Organisation	Date
Peer review	Matthias Prandtstetter	AIT Austrian Institute of Technology	24.07.2026
Peer review	Kevin Riehl, Julius Schlapbach	ETH Zürich	28.07.2026

VERSION HISTORY

Version	Date	Author	Summary of changes
1.0	1/08/2026	Stephan Krause	Initial Version for review
2.0	24/08/2026	Stephan Krause	Updated version after review
3.0	27/08/2026	Stephan Krause	Added software repositories to annex

LEGAL DISCLAIMER

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

TABLE OF CONTENTS

TABLE OF CONTENTS

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

LIST OF FIGURES

Funded by the European Union

Project funded by Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra Swiss Confederation Federal Department of Economic Affairs, Education and Research EAER State Secretariat for Education, Research and Innovation SERI

LIST OF TABLES

Funded by the European Union

Project funded by Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER State Secretariat for Education, Research and Innovation SERI

LIST OF ABBREVIATIONS AND ACRONYMS

Acronym	Meaning
AI	Artificial intelligence
AIS	Automatic Identification System
AMQP	Advanced Message Queuing Protocol
AP	Application Profile
API	Application Programming Interface
ASN.1	Abstract Syntax Notation One
C-ITS	Cooperative Intelligent Transport Systems
CKAN	Comprehensive Knowledge Archive Network
CSDGM	FGDC Content Standard for Digital Geospatial Metadata
CSV	Comma-separated Values
CSW	Catalogue Service for the Web
DCAT	Data Catalog Vocabulary
DOI	Digital Object Identifier
DSSC	Data Space Support Centre
EC	European Commission
EDC	Eclipse Dataspace Components
EMDS	European Mobility Data Space
EV	Electrical Vehicle
FAIR	Findable, Accessible, Interoperable and Reusable
FCD	Floating Car Data
FGDC	Federal Geographic Data Committee
GBFS	General Bikeshare Feed Specification
GeoJSON	Geographical features based on JSON
GeoTIFF	Geographical TIFF
GIF	Graphics Interchange Format
GIS	Geographical Information System
GLSKF	Generalized Least Squares Kernelized Tensor Factorization
GML	Geography Markup Language
GTFS	General Transit Feed Specification
GTFS-RT	GTFS Realtime
HDF5	Hierarchical Data Format, version 5
HTTP	Hypertext Transfer Protocol

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

HTTPS	Hypertext Transfer Protocol Secure
IDSA	International Data Spaces Association
IDS-RAM	International Data Spaces Reference Architecture Model
IEC	International Electrotechnical Commission
INSPIRE	Infrastructure for Spatial Information in the European Community
IoT	Internet of Things
IPR	Intellectual Property Rights
ISO	International Organization for Standardization
ITS	Intelligent Transport Systems
JPEG	Joint Photographic Experts Group
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation for Linked Data
KML	Keyhole Markup Language
LEZ	Low Emission Zone
LIDAR	Laser Imaging Detection and Ranging
LLM	Large Language Model
LRTL	Low-rank tensor learning
MaaS	Mobility as a Service
MDHESE	Mobility Data Harmonisation, Enrichment, and Synthetisation Engine
MQA	Metadata Quality Assessment
MQTT	Message Queueing Telemetry Transport
MTM	Multimodal Traffic Management
NAP	National Access Point
NAPCORE	National Access Point Coordination Organisation for Europe
NetCDF	Network Common Data Format
NeTEx	Network Timetable Exchange
OGC	Open Geospatial Consortium
OSM	Open Street Map
PDF	Portable Document Format
PNG	Portable Network Graphics
RAG	Retrieval-Augmented Generation
RDF (1)	Resource Description Framework
RDF (2)	Rich Data Format
REST	Representational State Transfer
RIS	River Information Services
RM-ODP	Reference Model of Open Distributed Processing

Funded by the European Union

Project funded by Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra Swiss Confederation Federal Department of Economic Affairs, Education and Research EAER State Secretariat for Education, Research and Innovation SERI

RMSE	Root Mean Square Error
RSS	Really Simple Syndication
SHACL	Shapes Constraint Language
SIRI	Service Interface for Real-time Information
SotA	State of the Art
SPARQL	SPARQL Protocol and RDF Query Language
SSH	Secure Shell
ST-LRTC	Spatio-Temporal constrained Low-Rank Tensor Completion
SUMO	Simulation of Urban Mobility
SVG	Scalable Vector Graphics
TBD	to be defined
TIFF	Tagged Image File Format
TNO	Nederlandse Organisatie voor toegepast-natuurwetenschappelijk onderzoek
VRU	Vulnerable Road User
W3C	World Wide Web Consortium
WFS	Web Feature Service
WMS	Web Map Service
WMTS	Web Map Tile Service
XML	Extensible Markup Language

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

PROJECT EXECUTIVE SUMMARY

Lack of orchestration, structured and standardized integration protocols and metadata descriptors, and the incorporation of real-world traffic complexities and nuances leads to suboptimal management of mobility services and infrastructure. Additional challenges, including the underutilization of valuable resources, model uncertainties and the limited integration of micro-mobility services and vulnerable road users (VRUs), further contribute to a divergence from EU's sustainable mobility targets. FEDORA aims to pave the way towards advanced traffic and network management through the development of a federated spaces platform offering a holistic framework of innovative solutions and services that enable precise and pro-active sensing of supply and demand, facilitate optimal operation of transport services, and advances learning and evolution in complex environments. At the operational level, FEDORA offers a collaborative space of data that can realize advanced data alchemy processes using interconnected services and tools, a space of advanced traffic management optimisation services and a multi-modal simulation space to create and assess future mobility scenarios. The approach is validated in six thematic demonstrations in Vienna, the Basque country, Reggio Emilia, Nicosia, Budapest and Denmark, covering varying EU urban and rural contexts, infrastructure maturity levels, multimodal mobility services availability, organisation/operational structures and social conditions. Interaction with existing programmes on road mapping and recommendations at national, EU and global level will be promoted, allowing a multiplication effect of project's results.

Social Media link:

@fedora-euproject

For further information please visit https://fedora-horizon.eu

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

DELIVERABLE EXECUTIVE SUMMARY

This document, together with deliverable D3.2 “AI-Enhanced Data Utilization and Smart Contracting” (Chevallier, et al., 2026), describes the **FEDORA Data Space**, a federated environment for multimodal mobility data sharing that enables participating organisations to discover, access, exchange, harmonize, enrich, and combine data while preserving data sovereignty.

Rather than creating a centralized repository, the FEDORA Data Space acts as an integration and collaboration layer that supports:

- Secure data exchange between project partners
- Metadata interoperability
- Data discovery and cataloguing
- AI-enhanced dataset search
- Data harmonization, enrichment, and synthetization

The data exchange capabilities of the FEDORA Data Space rely on the conceptual framework of the Data Spaces Support Centre (DSSC) (Data Space Support Centre, 2026) Blueprint, the International Data Spaces Association Reference Architecture Model (IDS-RAM) (International Data Spaces Association, 2026), and Eclipse Dataspace Connectors (EDC) (Eclipse Contributors, 2026) as the corresponding software implementation, ensuring alignment with the objectives of the European Strategy for Data (EC, A European strategy for data, 2020) and initiatives to create a European Mobility Data Space (EC, Creation of a common European mobility data space, 2023).

While these frameworks support interoperable data exchange on an organizational and technical level, semantical interoperability of data and metadata must be ensured by dedicated measures. This document proposes recommended data formats, open interface standards and metadata schemas to facilitate interoperable data exchange.

Data discovery is supported by the FEDORA Public Catalogue, a CKAN catalogue that harvests metadata from a wide range of multimodal mobility related data sources and provides not only federated search services, but also the base for AI-enhanced dataset search performed by a separate RAG (Retrieval-Augmented Generation) component.

In addition to the data exchange capabilities, the FEDORA Data Space includes the Mobility Data Harmonisation, Enrichment, and Synthetisation Engine (MDHESE), a dedicated architectural module engineered to standardise and elevate the quality of shared information. By establishing a robust data refinement pipeline, the MDHESE component ensures that output datasets are fully compliant with European data standards, providing a secure and reliable foundation for cross-platform integration and advanced analytics.

This document’s companion, deliverable D3.2 “AI-Enhanced Data Utilization and Smart Contracting” (Chevallier, et al., 2026), goes into more detail on the technical implementation of the FEDORA Data Space as well as blockchain-based Smart Contracting for secure data exchange and the Retrieval-Augmented Generation component for AI-enhanced search. Both documents describe the current state of development of the FEDORA Data Space as of Milestone 3 SPARK. They provide the foundation for further development of the Data Space components and are going to be updated by the final deliverable D3.3.

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

1 INTRODUCTION

Multimodal traffic management and planning require data from organisations that operate in different transport domains. However, these data sources are often distributed across multiple organisations and differ in format, structure, semantics, quality and access conditions. This makes it difficult to combine data consistently across multiple organisations for analysis, simulation, optimisation and the development of advanced traffic-management services.

FEDORA addresses these challenges by establishing a federated Data Space that enables participating organisations to make relevant data discoverable, and accessible under agreed technical and governance conditions. Furthermore, this allows them to retain control over their data assets. The Data Space is not intended as a centralised repository or a real-time operational control platform. Instead, it provides an integration and collaboration layer that allows heterogeneous multimodal datasets to be shared, described, harmonised, enriched and combined for the use in FEDORA services and use cases.

This deliverable, together with Deliverable D3.2 (Chevallier, et al., 2026), describes the FEDORA Data Space. The report first introduces the concepts and reference architectures of data spaces in general and then specifies the high-level use cases, requirements and structure of the FEDORA Data Space in particular. It describes the data and metadata handled by the components of the Data Space, goes into more depth on the functionality of the FEDORA Public Catalogue and explains the methodology behind the Harmonisation, Enrichment and Synthetisation component. Its companion, Deliverable D3.2 (Chevallier, et al., 2026), details the technology and implementation of the Data Space and its components.

While the FEDORA Data Space implements a subset of general data space concepts tailored to the scope and needs of the project, it also includes AI-enhanced catalogue search as well as mechanisms for harmonising, enriching, and synthetising heterogeneous mobility data. These functionalities go beyond the core concept of a conventional data space and aim to facilitate the work with distributed multimodal data.

The documents D3.1 and D3.2 describe the current state of development of the FEDORA Data Space as of Milestone 3 SPARK. They provide the foundation for further development of the Data Space components and are going to be updated by the final deliverable D3.3.

1.1 MAPPING FEDORA OUTPUTS

Deliverable D3.1 “INTEGRATED CROSS-SECTORAL DATA SPACE” covers tasks T3.1 (Standardised and interoperable future mobility data management), T3.2 (Data discovery, cataloguing and publication), and T3.3 (Dataset harmonisation, enrichment, and synthetisation for complex data exchanges).

The document is part of the first round of deliverables for the technical work packages due by the end of M15 (August 2026). It gives an overview of the concepts and components of the FEDORA Data Space at the current stage of development. It will be updated by Deliverable D3.3 by the end of the project.

The report covers one demonstrator, namely the FEDORA Public Catalogue. The data space concepts are demonstrated by the Minimum Viable Dataspace implementation described in Deliverable D3.2 (Chevallier, et al., 2026).

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

Table 1: Adherence to FEDORA GA Deliverable & Tasks Descriptions

FEDORA GA COMPONENT TITLE	FEDORA GA COMPONENT OUTLINE	RESPECTIVE DOCUMENT CHAPTER(S)	JUSTIFICATION
	DELIVERABLE		
D3.1	Integrated Cross-sectoral Data Space	Whole document	Gives an overview of the FEDORA Data Space
	TASKS		
T3.1	Defines interoperable and standardised mobility data interfaces/formats to enable cross-sectoral data exchange and support the implementation of the data space and data catalogue.	Chapter 2 Chapter 3	Chapter 2 gives a high-level overview of the Data Space that enables data exchange. Chapter 3 describes interoperable formats and interfaces for data exchange.
T3.2	Deploys and configures the FEDORA data catalogue for discovery, exchange, and federated access to static and real-time multimodal mobility datasets using interoperable metadata standards.	Chapter 3 Chapter 4	Chapter 3 describes metadata interoperability standards. Chapter 4 describes the catalogue in detail.
T3.3	Develops methods for harmonising, enriching, and synthesising heterogeneous mobility datasets to improve data quality and enable seamless complex data exchange across platforms.	Chapter 5	Chapter 5 describes the methodology for harmonisation, enrichment and synthetisation.

1.2 DELIVERABLE OVERVIEW AND REPORT STRUCTURE

The structure of both deliverables D3.1 and D3.2 is based on the Reference Model for Open Distributed Processing (RM-ODP).

RM-ODP is an ISO/IEC standard (ISO/IEC 10746) (International Organization for Standardization (ISO)) that provides a structured framework for specifying complex distributed systems through a set of complementary viewpoints. It separates concerns into five perspectives—enterprise, information, computational, engineering, and technology—allowing stakeholders to describe objectives, data semantics, functional decomposition, infrastructure, and implementation choices in a consistent and interoperable manner.

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

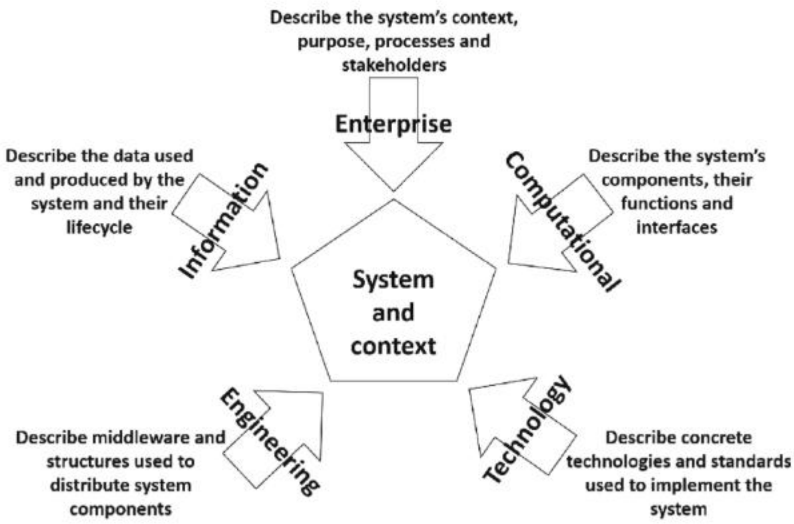

Figure 1 The five viewpoints of RM-ODP

Applying RM-ODP to the specification of the FEDORA dataspace, it provides a clear alignment between architectural design and specific requirements of the dataspace. More specifically it provides a methodological approach to organize heterogeneous actors, data flows, and services. **The enterprise viewpoint** formalizes governance, roles, and data-sharing agreements between public authorities, transport operators, and private service providers. **The information viewpoint** structures mobility data assets, including static supply data, dynamic flows, and metadata aligned with standards such as DCAT-AP. **The computational viewpoint** defines interoperable services for data exchange, discovery, and processing, which are essential for enabling e.g. multimodal journey planning, real-time analytics, dynamic pricing, network services optimisation, anomalies detection, etc. **The engineering viewpoint** addresses the distributed nature of the dataspace, ensuring secure, scalable, and sovereign data sharing across platforms. Finally, **the technology viewpoint** maps these specifications to concrete implementations, such as API, data formats, and cloud or edge infrastructures.

This document covers primarily the enterprise and information viewpoints of the FEDORA data space, whereas deliverable D3.2 (Chevallier, et al., 2026) goes into more depth on the computational, engineering and technology viewpoints.

Chapter 2 gives an overview of data spaces in general and lines out the role of the FEDORA Data Space within the project and describes its use cases and high-level design, covering the enterprise viewpoint.

Chapter 3 focuses on the information viewpoint, data and metadata, and interoperability.

Chapter 4 describes the FEDORA public catalogue being developed in Task T3.2

Chapter 5 describes the methodology for Dataset Harmonisation, Enrichment and Synthetisation being developed in Task 3.3.

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

2 DATA SPACE OVERVIEW

This chapter introduces the FEDORA Data Space, its role within the project, governance model, requirements, and key use cases. It also discusses the architectural principles adopted for its design and situates the FEDORA approach within the broader European data space landscape.

2.1 CONTEXT

The mobility sector relies on both static data (such as maps, transport schedules, infrastructure information) and dynamic data (such as traffic conditions, vehicle locations, real-time service updates). Exchange of this data is crucial to enable services, such as smarter routing, optimized public transport, efficient parking, multimodal trip planning, and support for connected and autonomous vehicles.

2.1.1 WHAT ARE DATA SPACES?

The CEN Workshop Agreement on Trusted Data Transactions (European Committee for Standardization, 2024) defines a data space as "*an interoperable framework, based on common governance principles, standards, practices and enabling services, that enables trusted data transactions between participants*". Various other definitions of data spaces exist. These share the following core principles:

- Sovereignty over data remains at the data provider
- Trust between participants
- Secure data transfer
- Interoperability and standardisation
- Services, such as onboarding, identification, data discovery, data publication, …
- Optionally additional services, such as tracking data transfers, data enrichment, …

From these core principles it follows that data spaces implement a decentralized approach, as using a central authority for data collection and distribution would violate data sovereignty. In principle, data is shared between organizations following agreed upon policies, protocols and processes. Data is advertised together with corresponding access and usage policy by data providers, while data consumers can request the data and access them directly from the data provider, given they agree to the data sharing contract. A governance authority enforces the policies and rules of the data space. Validation and verification of data space participants is based on trust framework.

Definitions of data space terms can be found in ISO/IEC DIS 20151 (International Organization for Standardization (ISO), 2015) and DSSC Blueprint Glossary (Data Spaces Support Centre, n.d.). A review of recent contributions on data spaces with focus on the European context is given in D2.1 Best Practices & State of the Art (Fafoutellis, et al., 2025).

The goal of data spaces is to encourage collaboration by creating a unified data ecosystem where organizations can share and access data efficiently. The focus is on ensuring fair, transparent and sovereign data usage, while standardized formats and protocols enable interoperability. Data spaces aim to enable secure data sharing and support data-driving innovations.

Organizations part of a data space benefit by gaining access to external data while retaining control over their own. Moving from data silos to data spaces can lower costs by avoiding duplicate efforts. Data spaces also enable faster innovation and new business opportunities through collaboration. Standardized frameworks simplify integrating external data and ensure compliance with regulations. They ensure data access and use is under

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

transparent rules and fair conditions. Data spaces support better, data-driven decision making and increase competitiveness.

2.1.2 EU INITIATIVES

The European Union has introduced several key initiatives to support the development of data spaces as part of its broader data strategy. The following short description is taken from D2.1 Best Practices & State of the Art (Fafoutellis, et al., 2025), section 3.1.3; please refer to that document for a more in-depth-discussion.

The Data Governance Act (EP, REGULATION (EU) 2022/868 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL, 2022) establishes frameworks to enhance trust, enable data reuse, and increase data availability across sectors. The Data Act (EP, REGULATION (EU) 2023/2854 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL, 2023) builds on this by clarifying user rights, ensuring fair access to data and strengthening interoperability. Together, they support the establishment of common data spaces in domains such as mobility, energy and health. In this context, the Communication on a European Strategy for Data (EC, A European strategy for data, 2020) proposes the creation of a single market for data with focus on governance frameworks, infrastructure, interoperability, portability rights, and the creation of data spaces, while the Communication on the Creation of a Common European Mobility Data Space (EMDS) (EC, Creation of a common European mobility data space, 2023) outlines a framework to connect and integrate diverse and often fragmented transport data ecosystems.

2.1.3 REFERENCE ARCHITECTURES AND IMPLEMENTATIONS

Different reference architectures provide frameworks for the development of data spaces. They discuss organizational, legal and technical aspects, detailing the structure, components, roles and interactions of the system. Software frameworks supporting data space implementations rely on these reference architectures. The most notable are listed in Table 2.

Table 2 Reference Architectures in Data Space Design. For each reference a name and short description is given. Core components and services list the names of components and services relevant to the reference architecture. The implementation column lists some reference implementations. This is a non-exhaustive list.

NAME	SHORT DESCRIPTION	CORE COMPONENTS / SERVICES	IMPLEMENTATIONS
International data space reference architecture model (IDS-RAM) developed by the International Data Spaces Association	It defines technical, organizational, and legal components to build trustworthy, sovereign, and interoperable data spaces. It provides guidance for roles, processes, functional requirements, and standards, enabling organizations to offer, discover, negotiate, and consume data while preserving participant autonomy and compliance with regulations.	• App Store • Clearing House • IDS Connector • Identity Provider • IDS Meta Data Broker	• Eclipse Dataspace Components (EDC) • Amadeus EDC Connector • Eclipse Tractus-X EDC Connector • TNO Security Gateway

Funded by the European Union
Project funded by Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER State Secretariat for Education, Research and Innovation SERI

Gaia-X developed by the Gaia-X European Association for Data and Cloud	It defines principles, standards, and modular components that enable organizations to share data and services securely while maintaining control and compliance.	• Federated Catalogue • Identity and Trust service • Data Sovereignty service • Compliance service • Gaia-X Portals and APIs	• Gaia-X Federation Services • Structura-X / Cloud Federation Middleware • Simpl
FIWARE supported by the FIWARE Foundation	It provides an open-source framework of modular components, called Generic Enablers, designed to enable interoperable, data-driven applications and platforms.	• Core Context Management • Context Processing, Analysis and Visualization • Interface with IoT, Robots and Third-Pary Systems • Context Data/API Management, Publication and Monetization • Deployment Tools • External Third-party Components	• FIWARE (provides implementation)
Big Data Value Reference Model developed by the Big Data Value Association	It provides a high-level conceptual model that outlines elements and concerns for big data systems across the data lifecycle.	• Data Visualization and User Interaction • Data Analytics • Data Processing Architectures • Data Protection • Data Management	-

2.1.4 RELATED MOBILITY DATA SPACES AND PORTALS

Besides general domain-unspecific implementations listed Table 2, there exist mobility specific data spaces. Table 3 includes both operational data spaces, projects and platforms that contribute to mobility data space deployment.

Table 3 Existing mobility-domain related data spaces and portals. For each space or portal a name and short description is given, along with the used reference architecture.

NAME	SHORT DESCRIPTION OF MAIN PURPOSE	REFERENCE ARCHITECTURE USED
EONA-X	EONA-X aims to develop a cross-sector data space to make tourism, mobility, and logistics better for everyone. Facilitate collaboration in a trusted digital infrastructure helping public and private actors get the best from data-sharing.	GAIA-X; IDSA
Mobilithek	Mobilithek is Germany's national mobility data access platform, designed to provide standardized access to mobility data for transport providers, authorities, and service developers.	IDSA
MobiSpaces	MobiSpaces is an EU Horizon Europe project developing new data spaces for green mobility that manage large volumes of mobility data	IDSA

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

	from sensors, vehicles, and infrastructure to support intelligent transport services and environmental sustainability.	
Data space traffic measurements	The Traffic measurement data spaces collect sensor-based traffic data and makes them available for analytics, traffic management, and smart-city services. This Flemish Smart Data Space created a worn agreement framework with open standards and reusable building blocks. It is an open ecosystem.	IDSA
Mobility Data Space (MDS)	The MDS is a data marketplace where partners from the mobility sector can exchange data. The goal is to create a cross-company data economy to realise and further develop innovative, environmentally and user-friendly mobility concepts. In doing so, the MDS brings together those who want to monetise their data and those who need data for mobility solutions, but in many cases also to create an exchange for mutual benefit. The MDS creates the framework for a trusting cooperation. It grew out of a project initiated by the acatech Foundation and is funded by the German Federal Ministry of Transport (BMV).	IDSA
Predictive Unit Realtime Information Service (PURIS)	PURIS aims to enrich companies' resilience strategies through standardized data sharing, giving stakeholders heightened transparency and comprehensive information. This allows PURIS users to detect supply chain issues earlier. PURIS is designed to provide real-time predictions of transport or traffic conditions based on live and historical mobility data. These services are typically used in intelligent transport systems and Mobility as a Service (MaaS) platforms. PURIS supports the supply chain across pre-, during-, and post-disruption phases, thereby improving operational efficiency and resilience within the Catena-X network.	GAIA-X; IDSA
European Mobility Data Space (EMDS)	The EMDS aims to implement a common European mobility data space by harmonizing common building blocks and working on connecting data spaces, to enable data transfer between them. It was developed by the PrepDSpace4Mobility Coordination and Support Action, (which prepared the deployment by mapping existing data ecosystems, identification of gaps and overlaps, and proposing common building blocks and governance frameworks) and the deployEMDS project (which is actively testing the deployment of EDC based mobility data spaces in various European cities and regions, with contributions such as a requirements analysis of the technical infrastructure and a technical specifications report).	GAIA-X; IDSA

For data spaces from all sectors the International Data Spaces Association (IDSA) provides the data space radar, giving an overview of data spaces in different development stages as well as use cases.

While Table 3 focuses on operational data spaces, deployment initiatives, and platforms that implement or actively contribute to mobility data spaces, the following examples are not data spaces themselves. Instead, they provide supporting frameworks, methodologies, standards, governance guidance, interoperability recommendations, and implementation resources that facilitate the development and operation of data spaces.

- The Data Space Support Centre offers definitions of data space related terms, methods for stakeholder alignment, building blocks for implementing data spaces, service examples and references to

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

implemented data space tools. For details on relation between these resources and the FEDORA data space see sections 2.2 and 2.4.

- NAPCORE coordinates and harmonizes national access points across Europe. These efforts include work on the DATEX II standard for road traffic and travel information (DATEX II, 2026).
- Simpl provides a common, open-source software solution for data access and interoperability among European data initiatives. It uses existing components such as the EDC connector, ensuring compatibility with other implementations.
- The International Data Spaces Association offers a Knowledge Base defining key principles and laying out requirements and models. A more detailed discussion of how these models play into the FEDORA data space implementation is given in Section 2.2.

2.2 THE FEDORA DATA SPACE

The FEDORA Data Space, as laid out in the Spaces Definition of D2.2 Living Labs Setup, Spaces Definition, and Evaluation Models (Sáenz de Ugarte, Cortés, & Cheliotis, 2026), provides the framework through which all data sharing within FEDORA will take place. It aims to facilitate secure and trusted data sharing within the FEDORA platform, making data from external sources available to the partners and providing harmonized and enriched data sets.

When want to reiterate here that whereas on the one hand, the FEDORA Data Space implements a subset of general data space concepts tailored to the scope and needs of the project, on the other hand, it provides additional features for AI-enhanced catalogue search and Harmonisation, Enrichment and Synthetisation that go way beyond the core functionalities defined in the reference architectures.

The basic components of the FEDORA Data Space are a public catalogue including a Retrieval-Augmented Generation (RAG) search system, connectors, smart data contracts modules and mobility data harmonization, enrichment and synthetisation engines. While metadata discovery is centralized through the public catalogue, data exchange itself remains decentralized and takes place directly between participants via their connectors.

Before going into more detail on its role, goals, governance, use cases and requirements in the subsequent sections, in the next section we briefly introduce the reference architecture, the conceptual framework on which the discussion and the implementation of the FEDORA Data Space is based.

2.2.1 REFERENCE ARCHITECTURE

The FEDORA Data Space follows the International Data Spaces Reference Architecture Model (IDS-RAM), version 4 (International Data Spaces Association, 2025). The reference architecture has been selected based on the goals and objectives of the FEDORA Data Space and the assessment of data space implementations by the deployEMDS project (Bozzi, De Cock, Van Gheluwe, & Pretzsch, 2025) which is discussed in more depth in Deliverable D3.2 AI-Enhanced Data Utilization and Smart Contracting (Chevallier, et al., 2026). The selection criteria were based on maturity of technical development, support for data sovereignty, connector-based exchange, trust management and alignment with European mobility data space initiatives. Among the reviewed alternatives, IDS-RAM provided the best match for the FEDORA requirements, particularly through its mature connector-based architecture and its widespread adoption within emerging European mobility data space initiatives.

The IDS-RAM Introduction states:

The IDS aims at meeting the following strategic requirements:

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

- ***Trust***: *Trust is the basis of the International Data Spaces. Each participant is evaluated and certified before being granted access to the trusted business ecosystem.*
- ***Security and data sovereignty***: *All components of the International Data Spaces rely on state-of-the-art security measures. Apart from architectural specifications, security is mainly ensured by the evaluation and certification of each technical component used in the International Data Spaces. In line with the central aspect of ensuring data sovereignty, a data owner in the International Data Spaces attaches usage restriction information to their data before it is transferred to a data consumer. To use the data, the data consumer must fully accept the data owner's usage policy.*
- ***Ecosystem of data***: *The architecture of the International Data Spaces does not require central data storage capabilities. Instead, it pursues the idea of decentralization of data storage, which means that data physically remains with the respective data owner until it is transferred to a trusted party. This approach requires a comprehensive description of each data source and the value and usability of data for other companies, combined with the ability to integrate domain-specific data vocabularies. In addition, Metadata Brokers in the ecosystem provide services for real-time data search.*
- ***Standardized interoperability***: *The International Data Spaces Connector, being a central component of the architecture, is implemented in different variants and can be acquired from different vendors. Nevertheless, each Connector is able to communicate with any other Connector (or other technical component) in the ecosystem of the International Data Space.*
- ***Value adding apps:*** *The International Data Spaces allows to inject apps into the IDS Connectors in order to provide services on top of data exchange processes. This includes services for data processing, data format alignment, and data exchange protocols, for example. Furthermore, data analytics services can be provided by remote execution of algorithms.*
- ***Data markets***: *The International Data Space enables the creation of novel, data-driven services that make use of data apps. It also fosters new business models for these services by providing clearing mechanisms and billing functions, and by creating domain-specific metadata broker solutions and marketplaces. In addition, the International Data Spaces provides templates and other methodological support for participants to use when specifying usage restriction information and requesting legal information.*

The IDS-RAM provides a generic conceptual framework for the development of data spaces. These concepts will be tailored to the role of the FEDORA data space within the project, its goals and objectives, its uses cases and requirements as laid out in the following subsections.

2.2.2 ROLE OF THE DATA SPACE WITHIN THE PROJECT

The Spaces Definition of D2.2 (Sáenz de Ugarte, Cortés, & Cheliotis, 2026) describes the role of the Data Space as follows:

> *The Data Space provides the framework through which all data sharing within FEDORA will take place. The workflow of the FEDORA Data Space is to first receive datasets from relevant trusted sources; secondly to enrich and harmonize all data sets so that they are available at the FEDORA standards; and finally, to make them available to the various FEDORA solutions in the appropriate data formats.*
>
> *The Data Space is the Space within the FEDORA system that connects most with systems external to FEDORA, identified mainly as the various data sources that feed into the various FEDORA solutions. Additionally, it provides traffic and other relevant data to the Network Management Services Space to be used for analysis and optimization strategy calculation, and it feeds data (geospatial, traffic, and others as needed) to the Simulations Space, to be used for the set up and calibration of simulation tools and models.*

The core functionality of each space developed within FEDORA is described in Figure 2.

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

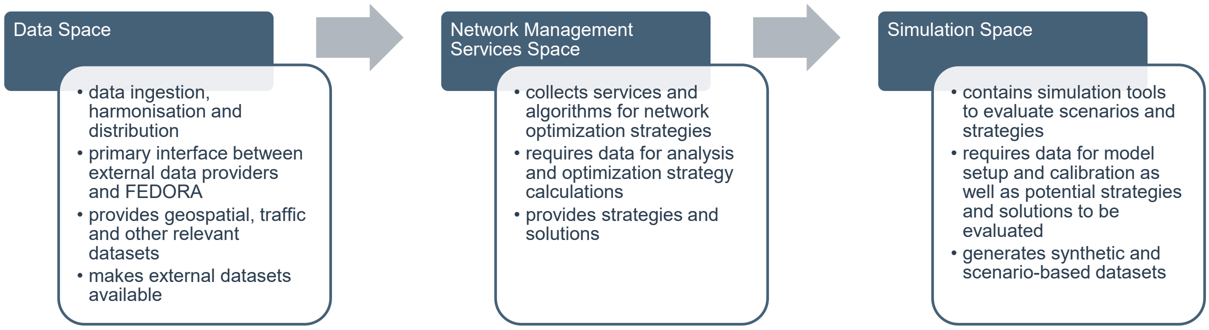

Figure 2 Interaction of Data Space with other FEDORA Spaces. The Data Space provides input data for the Network Management Services Space and Simulation Space.

According to deliverable D2.2 (Sáenz de Ugarte, Cortés, & Cheliotis, 2026) the flow of data within the FEDORA spaces architecture is primarily from the Data Space to the other spaces and further to the pilots, as seen in Figure 2. This does not prevent the project partners from using the FEDORA Data Space for sharing data between them regardless of the conceptual boundaries between FEDORA spaces and pilots, though.

At the current stage of development, it is not foreseen that data generated within the project should be made available to external stakeholders using the Data Space, because this data is often very specific to the pilot setting and therefore cannot be easily used for other purposes. In case the pilots require it, this will be revisited in the next stages of development in alignment with the recommendations of European flagship initiatives (European Mobility Data Space, deployEMDS project), see also section 2.2.4 on constraints and section 2.2.6 on governance.

The main role of the Data Space is thus:

- Providing external data to the two other FEDORA spaces and the pilots
- Sharing existing data between partners
- Enabling harmonisation, enrichment and synthetisation of data

Data sharing – the main role of the FEDORA Data Space – is the core function of a data space in the sense of the IDS-RAM and other reference architectures. But the data space idea as laid out in these reference models goes further by adding trust relationships and data sovereignty as core design principles. FAIR principles (Findability, Accessibility, Interoperability, Reusability) are further aspects guiding the design and development of the FEDORA Data Space. These are further discussed in Section 2.3.1. On the other hand, the FEDORA Data Space adds functionalities that are not required by the reference architecture, especially Harmonisation, Enrichment and Synthetisation of data. In terms of the IDS-RAM, these could be seen as Value Adding Apps.

2.2.3 GOALS AND OBJECTIVES

This section outlines the key goals and objectives of the FEDORA data space implementation. It describes core functionalities, principles and expected capabilities the data space should provide. The following list summarizes the primary objectives identified so far:

O1: Provide a Public Catalogue: The publicly accessible metadata catalogue shall act as an entry point to the data space. It includes selected datasets (e.g. publicly available datasets as well as from FEDORA-specific sources) to support discovery and attract potential participants, who can join the data space to request access to the underlying data.

O2: Ensure Secure Data Transfer: Ensure that non-public data is exchanged securely between participants. This includes reliable cross-organisational authentication and authorisation mechanisms, as well as protected data flows (e.g. encryption) during transfer.

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

O3: **Use Standardised Data Formats**: The use of common formats from the mobility domain as well as common general-purpose formats should be encouraged. Interoperability will be promoted by supporting common data formats and standards. Less common formats will be allowed if the alternative is a data exchange solution outside of the data space.

O4: **Enable Interoperability to other Data Spaces**: A data space connector should enable integration with external data spaces and systems, allowing access to data across domains and between different data space instances.

O5: **Support Identity and Access Management**: Provide a comprehensive system for managing identities, roles, and permissions across organizations. This must include structured processes for integrating and removing participants describing the technical and organizational steps.

O6: **Enable Contract and Policy Management**: Supports machine readable contracts, contract lifecycle management and enforcement. The data space must also support systematic logging of data access and exchanges to ensure traceability, accountability, and compliance.

O7: **Facilitate data harmonisation, enrichment and synthetisation**: Provides and integrates data processing functionality, including transforming datasets to different formats and standards, enriching it and quality assurance via validation mechanisms.

2.2.4 CONSTRAINTS

This section specifies constraints of the FEDORA Data Space implementation at the current stage of development. These are due to organizational and technical limitations and may be revisited for the final version considering the needs of the pilot sites.

In the first version, the FEDORA Data Space **does not offer** a streaming solution for **real-time data exchange**, as this is not part of the Minimum Viable Data Space concept and not supported off-the-shelf by EDC, the software framework used for the implementation of the FEDORA Connector. Stream processing is, however, investigated within the scope of the project (Task 3.6, see D3.2 (Chevallier, et al., 2026)).

As stated in section 2.2.2 on the role of the FEDORA Data Space, it is **not foreseen** that the data exchange capabilities should be used **to share data with external users** at the current stage. We want to emphasize here that this relates only to the functionality implemented by the FEDORA Dataspace Connectors (see section 2.4.1 on components). This constraint does neither restrict the use of the FEDORA Public Catalogue by project partners and external users to advertise and find public data offerings, nor does it restrict sharing data generated using value-adding features like Harmonisation, Synthetisation and Enrichment by other means (as long as regulations and third-party rights are respected).

There are two reasons for this restriction: first, as stated above, the data produced for the pilots is usually very specific to the sites and use cases of the pilots and may not be usable at all outside the original context. Second, ensuring trust between data space participants beyond the project partners requires a distinct governance structure which is not yet in place (see also section 2.2.6 on governance).

2.2.5 STAKEHOLDERS

The FEDORA Data Space will bring together a diverse range of actors across the multimodal mobility ecosystem. This stakeholder mapping aims to identify the main categories of actors involved, clarify their roles, and outline their relationships within the data space environment.

Interaction with the FEDORA Data Space will span both project-specific and broader ecosystem stakeholders. The DSSC blueprint introduces the following stakeholder categories: data holder, data space governance authority, data space operator and customers. In FEDORA, most of these roles are performed by the project partners themselves, see also Section 2.2.6 on Governance.

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

The following groups of FEDORA Data Space stakeholders have been identified:

- **FEDORA project partners**: Partners can act as pilot site leaders, providers of MTM solutions, and simulation providers; they can act as data providers or consumers or both; they can use the catalogue and the RAG service to find datasets; they can publish their own datasets within the FEDORA Data Space allowing other users to access them; they can use the Harmonisation, Enrichment and Synthetisation Engine to enhance their data
- **FEDORA Catalogue provider**: Provides and runs the catalogue services
- **FEDORA RAG provider**: Provide the RAG databases and interfaces
- **External data providers**: A wide range of organisations, such as authorities, mobility service providers, public transport and motorway operators publish their datasets and metadata on catalogues and open data portals (e.g. on NAPs); they are not involved in the FEDORA project, but their public offerings can be accessed by project participants, and their metadata records can be ingested into the FEDORA Catalogue
- **External catalogue providers**: Provide interfaces to harvest their metadata records; they are not involved in the FEDORA project, but their public interfaces can be used by the FEDORA Public Catalogue

2.2.6 GOVERNANCE

This section deals with the governance for the data exchange capabilities provided by the FEDORA Data Space. For the sake of clarity, we distinguish here between the *Public Layer* (the FEDORA Public Catalogue) and the *Private Layer* (the data sharing functionalities according to the more general data space concept, implemented by the FEDORA Dataspace Connector).

As the name suggests, the search features of the FEDORA Public Catalogue are available to the public without restrictions. The FEDORA Catalogue provider is responsible for configuring the data sources the catalogue harvests and presents, see chapter 4.

The preliminary governance rules for the Private Layer are described according to the DSSC Blueprint (Data Spaces Support Centre, n.d.) in the following paragraphs. The Blueprint defines two building blocks for data space governance:

- The Organisational Form and Governance Authority and,
- Participation Management

Being strictly project related, data exchange in the FEDORA Data Space Private Layer is governed by agreements between the consortium members of the FEDORA project. There is no specific legal entity acting as a Governance Authority (**unincorporated data space**). The regulations of the Grant and Consortiums Agreements regarding intellectual property rights, background, etc. shall apply.

During the FEDORA project, governance decisions regarding participant onboarding, metadata quality requirements, access policies, and dispute resolution are coordinated through the consortium governance structure, with WP3 providing operational coordination for Data Space related activities. The consortium authorizes Akkodis (WP3 leader) to operate the Trust/Issuing Service and issue Verifiable Credentials for the FEDORA Data Space.

As for **participation management**, the following rules have been defined:

- Every FEDORA project partner is a member of the FEDORA Data Space Private Layer.
- Participation in the FEDORA Data Space Private Layer is restricted to project partners.
- External stakeholders and the public shall only have access to the Public Layer.

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

- During the development phase, the data sharing capabilities of the FEDORA Data Space shall be accessible for participants, i.e. project partners, only
- Every FEDORA project partner should host their own data space connector; technical project partners may host connectors on behalf of non-technical partners.
- Project partners should use the FEDORA Data Space Private Layer for sharing data with other project partners.
- During the development phase, the FEDORA Data Space Private Layer shall be used for project related data exchange only.

The governance rules may be revisited in later stages of development and may be updated according to the needs of the pilot sites and following the recommendations of European flagship initiatives (EMDS, deployEMDS project).

2.3 NON-FUNCTIONAL REQUIREMENTS

Non-Functional requirements define qualities of the overall system, rather than functionality and behaviour of distinct components. Whereas functional requirements guide the design of the system and components, non-functional requirements guide the system architecture. The next two sections present requirements related to data exchange within the FEDORA Data Space: FAIR principles and requirements for data sovereignty.

2.3.1 FAIR PRINCIPLES

The FAIR Principles (GO FAIR, n.d.) provide guidelines to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of digital assets. The principles aim to ultimately improve the (re-)usability especially of scientific and government data sets. They define requirements for data, metadata, and infrastructure.

There is a certain tension between the FAIR principles of accessibility and reusability on one side and the idea of data sovereignty that is core to data spaces on the other side. The FAIR principles state clearly though, that the conditions of access and use (e.g. contacts and licenses) should be published in a standardized way while recognizing that data sets may be protected or too sensitive to be shared openly. A list of all FAIR principles is given in Table 4.

Table 4 FAIR principles and requirements as defined by GoFAIR (GO FAIR, n.d.).

PRINCIPLE	DESCRIPTION	REQUIREMENTS
Findability	These requirements intend to make it easy to find metadata and data for both humans and computers. They focus on machine-readable metadata.	*F1. (Meta)data are assigned a globally unique and persistent identifier.* *F2. Data are described with rich metadata.* *F3. Metadata clearly and explicitly include the identifier of the data they describe.* *F4. (Meta)data are registered or indexed in a searchable resource.*
Accessibility	These requirements deal with the description of access conditions; they do not impose any legal requirements for data access, or as GO FAIR (GO FAIR, n.d.) puts it: *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.*	*A1. (Meta)data are retrievable by their identifier using a standardised communications protocol.* *A1.1. The protocol is open, free, and universally implementable.* *A1.2. The protocol allows for an authentication and authorisation procedure, where necessary.*

Funded by the European Union

Project funded by Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER State Secretariat for Education, Research and Innovation SERI

		A2. Metadata are accessible, even when the data are no longer available.
Interoperability	These requirements focus on semantic interoperability of (meta-)data rather than structural interoperability of interfaces or organisational interoperability.	*I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.* *I2. (Meta)data use vocabularies that follow FAIR principles.* *I3. (Meta)data include qualified references to other (meta)data.*
Reusability	These requirements relate to community standards and the description of usage conditions.	*R1. (Meta)data are richly described with a plurality of accurate and relevant attributes.* *R1.1. (Meta)data are released with a clear and accessible data usage license.* *R1.2. (Meta)data are associated with detailed provenance.* *R1.3. (Meta)data meet domain-relevant community standards.*

2.3.2 REQUIREMENTS FOR DATA SOVEREIGNTY

Security

Data sovereignty requires that data be exchanged in a secure way:

- Secure data access and contract negotiation require secure, i.e. encrypted, communication
- Data access to restricted datasets (assets) requires authentication and authorization
- Secure authentication requires trusted identities (provided by a commonly trusted Identity Provider)
- Secure data transfer requires proofs of authorization (authorization tokens)

Access restrictions

Data sovereignty requires that the data provider must be able to define and enforce restrictions for data access. To this end, the data space architecture defines policies and contracts.

- The data space implementation must make sure that data transfer is only possible using a valid contract
- The data space implementation must make sure the conditions (policies) defined for a given asset are met for each requested data transfer

Observability and Traceability

Data sovereignty requires that data transfers be observable and traceable:

- Transfers must be recorded
- The records must be accessible to the partners involved
- The implementation must ensure the integrity of the records

2.4 COMPONENTS AND BUILDING BLOCKS

This section gives a short, high-level overview of the architecture and the components that make up the FEDORA Data Space. It intends to make it easier to understand the use case descriptions in section 2.5 which refer to

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

these components and show how users can interact with them. For more detailed technical descriptions of the components please refer to Deliverable D3.2 (Chevallier, et al., 2026) and chapters 4 and 5 about the FEDORA Public Catalogue and Harmonisation, Enrichment and Synthetisation, respectively.

2.4.1 FEDORA DATA SPACE COMPONENTS

The FEDORA data space is composed of several core components, each fulfilling a specific role within the data space architecture. Detailed component-level descriptions can be found in D2.2 Living Labs Setup, Spaces Definition, and Evaluation Models (Sáenz de Ugarte, Cortés, & Cheliotis, 2026); the focus of this section is on how they integrate into a unified system. A short overview of the FEDORA Data Space components is given in Table 5.

Table 5 List of core FEDORA data space components with short description.

DATA SPACE COMPONENT	SHORT DESCRIPTION
FEDORA Public Catalogue	A publicly accessible catalogue that lists data available within the data space, without disclosing the providers. It serves as a central component, providing a single, publicly visible point of access and is intended to attract potential participants to the dataspace.
RAG Search System	It provides a natural language conversational interface, extending the CKAN dataset metadata search. Metadata relevant to the search query is retrieved and ranked using a hybrid approach (semantic, keyword search, and knowledge graph, vector database).
Dataspace Connector	It performs all key operations related to data sharing between participants of the data space, such as publishing of metadata and offerings to the public catalogue; discovery of those offerings; negotiation; enforcement of use policies; and the secure transfer of data assets between participants.
Smart Data Contracts Module	A decentralized blockchain system for keeping records of all data transactions taking place in the FEDORA Data Space. It provides data space participants with a method to verify transactions and agreements. All data transactions will automatically be recorded in an immutable format and can be viewed in a user-friendly interface.
Mobility Data Harmonisation, Enrichment & Synthetisation Engine	It is responsible for transforming heterogeneous datasets into harmonized (e.g. normalisation of formats, resolution of ambiguities), enriched (e.g. integration of contextual metadata, linking with external datasets, augmentation of missing fields), and synthetically augmented (e.g. creation of realistic synthetic mobility data) data products.

Some of these components (as the FEDORA Public Catalogue and RAG Search System) will be deployed centrally, while others need to be deployed for every data space participant. Figure 3 gives an overview of the deployment plans.

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

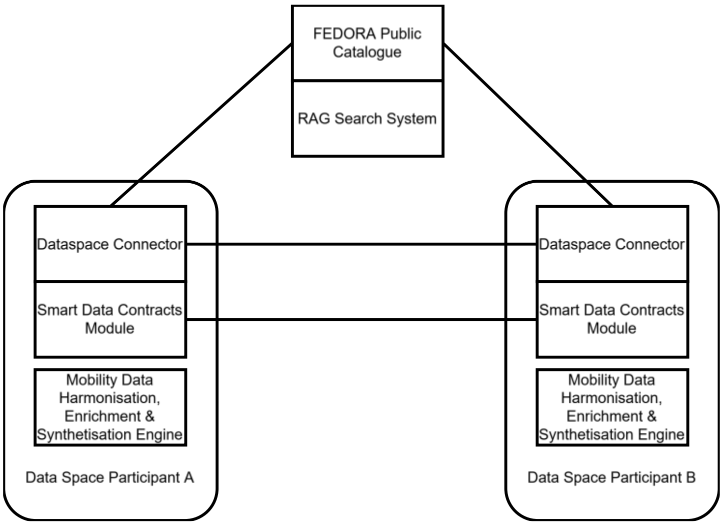

Figure 3: Deployment plans of the FEDORA data space with two example participants. The FEDORA Public Catalogue and RAG Search System are deployed centrally, while the Dataspace Connector, Smart Data Contracts Module and Mobility Data Harmonisation, Enrichment & Synthetisation Engine are deployed decentralized, once for every participant.

The components of the data space interact through interfaces exchanging data and metadata, as detailed in Table 6.

Table 6 Short overview of interaction between the data space components. For each connection the exchanged data and a short description is given.

CONNECTION	(META-)DATA EXCHANGED	SHORT DESCRIPTION
FEDORA Public Catalogue <-> Dataspace Connector	Meta data on data and offerings	Synchronization of data offers, ensuring consistency between source connectors and public catalogue.
Dataspace Connector <-> Dataspace Connector	Data	Secure transfer of data assets between participants.
Smart Data Contracts Module <-> Smart Data Contracts Module	Meta data on data transactions	Synchronization of records related to data transactions between participants, ensuring no tampering may take place.

2.4.2 IMPLEMENTATION GUIDELINES

The DSSC and IDSA both have resources available which support the creation and implementation of data spaces. The IDSA Rulebook (International Data Spaces Association, 2026) describes fundamental concepts, lists requirements and optional practices in the technical, commercial and legal dimensions of data spaces. The DSSC Blueprint (Data Spaces Support Centre, n.d.) offers business, organizational and technical best practices, alongside a co creation method, definitions and examples of services and a toolbox containing data space component implementations.

For the FEDORA data space implementation, some of the IDSA Rulebook sections and some of the DSSC blueprint blocks map to technical components deployed as part of the data space. The mapping can be found in Table 7.

Table 7 Mapping of FEDORA data space components and IDSA Rulebook and DSSC Blueprint sections. For each component the relevant sections in the IDSA Rulebook and DSSC Blueprint are given.

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

FEDORA DATA SPACE TECHNICAL COMPONENT	IDSA RULEBOOK SECTION	DSSC BLUEPRINT SECTION
FEDORA Public Catalogue	Marketplaces, Vocabularies, Data Discovery Services	Data, Services, and Offerings Descriptions; Publication and Discovery
Mobility Data Harmonisation, Enrichment & Synthetisation Engine	Processing Services	Value creation services
RAG Search System	AI Agents	Value creation services
Smart Data Contracts Module	Marketplaces, Observability, Data Sharing	Contractual Framework; Provenance, Traceability & Observability; Access & Usage Policies Enforcement
Dataspace Connector	Connector, Data sharing, Planes	How a Data Plane and Control Plane Work Together, Data Exchange

2.5 DATA SPACE USE CASES

The use cases identified for the FEDORA data space can be broadly separated into two categories: use cases related to data publication, processing and retrieval, and use cases related to user and contract management. The first category focuses on the core functionality associated with the data lifecycle, from making data available to processing and consuming it, while the second category addresses the governance, organizational and trust-related aspects required to enable secure and compliant data sharing.

For data publication, processing and retrieval, the use cases are classified according to data ownership and access:

- Data from external providers
- Input data owned by project partners to be shared with other project partners

Use cases for the different types of data differ with respect to governance and the components involved. As stated in Section 2.2.2, the FEDORA Data Space will not be used to publish data to external stakeholders; therefore, data publication to and data access by external stakeholders is not yet considered.

2.5.1 FINDING DATA FROM EXTERNAL SOURCES

Overview

This use case is intended to facilitate finding open data from external sources. The FEDORA Public Catalogue component harvests external catalogues and provides interfaces to search this data. The FEDORA RAG component uses catalogue interfaces to provide enhanced and more comfortable search options. The catalogue provides metadata on how to access the external data. As the data exchange capabilities of the FEDORA Data Space are restricted to project partners for the time being, these mechanisms cannot be used to access external data sources.

Funded by the European Union

Project funded by Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra Swiss Confederation Federal Department of Economic Affairs, Education and Research EAER State Secretariat for Education, Research and Innovation SERI

Stakeholders

- **External data providers**: Publish metadata records to external catalogues; they are not involved in the FEDORA project
- **External catalogue providers**: Provide interfaces to harvest their metadata records; they are not involved in the FEDORA project
- **FEDORA Catalogue provider**: Provides and runs the catalogue services
- **FEDORA RAG provider**: Provide the RAG databases and interfaces
- **FEDORA project partners**: Can access the catalogue and the RAG service

Supporting Components

- **FEDORA Public Catalogue**: harvests the external sources and provides interface to search the metadata records, see Chapter 4
- **FEDORA RAG Service**: uses the interfaces of the FEDORA Public Catalogue to provide AI enhanced search capabilities, see Deliverable D3.2 (Chevallier, et al., 2026)

2.5.2 ACCESSING DATA PROVIDED BY EXTERNAL SOURCES

Overview

Once the users have found a dataset they need, using the FEDORA Public Catalogue and/or the RAG service, it can be accessed using the method described in the metadata. The catalogue only acts as an intermediary providing information on how to access the data provided by external sources, but does not provide data access itself.

Stakeholders

- **External data providers**: provide data access; publish metadata records to external catalogues; they are not involved in the FEDORA project
- **External catalogue providers**: provide interfaces to harvest their metadata records; they are not involved in the FEDORA project
- **FEDORA Catalogue provider**: provides and runs the catalogue services
- **FEDORA RAG provider**: provide the RAG databases and interfaces
- **FEDORA project partners**: can access the catalogue and the RAG service

Supporting Components

- **FEDORA Public Catalogue**: harvests the external sources and provides interface to search the metadata records, see Chapter 4
- **FEDORA RAG Service**: uses the interfaces of the FEDORA Public Catalogue to provide AI enhanced search capabilities, see Deliverable D3.2 (Chevallier, et al., 2026)

2.5.3 PUBLISHING DATA TO PROJECT PARTNERS

Overview

Data to be shared with project partners can be published using the FEDORA Data Space Connector. A project partner who provides access to datasets needs to set up a provider connector. Using this component he/she can publish the asset (the dataset itself), the policies (rules and conditions for data access) and the contracts (linking policies to assets).

Funded by the European Union

Project funded by Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER State Secretariat for Education, Research and Innovation SERI

This use case is intended for (existing) input data provided to other project partners. For the time being, publication of results to external stakeholders is not part of this use case.

The Minimum Viable Data Space concept implemented in the first stage of development does not include a central repository of endpoints and metadata (Metadata Broker). Participants of the MVD communicate peer-to-peer and exchange their endpoint URLs by some other means, e.g. by email. Obviously, this approach does not scale up to larger numbers of participants. FEDORA aims to integrate the Public Catalogue with the Data Space, though (see also Section 2.4 and D3.2 Section 5.3), providing a mechanism for metadata publication and retrieval that can be used by project partners to connect with their peers. This Public Catalogue will be deployed centrally and will be accessible by the public. A private metadata broker might be needed if the metadata itself is sensitive and therefore cannot be shared with the public. After metadata discovery the data exchange is performed peer to peer.

Stakeholders

- **FEDORA project partners (provider side)**: publish assets, policies and contracts
- **FEDORA project partners (consumer side)**: can access these datasets according to the conditions of the contracts

Supporting Components

- **FEDORA Connector:** provides functionality to publish and access data

2.5.4 FINDING DATA PUBLISHED BY PROJECT PARTNERS

Overview

Project partners can mutually access their data offerings using the Data Space connectors. Each connector publishes its own *Data Catalog* (not to be confused with the FEDORA Public Catalogue) that describes the datasets that can be obtained from the connector and the corresponding access conditions based on the assets, policies and contracts defined by the provider (see section 2.5.3 on publishing data). Again, this approach does not scale up to larger numbers of participants. The integration of the FEDORA Public Catalogue with the Data Space allows the project partners to benefit from the advantages of the AI enhanced search facilities of the Catalogue and RAG systems.

Stakeholders

- **FEDORA project partners (provider side)**: publish assets, policies and contracts
- **FEDORA project partners (consumer side)**: can search for data connecting either to partners' connectors or using the Catalogue

Supporting Components

- **FEDORA Connector**: provides interfaces presenting information about its data offerings
- **FEDORA Public Catalogue**: harvests the external sources and provides interface to search the metadata records, see Chapter 4
- **FEDORA RAG Service**: uses the interfaces of the FEDORA Public Catalogue to provide AI enhanced search capabilities, see D3.2 (Chevallier, et al., 2026).

2.5.5 ACCESSING DATA PUBLISHED BY PROJECT PARTNERS

Overview

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

Once they have found the data they need, project partners can access the offerings of other project partners using their FEDORA Connectors. Data consumers must authenticate and obtain an authorization token for data access. The contracts linking assets to policies and conditions that the access request must fulfil are published by the data provider. The provider connector grants or denies authorization based on an evaluation of the request context against these rules.

Using the authorization token data can be retrieved from an API implemented by the provider. APIs can be very simple (e.g. HTTPS Download) or more elaborate (e.g. REST APIs).

Stakeholders

- **FEDORA project partners (provider side)**: publish assets, policies and contracts
- **FEDORA project partners (consumer side)**: can search for data connecting either to partners' connectors or using the Catalogue

Supporting Components

- **FEDORA Connector**: provides interfaces for contract negotiation and data retrieval

2.5.6 USING HARMONISATION, ENRICHMENT AND SYNTHETISATION

Overview

Harmonisation, Enrichment and Synthetisation include a wide range of value adding methods that can be applied to data sets to improve data quality and usability (see Chapter 5).

Typical tasks include:

- Dataset analysis to identify possible problems like deviations from the data schema
- Correction of character set encoding errors
- Duplicate removal
- Harmonisation and generalisation of temporal, spatial and spatio-temporal resolution
- Generation of contextual data
- Enrichment with metadata
- Imputation of missing data
- Generation of representative test datasets

Data harmonisation, enrichment and synthetisation can be performed at the provider side (data cleaning, anonymisation, data aggregation, etc.) or at the consumer side (e.g. contextual enrichment, data fusion).

Stakeholders

- **FEDORA project partners**: can use the Harmonisation, Enrichment and Synthetisation Engine to enhance their data

Supporting Components

- **MDHESE (Modular Data Harmonisation, Enrichment and Synthetisation Engine)**: provides functionality for harmonisation, enrichment and synthetisation, see chapter 5.

Funded by the European Union

Project funded by Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER State Secretariat for Education, Research and Innovation SERI

2.5.7 ONBOARDING DATASPACE PARTICIPANTS

Overview

Only FEDORA project partners are eligible to participate in the FEDORA Data Space. While all project partners are considered participants by default, the required technical components must be deployed and configured by each participant before they can exchange data. To support this process, setup guides, tutorials and technical documentation are provided to all project partners. Where required technical partners may deploy and configure the connector on behalf of another project partner.

Stakeholders

- **FEDORA project partners**: receive setup documentation and deploy the required technical components to participate in data sharing

Supporting Components

- **FEDORA Connector**: provides the functionality required for secure data exchange between project partners
- **Identity Provider**: provides authentication and identity management services
- **Smart Data Contracts Module**: provides additional traceability and auditability for data exchange

2.5.8 OFFBOARDING

Overview

No participant offboarding is foreseen during the lifetime of the FEDORA project, as project partners remain members of the consortium for its entire duration. Procedures for participant offboarding after the completion of the project have not yet been specified.

2.5.9 TRACING TRANSFERS USING SMART CONTRACTS

Overview

Data transfers within the FEDORA Data Space are automatically recorded by the Smart Contracts Module. To support transparency, accountability and traceability, Data Space participants can inspect records concerning them through a dedicated interface.

Stakeholders

- **FEDORA project partners**: can inspect transaction records to verify the origin, exchanged asset, applicable policies and other relevant information associated with each data transaction

Supporting Components

- **FEDORA Connector**: records the relevant information in the Smart Contracts Module for each data transfer
- **Smart Data Contracts Module**: stores transaction records and provides an interface for searching and displaying the information

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

3 DATA, METADATA AND INTEROPERABILITY

The FEDORA Data Space relies on a wide range of mobility-related datasets originating from various providers, pilot sites, domains and systems. To enable productive data exchange, a common understanding of data structures, descriptions, and interfaces is essential. In this context, interfaces describe how systems communicate and exchange data, formats define how the data itself is structured and encoded, and metadata provides descriptive information about the data, such as its content, scope, provenance, and access conditions. This chapter provides distinctions of content types of data typically handled by the data space, introduces relevant metadata standards and outlines relevant aspects on choosing appropriate data formats and standards. It follows with an overview of interoperable interfaces and closes with the inventory of data sets used in the FEDORA data space.

3.1 DATA TYPOLOGY

Mobility data encompasses a wide range of information types, reflecting the complexity of modern transportation systems. These datasets differ not only in what they describe but also in how they behave over time, ranging from static reference to rapidly changing real-time information. To bring structure to this diversity, mobility data can be organized along two complementary dimensions: content-based categories and temporal characteristics. With regards to the temporal characteristics, datasets can be classified according to their expected frequency of change. Historic or predictive datasets remain unchanged, as they refer either to past observations or future projections. Static data are updated only infrequently, typically once per month or year. Dynamic data are subject to more frequent and reoccurring updates, ranging from daily to monthly changes. Finally, real-time data are continuously refreshed at very short intervals, typically from every few minutes down to sub-second updates.

Table 8 gives examples of datasets following this classification.

Table 8 Data typology in content and temporal axes with example data categories.

	HISTORIC / PREDICTIVE	STATIC	DYNAMIC	REAL-TIME
INFRASTRUCTURE & SPATIAL DATA	• geospatial statistics	• infrastructure • network graph • stations • topography		
MOBILITY DEMAND & TRAVEL BEHAVIOR	• activity and travel diaries • travel surveys • bike, scooter usage	• origin destination flows • traffic model		• FCD • object lists of traffic participants • real-time locations • traffic count measures

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

TRANSPORT SYSTEM & OPERATIONS	• public transport statistics	• stops • fares • public transport schedules	• bike, scooter availability • parking availability	• real-time public transport schedules
TRAFFIC & MOBILITY CONDITIONS		• traffic rules & regulations • traffic network	• traffic notices & notifications	• traffic disruptions • traffic light status
ENVIRONMENTAL & EXTERNAL CONDITIONS	• satellite images • statistics	• environmental data • mechanisms / policies	• weather	

3.2 METADATA STANDARDS

Metadata standards ensure that datasets can easily be discovered and understood. In the mobility domain data stemming from various origins, such as sensors, vehicles, transport operators, infrastructure providers and traffic authorities, require different metadata standards to ensure coverage, interoperability and alignment with FAIR principles. A list of metadata standards applicable to FEDORA can be found in Table 9.

Table 9 Metadata standards, their domain and short description.

METADATA STANDARD	DOMAIN	SHORT DESCRIPTION
Dublin Core	General-purpose datasets	Dublin Core is a collection of metadata elements, properties, classes, datatypes, vocabulary and encoding schemes.
DCAT/DCAT-AP	General-purpose dataset	DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web. DCAT-AP is an application profile of DCAT used by public sector datasets in Europe.
GeoDCAT-AP	Geospatial data	GeoDCAT-AP is an extension of DCAT for describing geospatial datasets.
mobilityDCAT-AP	Mobility-specific datasets	mobilityDCAT-AP is a mobility extension of DCAT.
ISO 19115/19139	Geospatial data	ISO 19115 defines schemas for describing geographic information and services by means of metadata, while ISO 19139 defines corresponding encoding rules.
FGDC CSDGM	Geospatial data	FGDC CSDGM provides a common set of terminology and definitions for the documentation of digital geospatial data.
SensorML	Sensors and IoT	SensorML provides a means of defining processes and components associated with the measurement of observations.

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

OGC SensorThings API	Sensors and IoT	OGC SensorThings API provides a standard geospatial-enabled way to manage and retrieve observations.

Additionally, to the above-named metadata standards, self-describing data formats and standards are worth being mentioned. These data formats and standards embed metadata directly into the file structure and include NetCDF, HDF5, GeoPackage, OGC API Features/Tiles/Routes, DATEX II, NeTEx, SIRI, INSPIRE, GTFS, and GBFS.

When choosing the right metadata standard to use aspects to consider include

- the **domain** of data to be described,
- the intended **use** for the metadata,
- **interoperability** requirements and
- alignment with **FAIR principles**.

In addition to metadata standards that define the structure and content of metadata records, several frameworks and services are relevant for metadata exchange and discovery in the mobility and geospatial domain. In particular, the *OGC Catalogue Service for the Web (CSW)* is widely used to publish and harvest metadata records. Furthermore, the *INSPIRE Directive* establishes a European legal framework that mandates the use of harmonized metadata profiles and standardized network services to enable cross-border discovery and interoperability of spatial datasets.

3.2.1 METADATA QUALITY, NAPS AND MOBILITYDCAT-AP

The development of a federated metadata catalogue[1]—indexing more than 1000 datasets from diverse sources— is a key enabler for improved understanding, accessibility, and reusability of mobility data. The diversity of producers, production frameworks, and dataset maturity levels poses significant challenges for metadata harmonisation, semantic consistency, and technical interoperability.

In this context, metadata quality is an essential prerequisite for ensuring the catalogue's datasets scalability[2], discoverability[3] and comparability, as well as their understanding and automated integration into digital services. The alignment of FEDORA with the European data quality principles—particularly those promoted by NAPCORE—provides a foundational framework for ensuring regulatory compliance, cross-border reuse, and the long-term impact of the project's results within the European Mobility Data Space.

The adoption of a common metadata standard, such as MobilityDCAT-AP within FEDORA, will allow to structure dataset descriptions in a consistent way and to align with European practices that will facilitate interoperability as well as interconnection with other data portals and National Access Points.

---

[1] Federated metadata catalogue: a catalogue that does not physically centralize data, but rather references, describes, and makes discoverable datasets hosted and maintained by multiple sources through harmonized metadata. In other words, the federated catalogue acts as a metadata orchestration layer, not as a data warehouse.
[2] Catalogue scalability: the ability of the metadata catalogue to scale over time without loss of performance, consistency, or quality, despite an increase in the number of referenced datasets; the growing diversity of sources and producers; evolving standards, practices, and regulatory requirements; and an increase in the number of users and use cases. Scalability, therefore, concerns the architecture, governance, and quality of the metadata, and not solely the technical capacity of the system.
[3] Dataset discoverability: the ability of a catalogue to enable users to quickly identify, find, and understand the existence and relevance of a data set, without prior knowledge of the source or producer.

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

Based on this, FEDORA will incorporate a verification criterion on metadata quality, focusing on:

- The completeness of required fields, aims to verify that all mandatory and essential metadata fields are correctly filled in. This includes:
    - The dataset description (content, purpose, functional scope)
    - Spatial coverage (geographic area covered)
    - Temporal coverage (time period covered by the data)
    - License and reuse conditions
    - Point of contact (technical or subject matter expert)
    - Access/distribution methods (URL, API, available formats)
- Compliance with MobilityDCAT-AP schemas and profiles, ensuring technical interoperability. This includes:
    - Structural compliance (a dataset must include at least a unique identifier, a title, a description, and at least one valid access point).
    - Syntactic compliance (metadata adheres to the technical formats required to enable automatic ingestion into other catalogues or search engines).
    - Semantic compliance (fields are used in accordance with their meaning as defined in the standard to ensure consistent interpretation of metadata by third-party systems).
- The internal semantic consistency of the metadata and its alignment with the actual data made available.
- The relevance and freshness of the metadata (consistent with the announced frequency updates).
- Traceability, including clear identification of the provider, responsibilities, and versions.
- The existence of quality assessment documentation describing the criteria applied, the methods used, and, where applicable, the identified limitations, in accordance with NAPCORE recommendations.

This verification criterion will help ensure that the referenced datasets are described in a reliable, comparable, and exploitable manner; thereby enhancing transparency, trust, and the ability of re-users to assess data quality, while also serving as an essential prerequisite for the scalability and sustainability of the FEDORA catalogue. For more information on data and metadata quality in the context of the NAPs, please refer to the Annex "Data and metadata quality in the context of the growing number of national access points in Europe".

3.3 DATA FORMATS AND STANDARDS

The selection of appropriate data formats and standards is essential for ensuring that project data can be efficiently created, shared and preserved over time. A well-considered choice should take the following aspects into consideration:

- The standard should be well **documented**, with publicly available specifications.
- **Independence** should be given from **vendor** and from **platform**.
- While **compression** is beneficial, a **lossless** method should be used.
- Data formats should **avoid encryption** unless required for security or privacy reasons.
- **Open** standards should be used.
- The format should provide the data in a **machine-readable** way.
- **Versioning** and **backwards compatibility** should be given.
- The used standard should be widely used and have **community adoption**.
- **Validation** of correctness and **recoverability** of corrupted data should be supported.
- **Human readability** for inspection should be given.
- The format should be **scalable** and **efficient** for the expected **data volumes**.

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

In cases where the initially used format and standard are not optimal for sharing and preservation a conversion of the data to another format and standard may be beneficial.

Based on the initial list of preferred formats and standards presented in D1.2 Data Management Plan and Ethics (Obregon, Chevallier, Cheliotis, & Ziogou, 2025) the table in Section 3.3.1 gives suggestions for mobility-related standards and formats.

3.3.1 RECOMMENDED FORMATS

Each data typology has different needs for information and thus requires different data formats. For the data typologies from Table 8 we propose the data formats listed in Table 10. The list of recommendations is not finalized within the FEDORA project and will need to be adjusted as new needs arise from use cases and implementations.

Table 10 Recommended Standards and Data Formats for various data typologies.

DATA TYPOLOGY	RECOMMENDED DATA FORMAT
geospatial statistics	GeoPackage
infrastructure	*For C-ITS data:* C-ITS ASN.1 compliant with C-Roads specifications; *for others see network graph, stations, stops, parking availability*
network graph	OSM, GTFS, NeTEx, INSPIRE Transport Networks
stations	GTFS, NeTEx, INSPIRE Transport Networks
topography	INSPIRE Elevation, GeoTIFF
activity and travel diaries	SDMX
travel surveys	*For tabular data*: CSV, Parquet
bike, scooter usage	GBFS
origin destination flows	*There is no common interoperable format for OD matrices; import and export formats are usually software specific; text-based formats are recommended*: CSV; V- and O-Format *for PTV Visum;* Amitran OD format *for SUMO*
traffic model	*There is no common interoperable format for OD matrices; import and export formats are software specific; text-based formats are recommended:* .net-Files *for network objects,* .demd-Files *for demand,* V- and O-Format *for OD Matrices in PTV Visum; SUMO specific* XML Formats *for networks, trips, and states*
FCD	GPX (FCD trajectories)
object lists of traffic participants	*For C-ITS data:* C-ITS ASN.1 compliant with C-Roads specifications
real-time locations	*for C-ITS data:* C-ITS ASN.1 compliant with C-Roads specifications, *for public transport:* GTFS-RT, SIRI, *for ships:* AIS
traffic count measures	*For real-time data:* DATEX II, *for historical tabular data:* CSV, Parquet

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

public transport statistics	NeTEx, GTFS
stops	NeTEx, GTFS
fares	NeTEx, GTFS
public transport schedules	GTFS, NeTEx
bike, scooter availability	GBFS
parking availability	DATEX II, C-ITS ASN.1, APDS; *for C-ITS data:* C-ITS ASN.1 compliant with C-Roads specifications
real-time public transport schedules	GTFS-RT, SIRI
traffic rules & regulations	DATEX II, *for C-ITS data:* C-ITS ASN.1 compliant with C-Roads specifications
traffic notices & notifications & disruptions	DATEX II, *for C-ITS data:* C-ITS ASN.1 compliant with C-Roads specifications
traffic light status	*for C-ITS data:* C-ITS ASN.1 compliant with C-Roads specifications
satellite images	GeoTIFF, NetCDF
Statistics	*For tabular data*: CSV, Parquet
environmental data	*See Inspire Recommendations*
mechanisms / policies	*No recommendation; this is mostly provided as human-readable text*
Weather	DATEX II (*weather-related traffic conditions*); *see also Inspire Recommendations*

3.4 INTEROPERABLE INTERFACES

Interoperable interfaces are standardized technical access points that enable systems to exchange data and metadata across organizational and technical boundaries. In the context of a data space, they provide the means by which data providers, consumers, catalogues and services can interact in a consistent, machine-readable and automated manner.

Interoperable Interfaces are essential for effective data exchange, particularly in heterogeneous environments such as the mobility domain. Obstacles for interoperability often arise from a lack of standardised interfaces for certain types of data, issues with ambiguous metadata descriptions, and technical, semantic, and organizational challenges. These challenges can significantly increase integration efforts and limit data reuse.

From a technical perspective, interoperable interfaces are implemented using widely adopted communication protocols, depending on the type of data exchange required. Recommended example interfaces for various use cases are listed in Table 11.

Table 11 Technical interface types including typical use cases and examples.

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

INTERFACE TYPE	TYPICAL USE CASES	EXAMPLES
Batch Data Access	Static datasets, historical data	REST API, HTTP, OGC API
Streaming Data Access	Real-time data, sensor data	MQTT, AMQP, WebSockets
Metadata and Discovery	Discovery, searching of datasets	CSW (geospatial catalogues)
Dataspace Connector	Integration of data from other domains	EDC

Semantic interoperability ensures that exchanged data is interpreted consistently. This includes the use of stable and unique identifiers (e.g. dataset identifiers, object identifiers) as well as shared references systems. For the mobility domain spatial and temporal reference systems have high relevance.

Organizational interoperability is addressed through mechanisms which support authentication and authorization, as well as the enforcement of service contracts governing data access and usage conditions. These aspects are particularly relevant in federated data spaces, where participants retain sovereignty over their data while enabling controlled sharing.

To support interoperability of interfaces in practice, the following requirements should be met:

- Support for automation & machine-readable access
- Standardized authentication & authorization
- Sufficient scalability, reliability, and compatibility to operate across diverse systems and use cases

3.5 DATA SOURCES AND DATA PORTALS

Data sources are heterogenous and may differ according to each pilot of the project. An inventory was created based on the description of each pilot. This list started with the National Access Points (NAPs), regional, municipal and open access mobility (meta)data catalogues. Websites from public authorities, institutions and other (meta)data portals were added according to the geographical areas and subjects covered by each pilot.

3.5.1 INVENTORY OF DATA SOURCES FROM DATA PORTALS

A (meta)data inventory of data sources was created to track the most relevant data portals and the main information about the type of datasets available in their data portals or websites. Essential details about the identified datasets were added into fields such as name, description, data format, specification or standard, license, spatial coverage, and more. The details were selected and based on the mandatory fields recommended by the mobilityDCAT-AP specification of NAPCORE to comply with the FAIR principles and to facilitate the description and exchange of metadata. These details allow us to identify if the datasets are recent or updated frequently, if the catalogues are well maintained, but also to learn about what standards/specifications, API's or services are associated with the data catalogue or website to facilitate harvesting.

The data sources vary but among the major categories we can identify data portals, databases, data distributors, websites from governmental authorities, transport and logistics authorities and operators, organizations, associations and communities, academic research centre repositories, institutes, registries, agencies, private and commercial service providers, companies and diverse platforms.

These data sources covering multi-domains or specific ones such as transport, statistics, tourism, mobility, journey planning, traffic, infrastructure, road, tolling, smart cities, logistics, sciences, aviation, coordination, earth observation and environment (land, hydrology, water, meteorology, climate), geospatial (cadastral and mapping) and information services.

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

In Table 17 in the Data Sources Inventory in the Annex you can find some sources and some of the details gathered in our inventory. Datasets which come from open access (meta)data catalogues that comply with FAIR practices and use broadly used standards/specifications can be more easily harvested and they are prioritized sources if the quality and type of data needed is suitable.

Some of the most frequent transport and mobility themes of the datasets are: air and space travel, cycle network data, dynamic traffic signs and regulations, freight and logistics, general information for trip-planning, parking, service and rest area information, public transport (non-)scheduled transport, real-time traffic data, waterways, road events and conditions, road work information, sharing and hiring services, static traffic signs and regulations, static road network data and toll information.

3.5.2 INVENTORY OF DATA SOURCES FROM FEDORA PARTNERS

Additional datasets are continuously provided by the consortium partners, and the list of data sources refined with the help of the different pilot referents. This list will be updated as necessary and as new relevant information regarding the pilots arises. The table to track the metadata shared for these sources is composed of mandatory, recommended, optional and internal information fields based on the needs to comply with the FAIR principle, the Mobility-DCAT-ap specification and/or requisites of CKAN and project's needs to track and associate datasets with pilots. See the table below to have information about the fields requested and the level of importance to collect this metadata. Required metadata fields can be found in Table 12.

Table 12: Metadata fields required for the datasets proposed by the partners

FIELDS	OBLIGATION
Dataset Identifier, Name of the dataset, Spatial, Description, Dataset publisher, Dataset release date, Dataset modification date, Dataset update frequency, Dataset theme, Dataset keyword, Access policy, License, Dataset distribution, Format/ media type, Language, Ethics and legal aspects;	mandatory
Mobility data type, Mobility data standard/ specification Dataset spatial resolution, Dataset quality Information	recommended
Dataset provider, Dataset owner, Dataset documentation, Metadata file/ description.	optional
Related Task/pilot(s), Storage (location), Is the data freely accessible? Are data access conditions defined? If data is to remain restricted, is metadata accessible? Source code, Potential Harmonisation Issues, Enrichment Potential, Relevance, and Synthetic Data Need.	internal information

In Table 13 some examples of the metadata gathered and shared by the partners are listed.

Table 13: Metadata as shared by the partners

ORGANIZATIO N NAME /	NAME OF THE DATASET	SPATIAL COVERAGE	MOBILITY DATA TYPE	FORMAT / STANDARD / SPECIFICATIO	ACCESS POLICY (OPEN	DATASET UPDATE

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

DATASET PUBLISHER				N FOR METADATA SHARING	ACCESS Y/N)	FREQUENCY
Verkehrsverbund Ost-Region (VOR) GmbH, ITS Vienna Region	Traffic Demand Model for Eastern Austria	Eastern Austria (Wien, Niederösterreich, Burgenland)	historical, predictive	PTV Visum version file / GTFS / NETEX, DINO	restricted	IRREG
Verkehrsverbund Ost-Region (VOR) GmbH, ITS Vienna Region	Traffic participant counts as detected by cameras using a machine learning system	limited location within Vienna	real-time and historical	TBD	restricted	1MIN
Verkehrsverbund Ost-Region (VOR) GmbH, ITS Vienna Region	Traffic Counting Sensor Data	Eastern Austria (Wien, Niederösterreich, Burgenland)	real-time and historical	TBD	restricted	1MIN
Stadt Wien	Validated traffic counts	Vienna	historical	CSV	open	ANNUAL
Verkehrsverbund Ost-Region (VOR) GmbH, ITS Vienna Region	Online Traffic Model	Eastern Austria (Wien, Niederösterreich, Burgenland)	real-time traffic data	JSON, XML / DATEXII(?)	visualization is open, data is restricted	5MIN
Verkehrsverbund Ost-Region (VOR) GmbH, ITS Vienna Region	Traffic Management 2.0 Evaluation	Vienna		PDF	restricted	NEVER
DTU (DK)	Simulated Daily Activity Schedules	Denmark	simulated	CSV	open	TBD

Funded by the European Union

Project funded by Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER State Secretariat for Education, Research and Innovation SERI

4 THE FEDORA PUBLIC CATALOGUE

The FEDORA public-facing catalogue provides publicly accessible metadata about the data assets used in the FEDORA project. It serves as a central component, providing a single, publicly visible point of access, in contrast to decentralized connectors that operate behind the scenes. The aim is that all synchronized datasets—whether from dataspace connectors or harvested open data sources—are available in the same search interface. This allows seamless exploration and discovery without differentiating the source, while still ensuring metadata provenance is traceable internally.

It should be noted that even though data assets' metadata are listed in the catalogue, the corresponding data assets themselves are not necessarily publicly accessible. Access may remain restricted to authorised dataspace participants. Moreover, data assets for which the provider does not permit even the metadata to be made public are not included in the public-facing catalogue and are only registered in the access-controlled dataspace catalogue, at the connector's level (more details in D3.2 (Chevallier, et al., 2026)).

The public catalogue relies on CKAN (Comprehensive Knowledge Archive Network), an open-source solution for metadata management led by the Open Knowledge Foundation. CKAN is a system for data hubs with an active community. Many transport and many European national authorities use it as their official open data portal. Further details on the reasons why the CKAN solution has been chosen to build the FEDORA catalogue are provided in D3.2. Being based on CKAN, the FEDORA public catalogue provides metadata discovery, standardized metadata display and access, harvesting, and access control. The catalogue can handle DCAT and particularly DCAT-AP – a standard especially promoted by NAPCORE for metadata exchange and harmonization across NAPs. Therefore, the catalogue allows to discover datasets whether they originate from internal data offers or harvested external sources.

4.1 DATASETS METADATA COLLECTION

4.1.1 MANUAL COLLECTION

Metadata for datasets can be added one by one (via the user interface or the back end) or via a CSV file in the catalogue (via the CKAN API). Manual metadata entry is typically used when there are only a few datasets to be added on an ad hoc basis by organizations with administrator rights. However, when many datasets need to be collected and processed, the best approach is to use the harvesting mechanism, as manual addition could require a considerable time investment. Harvesting is also particularly useful when the datasets are not created by the organization managing the catalogue but are mostly collected into a single catalogue from many different external sources and harmonization is required to be considered in the collection process.

4.1.2 HARVESTING AND INTEROPERABILITY

Metadata can be added into the FEDORA catalogue via an automatic extraction from external sources. To this end, the catalogue implements a set of CKAN-based harvesting extensions that can ingest metadata at a predefined frequency (either manual, monthly, weekly, biweekly, daily or continuously). The harvesters operate through mechanisms that enable the exchange of metadata through standards and specifications that present information about datasets in a standardized way. These standards ensure interoperability.

4.1.2.1 INTEROPERABILITY: METADATA STANDARDS USED

The default harvester deployed in FEDORA is the CKAN harvester which ingests datasets from CKAN catalogues providing a CKAN API endpoint (it is important to note that not all CKAN catalogues provide this endpoint).

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

Most open data portals share their metadata through a DCAT or DCAT-AP endpoint, which can be provided in different formats such as JSON, JSONLD, RDF/XML, TTL, Turtle Notation, 3 N-Triples, SHACL or sector/national specific profiles such as GeoDCAT-AP, StatDCAT, and DCAT-AP national profiles (e.g. DCAT-AP-DE, DCAT-AP-IT, DCAT-AP-DK, DCAT-AP-ES, DCAT-AP-NO, DCAT-AP-SE, DCAT-AP-BE, DCAT-AP-CZ, etc.). This requires the use of the Ckanext-dcat extension which allows to "expose and consume metadata as serialized RDF documents using DCAT" (W3C, 2024).

Most geo-data portals share their metadata by using the CSW or INSPIRE services. Harvesting such data sources requires the installation of the ckanext-spatial (CKAN, 2026) extension which provides support for the OGC standard.

Other catalogues propose an API to expose their metadata, to harvest this metadata an additional extension can be installed. Ckanext-basket (CKAN) is an extension that increases the type of platforms that can be harvested, for instance: ODS, ArcGIS, Socrata, DKAN, Junar, and more.

Harvested datasets will be normalized and integrated into the catalogue, providing users with a unified view of both internal and external data offerings. This catalogue will propose a mobilityDCAT-AP endpoint for external users who wish to access the datasets in an interoperable and harmonized way. Metadata from harvested sources will be further enriched or tagged to align with the data space taxonomy, this process will be covered in the deliverable 3.2.

Given that mobilityDCAT-AP is an extension of DCAT-AP, the mobilityDCAT-AP endpoint will be implemented on top of the existing DCAT extension. Its compliance will be validated against the official mobilityDCAT-AP requirements and interoperability criteria, to ensure that the endpoint provides a consistent, interoperable, and harmonized interface for external users.

4.1.2.2 PROCESS TO ADD A HARVESTING SOURCE

To harvest datasets from another catalogue, the information about the organization must be registered first by an organization admin. Once there, a connected admin can add a harvesting source.

To add a source, the endpoint URL and the title of the source must be entered in the interface or through a SSH connexion. The source type and the update frequency must be selected, and the description and the configuration are optional.

4.1.2.3 CONFIGURATION AND FILTERS

Harvesters in FEDORA support several configuration options, which must be defined as a JSON object in the configuration form field. These include filtering groups and organizations to collect, adding or clearing tags, adding default groups, specifying an API key or version, adding or overriding extra key-value pairs, and assigning a user to the harvesting process.

Here is an example of a configuration object (see Figure 4):

Funded by the European Union

Project funded by Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra Swiss Confederation Federal Department of Economic Affairs, Education and Research EAER State Secretariat for Education, Research and Innovation SERI

Configuration:

```
{ "api_version": "3,
"default_extras": {"encoding":"utf8"},
 "organizations_filter_include": [],
 "remote_groups": "only_local",
 "user":"harverster-user",
 "api_key":"<REMOTE_API_KEY>",
 "remote_orgs": "create"
}
```

Figure 4: Harvesting configuration example

4.1.2.4 MASSIVE METADATA SOURCE MANAGEMENT

Managing metadata sources individually through the graphical user interface can become time-consuming and error-prone when working with a large number of sources. To address this limitation, a command-line utility has been developed to automate the management of harvest sources using a CSV file as the input. This approach enables administrators to perform bulk operations with a single command, improving efficiency, consistency, and scalability when configuring data harvesting.

Each row in the CSV file represents a harvest source and contains the configuration parameters required to create or modify that source. The tool processes the CSV file sequentially, allowing multiple harvest sources to be managed in a single execution. This significantly reduces the manual effort involved in deploying or maintaining large metadata catalogues.

Figure 5 shows the available commands and options provided by the harvesting management tool.

```
maheritiana@AK68051:~$ fedora-harvester --help
usage: fedora-harvester [-h] [--verbose] [--row_number ROW_NUMBER] {harvest,delete,update} ...

Automate CKAN harvest jobs from a CSV file.

positional arguments:
  {harvest,delete,update}
                        Sub-command
    harvest             Create/ensure harvest sources and trigger jobs
    delete              Delete harvest sources, datasets, and organizations
    update              Update harvest sources based on CSV definitions

options:
  -h, --help            show this help message and exit
  --verbose, -v         Enable debug logging.
  --row_number ROW_NUMBER, -n ROW_NUMBER
                        Row number to process.
```

Figure 5 Harvesting source management command line tool

The utility provides three main operations:

- **harvest** – Creates new harvest sources or validates the existence of previously configured ones and automatically triggers the corresponding harvesting jobs to populate the data catalogue.
- **update** – Modifies the configuration of existing harvest sources according to the definitions provided in the CSV file, allowing administrators to apply configuration changes consistently across multiple sources.
- **delete** – Removes the selected harvest sources together with their associated datasets and organizations, providing an efficient mechanism for cleaning up obsolete or deprecated metadata sources.

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

In addition, the tool supports optional parameters such as selecting a specific CSV row for processing and enabling verbose logging for debugging and monitoring purposes. These options facilitate testing, troubleshooting, and incremental deployment of harvest source configurations.

By automating bulk harvest source management, the utility reduces the need for manual configuration, lowers the risk of human error, and facilitates the tracking of data sources, improving overall governance.

4.1.2.5 AUTOMATED UPDATE AND VALIDATION

Besides harvested metadata sources, the FEDORA catalogue will also collect metadata from the decentralized dataspace connectors. This process is described in more details in D3.2 (Chevallier, et al., 2026). The frequency, depending on connector capabilities, will be set so that public users see up-to-date information. Validation checks will ensure completeness and quality of the metadata, maintaining a reliable user experience.

4.1.2.6 CUSTOM HARVESTER

CKAN provides a harvesting framework through extensions/plugins that allow ingestion of metadata from multiple source formats such as DCAT RDF, DCAT JSON, and native CKAN instances. However, not all external data sources are directly compatible with these standard harvesters. In particular, platforms like **data.gv.at** (Austria's open data portal) which is used by one of the six pilot sites, expose metadata in formats that require additional transformation before they can be ingested into CKAN.

To address this limitation, a custom harvesting plugin was developed to enhance CKAN's harvesting capabilities. This extension builds on top of the existing **ckanext-dcat** extensions and integrates the **dcat-piveau** plugin. The role of dcat-piveau is to consume piveau metadata formats and transform them into a structured representation that CKAN can harvest.

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

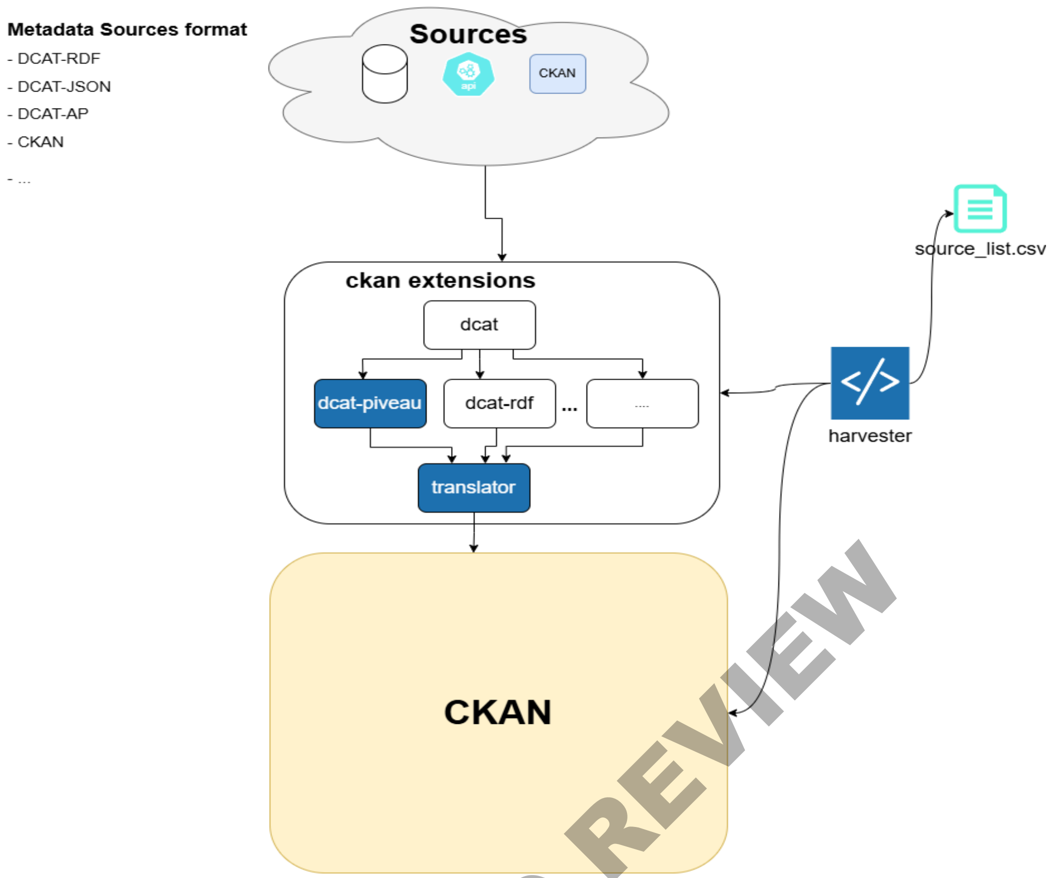

Figure 6 Harvesting & Extensions Architecture

Figure 6 presents our harvesting architecture. As described in Section 4.1.2.4, the process begins with a CSV file containing the harvest sources, this file is ingested into CKAN by the fedora-harvester (a custom component introduced in the same section), to initiate and trigger the harvesting process.

Once the harvesting process is triggered, the metadata content is retrieved by ckanext-dcat and passed to the dcat-piveau plugin, which transforms it into the required format and another custom plugin translates it into English (as presented in 4.3) before the final ingestion.

4.2 DATA VISUALIZATION AND SAMPLE DATA

According to the type and file content of the resources of datasets, data can be visualized in several ways. These visualizations are called resource views. A resource can propose multiple types of views for the same data such as graphs, maps, tables and bar charts, but it is also possible to display an image, a video, a pdf reader, a text view, a web page view or add an audio player.

4.3 MULTI-LINGUAL/TRANSLATION DISPLAY

When harvesting metadata from different regional portals (such as the Austrian government data portal), the datasets are often in their local language. If the catalogue contains mixed languages, core features like metadata search become more fragmented and difficult for users to navigate.

To solve this, we built a custom CKAN translation extension that automatically translates metadata into English as it enters the system. Under the hood, this plugin connects to a self-hosted instance of LibreTranslate, an open-

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

source machine translation API. Using self-hosting and an open-source project ensures data privacy, removes external dependencies, and most importantly, prevents us from hitting any external API rate limits during bulk data harvesting.

The plugin works by hooking into the CKAN's standard chained actions. It intercepts the incoming data payloads during the "package_create" and "package_update" API events. This means that whenever a new dataset is harvested or an existing one is modified, the plugin automatically extracts the main fields of the dataset like the Title, the Description, and the Tags and sends them through the translation pipeline.

To ensure the original source data is never lost, the translation process is non-destructive. The new English text is injected into the main CKAN fields so that it gets properly indexed by the Solr search engine[4]. At the same time, all the original texts are preserved and saved into the dataset's extras dictionary (CKAN's default key-value store for custom fields). This approach provides the best of both worlds: a unified, English-searchable catalogue, while keeping the original local metadata safely stored and accessible, and searchable in its native language.

To maintain the translation quality, the system uses LibreTranslate's language detection. Observational checks indicate an accuracy rate of approximately 90%. However, with very short text strings the language-detection confidence can drops, leading to skipped translations. We also implemented fallback and error-handling mechanisms: if an error occurs, the system falls back to displaying the original text. Due to those edge cases, a manual review is used to check the catalogue and manually update fields that failed to translate properly.

In the future, we are looking to directly an integrate a locally hosted Large Language Model (LLM) to improve this translation pipeline. This local LLM could be used to resolve low-confidence language detection edge cases or replace LibreTranslate entirely. This would improve translation accuracy while maintaining our requirement for data sovereignty and no reliance on external APIs.

4.4 SEARCH AND DISCOVERY

The metadata search function is available in the FEDORA catalogue as either a simple or advanced search. By default, on each page of the catalogue there are usually two search boxes available to search datasets.

A simple search is composed of a single or more keywords without characters. The plus ("+") or minus ("-") symbols can be used as modifiers in the search for a term (without their use, words are considered as optional). The plus symbol designates a mandatory term to be included and the minus symbol a term to be excluded in the search (they must precede the term or phrase). To consider an ensemble of words as a phrase they must be wrapped by quote signs (""); the search will look for the works in the exact same order.

An advanced search requires the use of a keyword and at least one colon (":"). This function allows to search terms by following a field:term basic syntax, which means that terms can be searched per field and can make use of wildcards ("*") and proximity matching ("~") terms and/or Solr features[4].

Simple search examples:

+ Nicosia will provide all the results of datasets containing the term Nicosia. However, if we search for Nicosia -city the results will show the same results, but without the datasets containing the word city.

"Municipality of Copenhagen" will show all the datasets comprising that sentence.

[4] https://solr.apache.org/features.html

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

Advanced search examples:

title: real-time will look for all the datasets containing in its title the word “real-time”

title:share* will look for all the datasets containing in its title a word that starts with “share” like “shared” and “sharing”.

title:Biscay || title:traffic look for datasets containing “Biscay” or “traffic” in its title.

title: "pedestrian zone" ~ 4 is a proximity search that looks for terms that are within a specific distance from one another. This example will look for datasets whose title has the words “pedestrian” and “zone” within a four words distance.

text:ycle~ CKAN supports fuzzy searches based on the Levenshtein Distance or Edit Distance algorithm. To do a fuzzy search use the “~” symbol at the end of a single-word term. In this example, words like “cycle” or “CLE” will also be found.

The search result can be filtered by organization(s), group(s), tag(s), format(s) and license(s). When selecting a dataset result, the metadata and resources attached to this dataset will be displayed.

The resources description can be displayed by a simple click on “Explore” + “More information” or it is also possible to be directly redirected to the source of the resource by a simple click on “Explore” + “Go to resource”. According to the type of resource, sometimes the option to download directly a file is available when clicking on “Explore”.

A dataset can have more than one resource. These resources can be explored or downloaded. Many types of resources can also be previewed directly on the resource page. Spreadsheets (.CSV and .XLS) can be previewed in a grid view. Map and graph views can also be previewed if the data format is suitable.

Datasets can be also discovered on the organizations and groups pages with the help of filters.

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

5 DATASET HARMONISATION, ENRICHMENT AND SYNTHETISATION FOR COMPLEX DATA EXCHANGE

In the context of modern digital infrastructures, the reliable exchange of high-quality data is a fundamental requirement. This section presents the Mobility Data Harmonisation, Enrichment, and Synthetisation Engine (MDHESE) component, a dedicated architectural module engineered to standardise and elevate the quality of shared information. Specifically, this section details the technical framework through which raw, fragmented inputs are systematically processed into structurally consistent and standardised data assets. By establishing a robust data refinement pipeline, the MDHESE component ensures that output datasets are fully compliant with European data standards, providing a secure and reliable foundation for cross-platform integration and advanced analytics.

The mobility sector is defined by a highly heterogeneous data landscape, combining static transport network graphs with continuous, real-time sensor streams. Extracting operational value from these sources involves navigating significant technical barriers. These include the absence of unique entity identifiers across independent databases, the prevalence of structural inconsistencies, and the inherent noise within temporal observations. Additionally, the processing of geolocation and travel logs is strictly constrained by privacy and data protection frameworks.

The MDHESE component is specifically designed to overcome these domain-specific challenges. To resolve representational ambiguities, the system incorporates advanced probabilistic data linkage techniques, whilst time-series filtering models are deployed to eliminate sensor noise and ensure temporal accuracy. Furthermore, to overcome privacy limitations and the unavailability of production data during system testing, the component integrates synthetic data generation capabilities. Ultimately, this structured approach ensures the delivery of refined, privacy-compliant datasets that are directly applicable to the training of artificial intelligence models and the advancement of data-driven mobility applications.

To maximize flexibility and adapt to diverse operational requirements, all core functionalities of the MDHESE component are encapsulated into an API architecture, Figure 7. This API will be seamlessly integrated into and accessible through the data space, allowing both data providers and data consumers to trigger specific processing workflows according to their needs—whether a provider requires pre-ingestion data cleansing or a consumer seeks post-acquisition data refinement. Currently, each functionality operates independently, requiring the user to iteratively chain multiple processes if several data treatments are needed. Furthermore, to address strict privacy constraints, an on-premises deployment option is currently under development for the future version. This local setup will enable users to execute anonymization methods directly within their own infrastructure, ensuring that sensitive data is fully anonymized before any cloud transmission takes place.

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

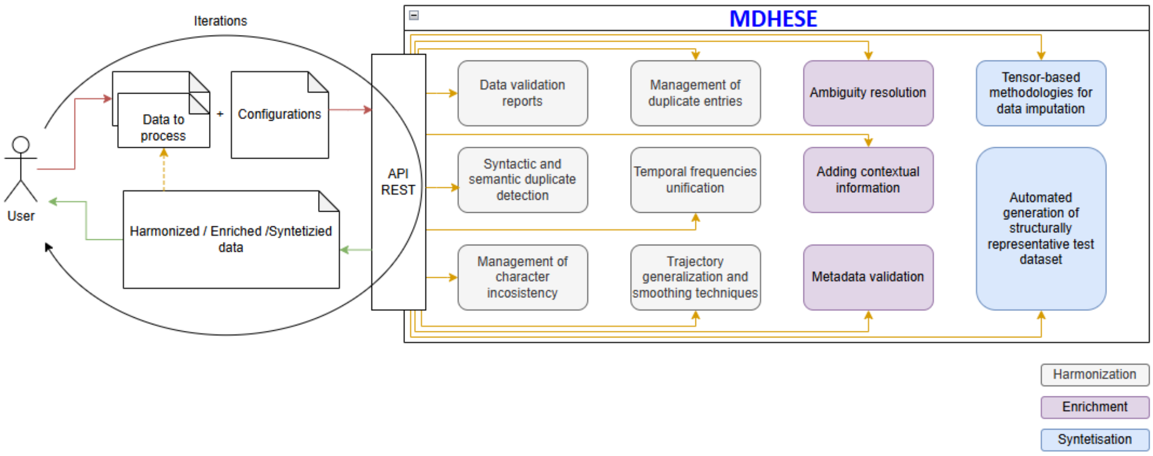

Figure 7 MDHESE architecture

To complement the API architecture, Table 14 outlines the primary input and output data types expected and generated by each MDHESE module. In addition to the raw input datasets—predominantly in CSV format or zipped GTFS feeds—certain functionalities require specific configuration parameters (such as specifying CSV delimiters, selecting the model to apply, thresholds …) to tailor the execution to the dataset's characteristics.

Table 14 Specification of the general input and output for each module

MODULE	INPUT DATA TYPE	OUTPUT DATA TYPE
Data validation report	CSV file	JSON
Syntactic and semantic duplicate detection Management of character inconsistency Management of duplicate entries	CSV file	CSV file
Temporal frequencies unification Ambiguity resolution Adding contextual information	CSV files	CSV file
Trajectory generalization and smoothing techniques	GTFS ZIP FILE	CSV file
Metadata validation	RDF file	JSON
Tensor-based methodologies for data imputation	CSV file	CSV file
Automated generation of structurally representative test dataset	JSON	CSV file

5.1 DATA HARMONIZATION

Harmonisation is the process of ensuring data consistency and compatibility by preparing the dataset for further analysis. This process improves and accelerates analysis, enhancing its efficiency and consistency. Within the harmonisation process, two main tasks stand out: data profiling to identify the issues, and data wrangling to

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

implement the solutions. MDHESE facilitates the preparation of the dataset, offering a range of functionalities for each step in the process.

5.1.1 DATA PROFILING

Effective data profiling is a fundamental step in the harmonization of datasets. In this context, the provision of thoroughly validated datasets is critical within any information exchange ecosystem. Data utility is significantly enhanced when datasets are accompanied by transparent reporting and metrics that certify their reliability. Without systematic profiling and validation, there is a substantial risk of propagating errors throughout the data lifecycle, which can compromise analytical models and downstream decision-making.

The MDHESE developed component incorporates a module to generate comprehensive validation reports to efficiently diagnose the state of datasets. It provides clear insights into any detected anomalies to ensure high data quality. This functionality uses the Pointblank (Posit, 2026) library, a data validation toolkit designed for the rigorous assessment and monitoring of data quality across diverse tabular formats.

By centralising this evaluation, MDHESE transforms data quality metrics into actionable insights. Specifically, this module enables relevant stakeholders to:

- Define and execute precise validation rules tailored to the dataset's structural requirements. This includes fundamental quality checks such as validating data schemas (e.g., verifying field constraints, ranges, and patterns), quantifying missing values to assess data completeness, and detecting duplicated rows to enforce record uniqueness.
- Establish customisable alert thresholds (e.g., 'warning', 'error', and 'critical' levels) adapted to these specific operational rules, allowing the automated monitoring of the data.
- Translate technical validation results into accessible formats for various stakeholders, ensuring quality issues are immediately identifiable and resolvable.
- Isolate and extract problematic data instances directly based on the validation outputs, facilitating rapid remediation before the data is utilised in downstream AI models or mobility applications.

Through this functionality, MDHESE ensures the robustness of the shared data, substantially maximising its reusability and utility for end-users across the mobility domain.

To enhance the resulting analysis, the component incorporates a dedicated module for syntactic and semantic duplicate detection. By leveraging PolyFuzz (Grootendorst, 2022), MDHESE can suggest potential duplicate entries for stakeholder review. This framework performs fuzzy syntactic and semantic matching, associating structural variations (e.g., “San Sebastián-Donostia” and “S. Sebastián-Donostia”), and semantic equivalencies (e.g., “Pamplona” and “Iruña”). There are three models to choose from to calculate the matching, as defined in Table 15:

Table 15 Specifications of the available models

MODEL	DESCRIPTION	EXAMPLE
EditDistance	Calculate the minimum number of character changes (insertions, deletions, or substitutions) needed to transform one string into another. Syntactic match.	Gipuzcoa | Gipuzkoa Miss-spelling error

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

TF-IDF	Breaks text into fragments (n-grams) to measure their statistical importance and compare the overlap between strings. Syntactic match.	San Cristóbal | S. Cristóbal Abbreviation
Embeddings	Transforms text into numerical vectors that represent the concept of the texts. Compare how similar both concepts are. Semantic match.	San Sebastián | Donostia Different languages

Figure 8 shows an example of the result obtained when using the 'TF-IDF' model, in which 'Gipuzkoa' is syntactically grouped with "Gipuzcoa" and 'S. Cristobal' with 'San Cristobal'.


```
[
  {
    "group_id": "Hondarribia Gipuzcoa",
    "latitude": 43.3625,
    "longitude": " -1.7915",
    "master_name": "Hondarribia Gipuzkoa",
    "matched_variants": [
      {
        "confidence_score": 1,
        "name_variant": "Hondarribia Gipuzkoa"
      },
      {
        "confidence_score": 1,
        "name_variant": "Hondarribia Gipuzcoa"
      }
    ]
  },
  {
    "group_id": "San Cristobal de La Laguna Santa Cruz de Tenerife",
    "latitude": 28.4873,
    "longitude": " -16.3159",
    "master_name": "S. Cristobal de La Laguna Santa Cruz de Tenerife",
    "matched_variants": [
      {
        "confidence_score": 1,
        "name_variant": "S. Cristobal de La Laguna Santa Cruz de Tenerife"
      },
      {
        "confidence_score": 1,
        "name_variant": "San Cristobal de La Laguna Santa Cruz de Tenerife"
      }
    ]
  }
]
```


Figure 8 Result example using the 'TF-IDF' model

5.1.2 DATA WRANGLING

Harmonizing the target datasets requires comprehensive data wrangling and analysis. The insights derived from data profiling directly inform the specific transformations needed for each dataset. The MDHESE component facilitates this by fixing character inconsistency, harmonizing dates, removing duplicate entries, normalizing

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

temporal frequencies, and correcting geographic route positions. Each of these capabilities is detailed in the subsequent sections.

CHARACTER INCONSISTENCY

Occasionally, input data suffers from incorrect decoding, resulting in text artifacts commonly known as mojibake. These encoding anomalies typically arise when data is transferred between systems with mismatched character encodings, such as legacy systems using Latin-1 and modern applications expecting UTF-8. If left unresolved, these corrupted Unicode characters introduce significant noise into the dataset, which can severely degrade the performance of downstream processes, including semantic deduplication algorithms and Natural Language Processing (NLP) tasks. To address this, the MDHESE component utilizes the ftfy (Speer, 2024) library, which applies advanced heuristics to automatically detect and repair encoding errors, restoring the proper Unicode formatting. For instance, when processing the corrupted dataset shown in Figure 9, MDHESE generates the corrected JSON output illustrated in Figure 10, ensuring high data quality for subsequent analysis.

```
id;station_name;city;status_message;sensor_type
101;"EstaciÃ³n de Atocha";"Madrid";"Operativo y funcionando al 100&percnt;";"CÃ¡mara TrÃ¡fico"
102;"CabaÃ±as de la Sagra";"Toledo";"Incidencia en el KM 24 &ndash; Carril cerrado";"Sensor AcÃºstico"
103;"PlaÃ§a de Catalunya";"Barcelona";"EstaciÃ³n prÃ³xima a su capacidad mÃ¡xima";"Contador Peatones"
104;"NuÃ±ez de Balboa";"Madrid";"Mantenimiento de seÃ±alizaciÃ³n acÃºstica";"TermÃ³metro"
105;"San SebastiÃ¡n-Donostia";"Gipuzkoa";"VÃa de acceso cortada por mar de fondo";"EstaciÃ³n
MeteorolÃ³gica"
106;"A CoruÃ±a - San CristÃ³bal";"A CoruÃ±a";"Trenes con demora de 15&apos;";"Panel Informativo"
107;"MÃ¼nchen Hauptbahnhof";"MÃ¼nchen";"GlÃ¼ckliche Reise - Buen viaje";"EscÃ¡ner LiDAR"
108;"ZÃ¼rich Flughafen";"ZÃ¼rich";"ConexiÃ³n con autobÃºs nÃºmero 824";"Radar Velocidad"
```

Figure 9 Example of corrupt records


```
[
  {
    "city": "Madrid",
    "id": 101,
    "sensor_type": "Cámara Tráfico",
    "station_name": "Estación de Atocha",
    "status_message": "Operativo y funcionando al 100%"
  },
  {
    "city": "Toledo",
    "id": 102,
    "sensor_type": "Sensor Acústico",
    "station_name": "Cabañas de la Sagra",
    "status_message": "Incidencia en el KM 24 - Carril cerrado"
  },
  {
    "city": "Barcelona",
    "id": 103,
    "sensor_type": "Contador Peatones",
    "station_name": "Plaça de Catalunya",
    "status_message": "Estación próxima a su capacidad máxima"
  },
  {
    "city": "Madrid",
    "id": 104,
    "sensor_type": "Termómetro",
    "station_name": "Nuñez de Balboa",
    "status_message": "Mantenimiento de señalización acústica"
  },
  {
    "city": "Gipuzkoa",
```


Funded by the European Union

Project funded by Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra Swiss Confederation Federal Department of Economic Affairs, Education and Research EAER State Secretariat for Education, Research and Innovation SERI

```
    "id": 105,
    "sensor_type": "Estación Meteorológica",
    "station_name": "San Sebastián-Donostia",
    "status_message": "Vía de acceso cortada por mar de fondo"
  },
  {
    "city": "A Coruña",
    "id": 106,
    "sensor_type": "Panel Informativo",
    "station_name": "A Coruña - San Cristóbal",
    "status_message": "Trenes con demora de 15'"
  },
  {
    "city": "München",
    "id": 107,
    "sensor_type": "Escáner LiDAR",
    "station_name": "München Hauptbahnhof",
    "status_message": "Glückliche Reise - Buen viaje"
  },
  {
    "city": "Zürich",
    "id": 108,
    "sensor_type": "Radar Velocidad",
    "station_name": "Zürich Flughafen",
    "status_message": "Conexión con autobús número 824"
  }
]
```

Figure 10 Results obtained with corrected data

DUPLICATE ENTRIES

The identification and removal of duplicate entries is an important step in data harmonization. Duplicated data introduces critical statistical biases that can affect analytical outcomes. By meticulously deduplicating datasets, the MDHESE component ensures analytical integrity, optimizes processing performance, and guarantees that downstream decision-making is based on accurate, non-redundant information. To achieve this, the MDHESE component leverages the Pandas (NumFOCUS, Inc., 2026) library to allow stakeholders to identify and remove exact duplicate rows.

TEMPORAL FREQUENCIES

Within the mobility domain, dynamic information—such as continuous sensor streams, vehicle geolocation logs, and real-time traffic updates—is predominantly structured as time-series data. Consequently, ensuring the structural consistency of these temporal observations is a critical requirement for accurate cross-platform analysis and predictive modelling.

When integrating these heterogeneous data sources, data normalization is required. In the context of time series, normalizing the integration of two or more streams entails the unification of their temporal indices; i.e., matching the data frequency. By selecting one time series as the baseline reference, the data resampling applied to the remaining series generally occurs in two ways:

- Downsampling: Reducing data frequency (e.g., aggregating from hourly to daily data).
- Upsampling: Increasing data frequency (e.g., interpolating from daily to hourly data).

To manage these operations, the developed component incorporates dedicated temporal normalization capabilities. The system systematically orchestrates these resampling processes, utilizing the Darts Python library (Unit8 SA, 2026) as the underlying computational engine. This package simplifies time series treatment

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

and unifies time series processing and forecasting models. The library has no limitations regarding its licensing (as it is licensed under the Apache License 2.0) or supported types of time series (as it supports both univariate and multivariate time series, with or without static or dynamic covariates). The component specifically extracts and applies the data processing modules of this library to enforce the structural alignment of the indices. Through this built-in functionality, the component evaluates and applies many methods for resampling. In the case of downsampling, the following methods are available:

- *first:* Take the first value in the original time period that spans over the new frequency unit.
- *last*: Take the last value in the original time period that spans over the new frequency unit.
- *max:* Take the maximum value along the original time period that spans over the new frequency unit.
- *median:* Take the median value along the original time period that spans over the new frequency unit.
- *min:* Take the minimum value along the original time period that spans over the new frequency unit.
- *prod:* Take the product (multiplication) of all the values along the original time period that spans over the new frequency unit.
- *quantile:* Take the x-th quantile along all the values of the original time period that spans over the new frequency unit.
- *std:* Take the standard deviation of all the values from the original time period that spans over the new frequency unit.
- *sum:* Take the sum (addition) of all the values along the original time period that spans over the new frequency unit.
- *var:* Take the variance of all the values from the original time period that spans over the new frequency unit.
- *mean:* Take the mean value along the original time period that spans over the new frequency unit.

On the other hand, the following methods are available for upsampling:

- *forwardfill*: Fill forward all the new values of the new time period by repeating the value from the original frequency unit.
- *backfill*: Fill backwards all the new values of the new time period by repeating the value from the original frequency unit.
- *nearest*: Fill the new values of the new time period by repeating the value from the nearest original frequency unit.
- *interpolation*: Fill the new values of the new time period by interpolation of the values from the original frequency units. Different types of interpolation are allowed: linear, splines…

Figure 11 shows a practical example, using GeoSphere Austria datasets. Both plots display the time series in their original frequency, the resampled frequency, and a validation target time series. The graph on the left illustrates hourly to daily downsampling, while the one on the right demonstrates daily to hourly upsampling.

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

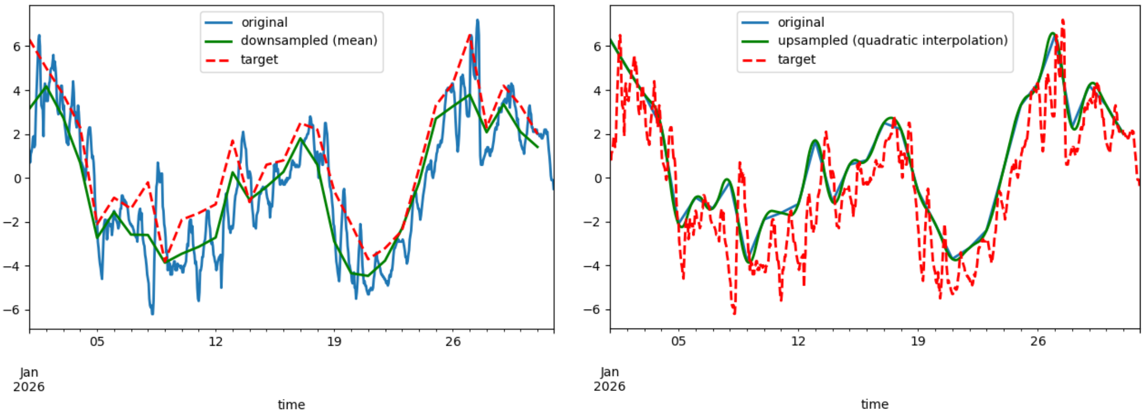

Figure 11 Plots for time series data from GeoSphere Austria with hourly to daily downsampling (left) and daily to hourly upsampling (right). Each plot shows the original time series, the resampled one and a validation target for the resampling.

GEOGRAPHICAL POSITIONS

Within the transport and mobility domain, analysing vehicle trajectories is a critical task. A trajectory consists of a sequence of geolocated positions mapped to specific timestamps, enabling the continuous tracking of a vehicle's location over time. To manage this specialized data, the MDHESE component integrates the MovingPandas (movingpandas) library, offering advanced trajectory generalization and smoothing techniques.

Generalizing a trajectory consists of reducing the number of data points in trajectory objects while preserving their essential characteristics. Three distinct generalization strategies can be applied:

- **Spatial generalization**: applies the Douglas-Peucker algorithm to simplify the trajectory's geometry based on a specified tolerance value.
- **Temporal generalization**: utilizes downsampling to increase the time interval between recorded points, similar to time series data, based on a maximum delta time between points.
- **Spatiotemporal generalization**: employs the Top-Down Time Ratio algorithm to simplify the route based on both spatial and temporal dimensions, depending on a tolerance value.

To illustrate this process, consider a bus trajectory from the NPT Nicosia GTFS dataset in the FEDORA catalogue (see upper left plot from Figure 12), subjected to the three generalization strategies. For spatial generalization, applying a tolerance of 0.001 reduces the original 456 points to 258, while preserving the overall trajectory length of 15.25 km (see upper right plot from Figure 12). In contrast, temporal generalization via downsampling with a one-minute delta time tolerance reduces the data points to approximately 8% of their original volume (see lower left plot from Figure 12). Although this substantial reduction implies a higher loss of information compared to spatial generalization, the essential geometry of the path remains. Finally, spatiotemporal generalization with a 0.001 tolerance yields a slightly smaller reduction (269 points) compared to spatial generalization, while similarly maintaining the 15.25 km route length (see lower right plot from Figure 12).

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

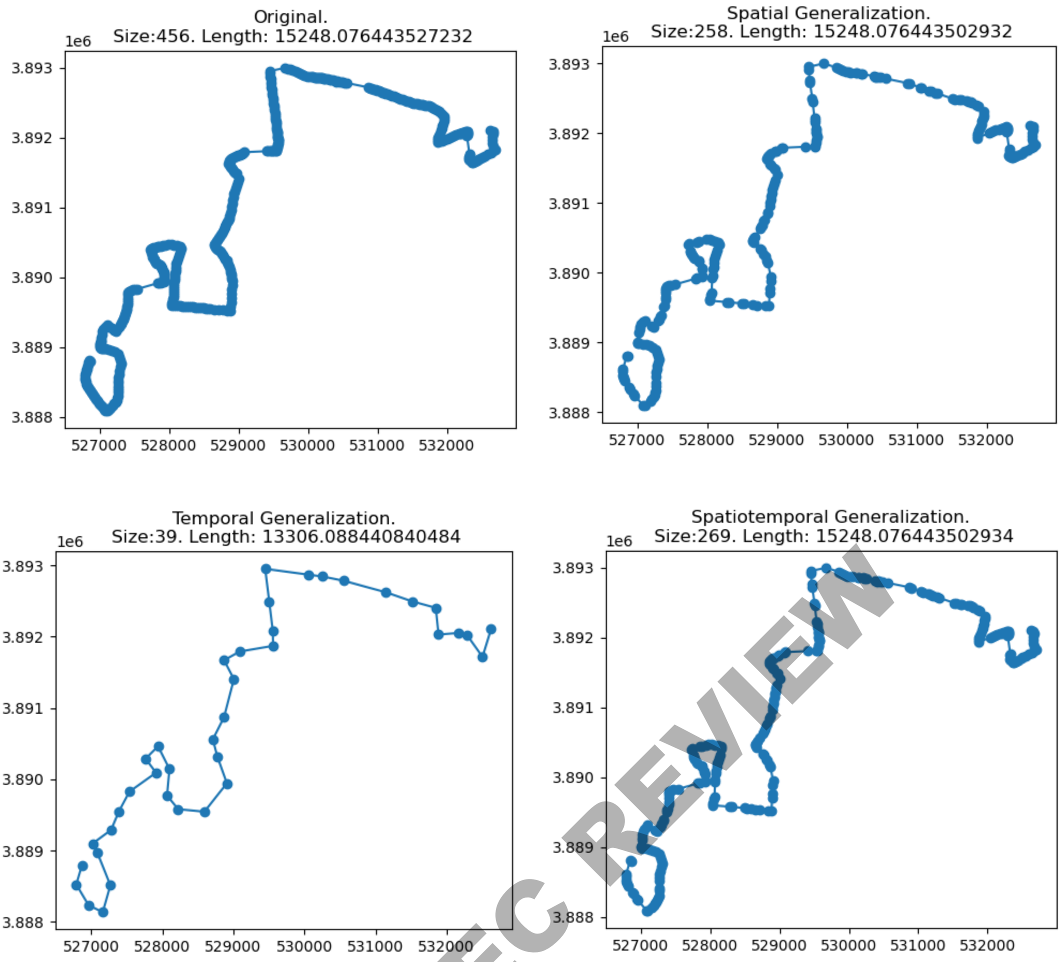

Figure 12 Geographical representations of a bus trajectory on its original shape (upper left), spatially generalized (upper right), temporally generalized (lower left) and spatiotemporally generalized (lower right).

Meanwhile, smoothing the trajectory consists of removing the location point noise. For this purpose, the component applies a Kalman filter based on the assumption of a nearly-constant velocity (CV) model. This model assumes that the velocity in the trajectory is almost constant, where the velocity is indeed the position derivative, and adds some white noise to consider physical relations of real-life scenarios. The filter can be tuned by adjusting the standard deviation for both process and measurement noise:

- **Process noise standard deviation**: governs the uncertainty associated with the trajectory's adherence to the CV model assumption. Higher values relax this assumption, which results in less-smooth trajectories, and vice-versa.
- **Measurement noise standard deviation**: controls the assumed error in the original data points. Higher values indicate that the input trajectory is expected to be noisy and less reliable, thereby prompting the algorithm to produce a smoother output, and vice-versa.

Returning to the previous example, Figure 13 illustrates a smoother and shorter trajectory achieved by setting a process standard deviation of 0.1 and a measurement standard deviation of 10.

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

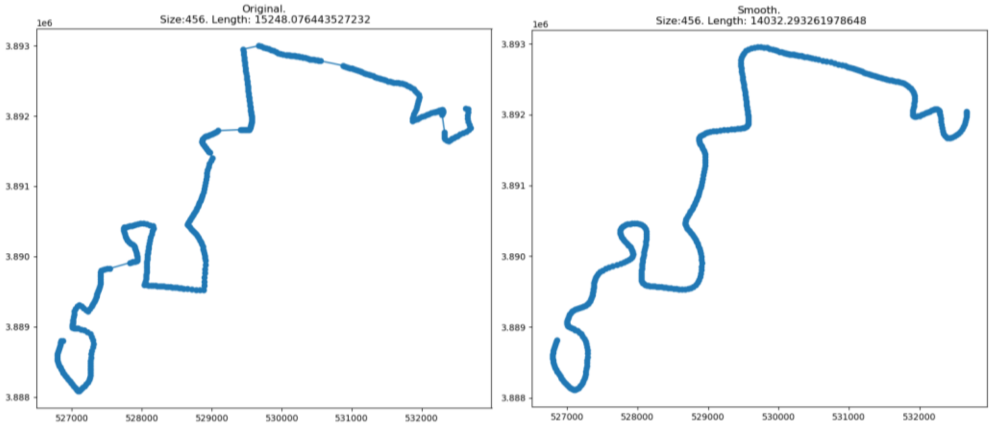

Figure 13 Geographical representations of a bus trajectory on its original shape (left) and smoothed (right).

For optimal smoothing, it can be helpful to first remove outliers. The speed-based outlier cleaner removes anomalous spikes in the trajectory when the speed exceeds a provided threshold value. In the same example as before, applying a threshold value of 2, 14 points are removed from the trajectory (see Figure 14). Subsequent smoothing then results in a shorter and more fluid curve.

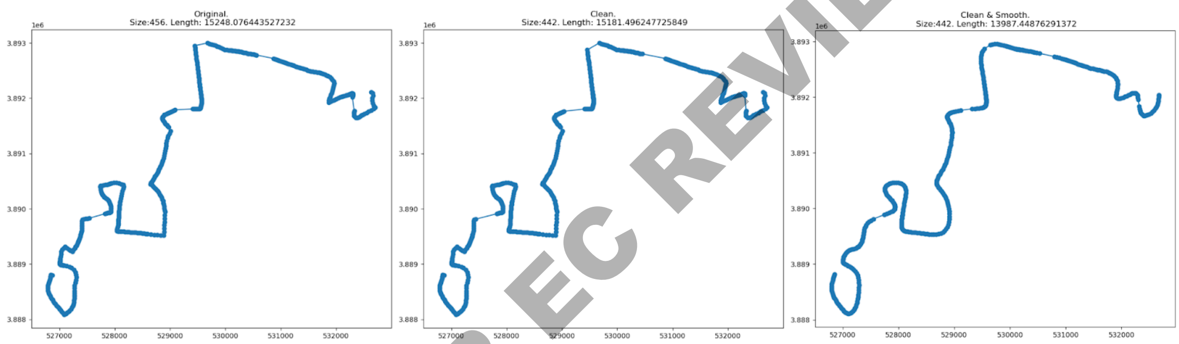

Figure 14 Geographical representations of a bus trajectory on its original shape (left), after outlier cleaning is applied (middle) and cleaned and smoothed (right).

In summary, the MDHESE component enables the harmonization of geographic trajectories through two distinct approaches: generalization (reducing the volume of data points) and smoothing (correcting noisy and erratic measurements).

5.2 DATA ENRICHMENT

Data enrichment is the process of augmenting a dataset by resolving existing inconsistencies and incorporating relevant information from internal or third-party sources. This practice significantly improves data quality, accuracy, and overall usability, laying the foundation for deeper analysis. To achieve this, the MDHESE component approaches enrichment through two main pillars: first, by resolving semantic and syntactic ambiguities to ensure accurate data linkage; and second, by appending valuable contextual information and metadata to the original records.

5.2.1 AMBIGUITY RESOLUTION

In the mobility sector, combining data from multiple sources is a standard practice to enhance datasets for subsequent analysis. However, direct column joins are often unfeasible due to syntactic or semantic discrepancies. A critical and highly common example is the integration of datasets based on public transport station or stop names.

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

This challenge arises from the diverse origins of the data—such as different transport operators, municipal councils, or third-party platforms (e.g., Google Maps, OpenStreetMap)—each employing distinct naming conventions. Consequently, advanced syntactic and semantic matching becomes essential in scenarios including:

- **Station and stop name variations**: reconciling different registered names that refer to the exact same physical location.
- **Address and street name discrepancies**: cross-referencing entities such as traffic accident records, electric vehicle (EV) charging stations, or Low-Emission Zones (LEZs).
- **Operator or fleet consolidation**: unifying reports from various shared mobility providers (e.g., bikes, scooters, car-sharing fleets).
- **Multilingual or multi-format equivalents**: harmonizing records in cross-border initiatives or regions with multiple official languages.

To address these challenges, the MDHESE component leverages the PolyFuzz (Grootendorst, 2022) library. Through robust syntactic and semantic matching, the component successfully enriches the core dataset with valuable external information. The process is the same as described in the previous section, with the option of using the same models for matching.

5.2.2 ADDING CONTEXTUAL INFORMATION AND METADATA

Beyond resolving structural ambiguities, raw mobility data often lacks the necessary context for advanced predictive analysis or seamless cross-border data sharing. A standalone GPS point or a timestamp provides limited value without understanding the vehicle's behaviour or the surrounding environment. To solve this limitation, the MDHESE component enriches the core dataset with additional layers of meaning by integrating internal and external context and standardizing its metadata.

INTERNAL CONTEXT

Trajectory data can also be enriched by deriving new variables directly from the timed locations. For instance, using the example trajectory presented in the 5.1.2 Data Wrangling section, the speed and acceleration of the vehicle can be mathematically deduced and appended to the dataset without requiring any external information sources. To achieve this, the MDHESE component leverages the functionalities provided by the MovingPandas (movingpandas) library. Figure 15 illustrates the vehicle's acceleration (left) and speed (right) throughout the trajectory, calculated exclusively from the dataset's intrinsic timed locations.

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

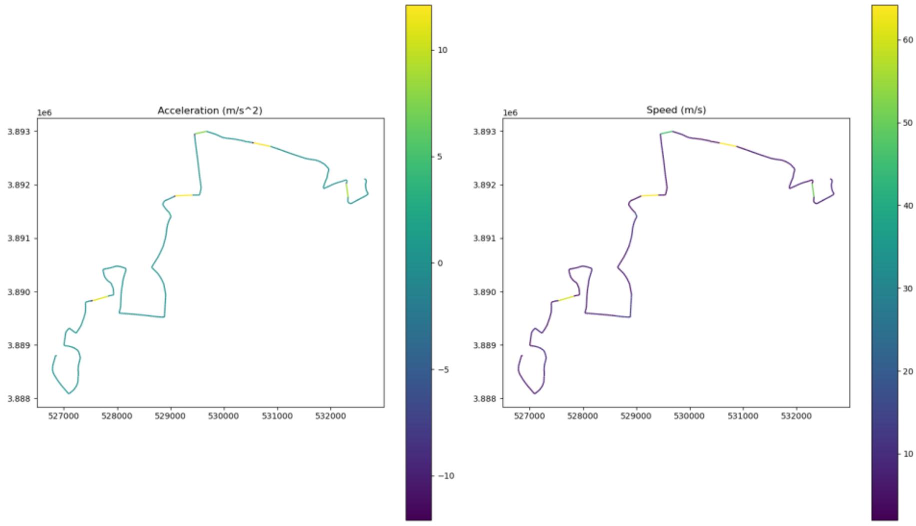

Figure 15 Geographical representations of a bus trajectory with a colour representation of its acceleration (left) and speed (right).

EXTERNAL CONTEXT

As mentioned above, in the context of transportation and mobility, time series data gain relevance because they describe the inherent temporal dimension of movement observations. In that sense, time series can be enriched with covariates, i.e., external variables added to the target variable that give context information and could describe some of the patterns of the principal data. Indeed, in the context of forecasting, covariate series are time series that the models can take as inputs but will not forecast.

The developed component facilitates this enrichment by incorporating dedicated capabilities for covariate processing by the Darts Python (Unit8 SA, 2026) library. Moreover, this package also provides forecasting models where these covariates can be used. Since it integrates a lot of different models, some of them support features that others do not, so they have to be carefully chosen depending on data. Nevertheless, adding contextual information to time series is an easier task. With Darts as its computational engine, the MDHESE component integrates additional context in three distinct ways:

- **Static covariates**: time-invariant external data about the series that can be used by some models to help improve predictions, especially when working with multivariate time series. These can represent metadata like, for example, sensor location, type or ID.
- **Past covariates**: time series whose values are not known into the future at prediction time. Those can for instance represent signals that have to be measured and are not known upfront.
- **Future covariates**: time series whose values are known into the future at prediction time (up until the forecast horizon). These can represent signals such as calendar information, holidays, weather forecasts, etc.

Past, future and static covariates provide additional information/context that can be useful to improve the prediction of the target series. The target series is the variable which is going to be predicted for the future. The covariates are not predicted themselves; they are only used for prediction of the target. As an example, consider a time series with hourly temperature values, rain data as a past covariate and air pressure data as future covariate. Consider also the Gradient Boosted Trees algorithm known as XGBoost model (XGB) for prediction tasks. With the aim of forecasting temperature values for the next day, we apply a XGB prediction model to the time series taking advantage of the prior and future knowledge of the past and future covariates, respectively

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

(see Figure 16). This prediction shows the contribution of external context as covariates when performing forecasts.

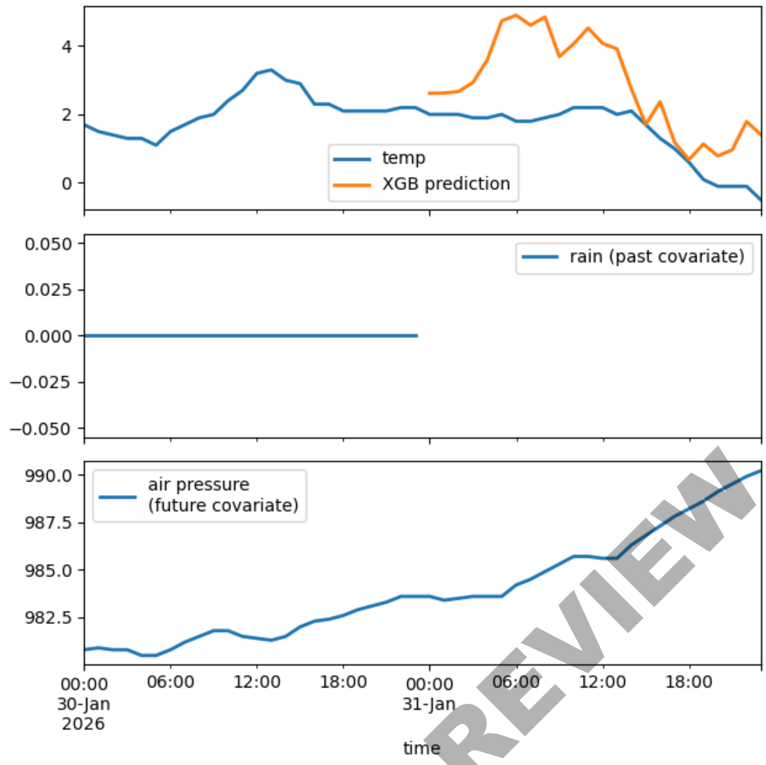

Figure 16 Graph of the temperature time series with its forecast (top), graph of the rain time series (middle) and graph of the air pressure time series (bottom).

Alongside the temporal dimension, mobility data is inherently spatial, requiring its own specific external contextualization. To achieve this, the MDHESE component leverages the GeoPy (GeoPy, 2026) library to perform spatial enrichment through third-party geocoding services. This allows the system to seamlessly translate raw geographic coordinates into human-readable addresses, cities, and administrative boundaries—or vice versa—enriching the dataset with critical real-world geographical context.

The geographic enrichment module uses the Geopy library to interact with the public Nominatim API (OpenStreetMap). During execution, the system transmits text addresses or coordinates outside the local network via secure HTTPS connections. To comply with the provider's usage policy, requests include a custom User-Agent header and enforce a minimum interval of 1.1 seconds between requests, preventing service overload and IP blocks.

To optimise network traffic, the code deduplicates input data using the Polars library, sending only unique records to the API and performing the final join locally. The information sent to the external service is strictly limited to the address or coordinate pair, omitting user identifiers, file names, or system metadata, thereby ensuring the isolation of sensitive information.

During this spatial enrichment, the component extracts two distinct categories of information: direct elements and breakdown fields.

- **Direct elements**: These are universally extracted from the primary location object. They include the latitude and longitude—which are particularly valuable when performing forward geocoding from a place name—and the address, which provides a single, fully formatted, and human-readable text string.

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

* **Breakdown fields**: These consist of granular data points parsed from the full address. They are grouped into specific categories: **Political Geography** (country, country code, state, and county), **Population and City** (city, town, and village), and **Neighbourhoods and Streets** (suburb, neighbourhood, road, house number, and postcode).

It is important to note that the component applies strict semantic sensitivity to the Population and City elements based on accurate demographic classifications. For instance, if the system queries for a 'city' but the coordinates correspond to a smaller settlement, the 'city' field will return a null value, and the data will be accurately populated in the 'town' or 'village' field instead.

The figures below illustrate an execution example of this process. Given the raw dataset (Figure 17) and the corresponding JSON input configuration (Figure 18), the MDHESE component generates the spatially enriched output presented in Figure 19.

```
id;poi_name;sensor_lat;sensor_lon;bike_availability
V01;Wien Hauptbahnhof;48.208174;16.373819;15
V02;Schönbrunn Palace, Vienna;48.198802;16.367401;2
V03;Belvedere, Vienna;48.210611;16.396315;8
```

Figure 17 Raw dataset.

```
{
  "data_filename": "data_mob_vienna.csv",
  "separator": ";",
  "tasks": [
    {
      "type": "geocode",
      "input_column": "poi_name",
      "extract": ["latitude", "longitude", "country"]
    },
    {
      "type": "reverse_geocode",
      "input_columns": ["sensor_lat", "sensor_lon"],
      "extract": ["address", "suburb", "postcode"]
    }
  ]
}
```

Figure 18 JSON input configuration

```
[
  {
    "bike_availability": 15,
    "id": "V01",
    "poi_name": "Wien Hauptbahnhof",
    "poi_name_country": "Österreich",
    "poi_name_latitude": "48.1849883",
    "poi_name_longitude": "16.3779391",
    "sensor_lat": 48.208174,
    "sensor_lat_address": "Stubenviertel, Katastralgemeinde Innere Stadt, Innere Stadt, Wien,
1010, Österreich",
    "sensor_lat_postcode": "1010",
    "sensor_lat_suburb": "Innere Stadt",
    "sensor_lon": 16.373819
  },
  {
    "bike_availability": 2,
    "id": "V02",
    "poi_name": "Schönbrunn Palace, Vienna",
```


Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

```
    "poi_name_country": "Österreich",
    "poi_name_latitude": "48.1849894",
    "poi_name_longitude": "16.3115684",
    "sensor_lat": 48.198802,
    "sensor_lat_address": "Wiedner Hauptstraße, Freihausviertel, Wieden, Wien, 1040, Österreich",
    "sensor_lat_postcode": "1040",
    "sensor_lat_suburb": "Wieden",
    "sensor_lon": 16.367401
  },
  {
    "bike_availability": 8,
    "id": "V03",
    "poi_name": "Belvedere, Vienna",
    "poi_name_country": "Österreich",
    "poi_name_latitude": "48.1939172",
    "poi_name_longitude": "16.3804024",
    "sensor_lat": 48.210611,
    "sensor_lat_address": "19, Schüttelstraße, Pratercottage, Leopoldstadt, Katastralgemeinde
Leopoldstadt, Leopoldstadt, Wien, 1020, Österreich",
    "sensor_lat_postcode": "1020",
    "sensor_lat_suburb": "Leopoldstadt",
    "sensor_lon": 16.396315
  }
]
```

Figure 19 Spatially enriched output

META-DATA

Within federated data ecosystems, interoperability relies strictly on the structural compliance of dataset metadata. To enforce this, the MDHESE component implements automated validation mechanisms, ensuring that the contextual information appended during the enrichment phase conforms to the standard specifications required by European data spaces.

Specifically, the system targets compliance with DCAT-AP, as the foundational metadata baseline for MobilityDCAT-AP datasets. To support this, the component systematically evaluates the associated RDF metadata against predefined SHACL shapes. Operationally, this is executed by interacting directly with the official Metadata Quality Assessment (MQA) SHACL validation API provided by data.europa.eu. The implemented endpoint submits RDF content (e.g., via the text/turtle format) and parameterises the validation profile through the *shapeModel* attribute (e.g., dcatap300level1, dcatap300level2 or dcatap300level3). This configuration allows the system to enforce different tiers of structural compliance by selecting from the following hierarchical levels:

- **Level 1**: Validates strictly against the foundational (Base) mandatory elements of the specification.
- **Level 2**: Extends the validation to enforce compliance with official controlled vocabularies alongside the base rules.
- **Level 3**: Incorporates the validation of recommended properties in addition to base rules and vocabularies.
- **Full Profile** (All shapes / no level): The most exhaustive configuration, evaluating the base structure, vocabularies, recommended properties, and specific data ranges.

Following the API execution, the component processes the SHACL validation output to function as a semantic data governance engine. It aligns the technical response with the European Union's Metadata Quality Assessment (MQA) methodology. Rather than merely validating the RDF syntax, the system systematically categorises the findings across five operational dimensions: Findability, Accessibility, Interoperability, Reusability, and Contextuality. By mapping specific DCAT-AP properties to these dimensions—for instance, evaluating titles and descriptions under Findability, or tracking usage licences under Reusability—the

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

component computes an objective, quantitative score for the dataset. Ultimately, this categorisation identifies exact metadata gaps, establishing a clear metric for semantic governance and ensuring the dataset meets the structural prerequisites for cross-border interoperability

The following figure, Figure 20, illustrates a compliant JSON output from the validation endpoint, confirming the structural alignment of the metadata with the selected SHACL profile:


```
{
  "quality_summary": {
    "conforms": true,
    "dimensions_scores": {
      "Accessibility": 100,
      "Contextuality": 100,
      "Findability": 100,
      "Interoperability": 100,
      "Reusability": 100
    },
    "global_score": 100,
    "message": "The file perfectly conforms to the DCAT-AP profile.",
    "violations_count": 0
  },
  "status": "success"
}
```


Figure 20 Compliant JSON output from the validation module

In contrast, the Figure 21 presents a non-compliant JSON output, detailing the specific schema violations and metadata gaps identified during the evaluation. This JSON snippet illustrates a failed compliance check (conforms: false). The 'Findability' score is reduced to 60 (out of 100) because two cardinality violations were detected: both the *description* and *title* fields are missing (count = 0). As a result, 2 total violations are recorded, resulting in a global score of 92:


```
{
  "quality_summary": {
    "conforms": false,
    "detailed_issues": {
      "Findability": [
        {
          "field": "description",
          "message": "minCount[1]: Invalid cardinality: expected min 1: Got count = 0",
          "severity": "Violation"
        },
        {
          "field": "title",
          "message": "minCount[1]: Invalid cardinality: expected min 1: Got count = 0",
          "severity": "Violation"
        }
      ]
    },
    "dimensions_scores": {
      "Accessibility": 100,
      "Contextuality": 100,
      "Findability": 60,
      "Interoperability": 100,
      "Reusability": 100
    },
    "global_score": 92,
    "violations_count": 2
  },
```


Funded by the European Union

Project funded by Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER State Secretariat for Education, Research and Innovation SERI

```
  "status": "warning"
}
```

Figure 21 Non-compliant JSON output

5.3 DATA SYNTHETIZATION

5.3.1 OVERVIEW OF SYNTHETIC DATA GENERATION

Synthetic data generation refers to the algorithmic creation of artificial datasets designed to accurately preserve the statistical properties, underlying relationships, and structural characteristics of real-world data, without containing any actual or identifiable personal observations. In the context of modern data ecosystems, this process has become an essential capability to overcome critical challenges related to data scarcity, strict privacy constraints, and the necessity for rigorous system testing prior to production deployment.

In order to evaluate the quality of the synthetic data, the similarity to the real data that wants to be imitated will be assessed by measuring the difference between the actual and the created data (for example, by measuring the Root Mean Square Error – RMSE).

Within the mobility domain, the reliance on sensitive and highly granular information—such as passenger travel logs, vehicle geolocation coordinates, and infrastructural capacities—presents significant barriers to open data sharing and collaborative system development. Synthetic data provides a robust and compliant solution to these barriers. By generating structurally representative data, complex urban mobility scenarios can be safely modelled without exposing actual user behaviour. This approach accommodates the inherent heterogeneity of mobility data formats, enabling the simulation of both dynamic and static data types, including continuous sensor data streams, complex network graphs representing transport infrastructures, and real-time traffic information.

As a result, the synthetisation of these diverse datasets enables the continuous validation of analytical models and data-driven mobility applications, ensuring robust interoperability across different platforms whilst strictly adhering to European data protection regulations. The Folkwang sections detail the specific techniques and tools utilised for this synthetic generation, as well as their direct integration with dynamic and static mobility data types.

5.3.2 TECHNIQUES AND TOOLS FOR SYNTHETIC DATASET CREATION

Within the scope of this project, the generation of synthetic data serves a dual operational purpose. Firstly, synthetic data techniques can be utilised for data imputation within incomplete datasets; to address this requirement, this section will explore advanced methodologies relying on tensor-based operations to accurately reconstruct and estimate missing information. Secondly, synthetic generation is employed to create entirely

Funded by the European Union
Project funded by
Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

complete test datasets from scratch. To facilitate this second objective, the developed MDHESE component incorporates a dedicated functionality that enables the automated creation of structurally representative test datasets, ensuring robust system validation even when production data is restricted or unavailable.

5.3.2.1 TENSOR-BASED METHODOLOGIES FOR DATA IMPUTATION

Spatiotemporal traffic (speed, flow, occupancy, density) and mobility data (timestamped origin–destination trips) are high-dimensional, incomplete, and noisy, yet exhibit strong multiway dependencies across networks and time. Low-rank tensor learning (LRTL) exploits this latent structure to enable completion at scale.

Tensors are multidimensional generalizations of matrices which provide a natural representation for structured data. LRTL leverages intrinsic low-dimensional structure by injecting domain knowledge as spatial and temporal constraints so that models respect network topology and recurrent temporal patterns.

- **LRTL** techniques can be used in many tasks: imputation, spatiotemporal extrapolation, temporal prediction/forecasting, pattern discovery/anomaly detection, and regression. In the case of data imputation, two studies which provide the code for its algorithms have been tried: ST-LRTC from Zhao (2023) and GLSKF from Lei and Sun (2024).
- **ST-LRTC** (Spatio-Temporal constrained Low-Rank Tensor Completion) depicts the local geometric structure of spatiotemporal domains. Specifically, under the low-rank assumption, the method introduces temporal constraints based on the continuity and periodicity of traffic flow and a spatial constraint matrix reflecting the traffic flow transmission mechanism.
- **GLSKF** (Generalized Least Squares Kernelized Tensor Factorization) captures global and long-range correlations by a low-rank, covariance-constrained/smoothed tensor factorization; and captures short-scale, high-frequency variations in the data by a correlated residual module.

Using vehicle flow data recorded by induction loops from a northern Spanish city, ST-LRTC and GLSKF have been tried for imputing data missingness. This dataset comprises 339 induction loops for a month (January 2023) with a frequency of a minute, which makes a total of 44640 records for each location. Thus, the data tensor has a shape of 339 x 31 x 1440. Note that, in this case, a spatial constraint matrix could be computed for ST-LRTC. See results on Table 16 for single runs of 20 iterations with each combination of type of missingness, proportion of missingness and algorithm used. In this case, there are no data splits for training and testing, since these algorithms do not rely on any training as Machine Learning and other types of models do.

Table 16 Process time and error of each imputation by type of missingness, proportion of missingness and algorithm used.

# DAYS	TYPE OF MISSINGNESS	MISSINGNESS	ITERATIONS	ALGORITHM	TIME (MIN)	RMSE (VEH/H)
31	Random Missing (RM)	90%	20	ST-LRTC	11	202.67
				GLSKF	10	60.02
		50%	20	ST-LRTC	12	63.71
				GLSKF	9	56
		10%	20	ST-LRTC	12	18.67
				GLSKF	9	55.21

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

	Dimension 0 Missing (D0M) - Entire locations missing	90% (305 ind. loops)	20	ST-LRTC	12	318.58
				GLSKF	12	305.15
		50% (170 ind. loops)	20	ST-LRTC	12	301.94
				GLSKF	13	225.89
		10% (34 ind. loops)	20	ST-LRTC	11	210.5
				GLSKF	10	206.36
	Dimension 1 Missing (D1M) - Entire day missing among every location	90% (28 days)	20	ST-LRTC	12	322.87
				GLSKF	9	264.73
		50% (15 days)	20	ST-LRTC	11	314.89
				GLSKF	9	147.68
		10% (3 days)	20	ST-LRTC	13	310.9
				GLSKF	10	149.96
	Dimension 2 Missing (D2M) - Same time of the day missing among every location and day	90% (1296 timestamps)	20	ST-LRTC	12	321.34
				GLSKF	9	369.93
		50% (720 timestamps)	20	ST-LRTC	12	315.88
				GLSKF	10	325.68
		10% (144 timestamps)	20	ST-LRTC	12	304.5
				GLSKF	10	315.53
	Dimension 0 Partial Missing (D0PM) - In every location same time of the day missing	90%	20	ST-LRTC	11	294.73
				GLSKF	11	954.82
		50%	20	ST-LRTC	13	225.73
				GLSKF	10	66.13
		10%	20	ST-LRTC	11	114.05

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

				GLSKF	9	71.17
	Dimension 1 Partial Missing (D1PM) - Everyday same time of the day missing in each location	90%	20	ST-LRTC	11	286.86
				GLSKF	11	387.98
		50%	20	ST-LRTC	12	238.84
				GLSKF	11	58.62
		10%	20	ST-LRTC	11	204.45
				GLSKF	11	56.29
	Dimension 2 Partial Missing (D2PM) - Different entire days missing among every location	90%	20	ST-LRTC	12	302.19
				GLSKF	14	235.93
		50%	20	ST-LRTC	13	229.24
				GLSKF	11	69.47
		10%	20	ST-LRTC	12	107.18
				GLSKF	11	58.24

The first conclusion is that both ST-LRTC and GLSKF last the same time for the same number of iterations: around 10 minutes for 20 iterations, regardless of the percentage of missing data.

The first type of missingness addressed is uniformly random missing across all the datasets. These imputations are getting the best results comparing with missingness by patterns. GLSKF behaves similarly in both three proportions of missing data, with an error between 55 and 60 veh/h. ST-LRTC, however, changes a lot its performance: the less the missing data, the better the RMSE. Moreover, its best result at 10% of missing data outperforms GLSKF.

Taking care of patterns of missingness, entire data dimensions missing were first handled. As there are three dimensions in this tensor, three types of dimension missing are managed:

- Dimension 0 Missing (D0M) - Entire locations missing. Both algorithms behave similarly (around 200-300 veh/h of RMSE), although GLSKF works better whatever the number of induction loops missing.
- Dimension 1 Missing (D1M) - Entire day missing among every location. In each case the difference between ST-LRTC and GLSKF is larger, as the former gets an error between 310 and 320 veh/h in both three amounts of days missing while the latter 265 veh/h for 28 days missing but near 150 veh/h for 15 and 3 days.
- Dimension 2 Missing (D2M) - Same time of the day missing among every location and day. Even if in both three proportions of missing the error is between 300 and 370 veh/h, in this case GLSKF gets higher errors than ST-LRTC.

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

These missing patterns could happen not only in the entire dimension, but also partially:

- Dimension 0 Partial Missing (D0PM) - In every location same time of the day missing. GLSKF gets an extremely big error for 90% of missing data, but great results of around 70 veh/h for 50% and 10%. ST-LRTC is more stable, but with errors ranging between 110 and 300 veh/h.
- Dimension 1 Partial Missing (D1PM) - Everyday same time of the day missing in each location. Once again, GLSKF gets good results for 10% and 50% of missingness, and a lot worse result for 90%, while ST-LRTC stabilizes between 200 and 290 veh/h of RMSE.
- Dimension 2 Partial Missing (D2PM) - Different entire days missing among every location. In this case GLSKF also performs much better with 50% and 10% of data missing but gets a better result than ST-LRTC for 90%. Indeed, ST-LRTC changes more between the three proportions than before.

Taking these results into account, it can be said that when missingness is randomly distributed across data in higher proportions, GLSKF performs better than ST-LRTC. When entire locations or entire days are missing, GLSKF also performs better, no matter the proportion of missingness. Nevertheless, when a specific time of the day is repeatedly missing, ST-LRTC works better. Finally, in partial missing cases, GLSKF gets again higher performance, especially with 10% and 50% of missing proportions, with acceptable errors below 100 veh/h. When the missing percentage is higher, ST-LRTC gets better results, probably because of its use of spatial correlation.

5.3.2.2 AUTOMATED GENERATION OF STRUCTURALLY REPRESENTATIVE TEST DATASETS

In various development and testing scenarios, synthetic data that closely mimics real-world conditions is required to validate system behaviour. To address this, the MDHESE component incorporates a dedicated functionality for generating realistic test datasets, implemented through the Pointblank (Posit, 2026) toolkit. This functionality enables the generation of test data based on precise schema definitions, which is particularly useful for prototyping workflows when production data is unavailable.

The module allows for the creation of synthetic datasets adapted to different regional configurations, offering support for 100 countries. It also provides standard presets to generate common data fields—such as names, email addresses, physical addresses, and 'user agent' strings—whilst enforcing strict field constraints, including numerical ranges, specific text patterns, and value uniqueness. Furthermore, in the context of the mobility domain, the use of time-series datasets (such as sensor data and real-time traffic information) is highly frequent. To accommodate this requirement, the component includes the capability to generate synthetic datasets by defining the specific temporal frequency of the series, ensuring the resulting data accurately reflects the chronological structure of mobility events.

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

6 CONCLUSION

This report has described the context, high-level requirements, use case and architecture of the FEDORA Data Space, covering the RM-ODP enterprise viewpoint. It has shown how its components work together to support use cases for data exchange between the project partners, and how data space technologies can be applied to enable these use cases. The FEDORA Public Catalogue is integrated into the data space and extends its capabilities by facilitating search and access to publicly available datasets. As a value adding service, the Mobility Data Harmonisation, Enrichment and Synthetisation Engine adds functionality to ensure and improve data quality and to enrich datasets.

Covering the information viewpoint, Chapter 3 has described the typology of the data relevant to the project, as well as the applicable metadata standards, data formats and standards for interoperable interfaces. Chapter 4 has described the functionality of the FEDORA Public Catalogue, the way metadata can be ingested and harvested and the way project partners and the public can use it to find and show information about the datasets they need. Chapter 5 has introduced the methodology for Harmonisation, Enrichment and Synthetisation and the tools to be used for the implementation.

Being part of Milestone 3 SPARK, this document describes the high-level concepts of the FEDORA Data Space and the first version of its Catalogue and MDHESE components. In the next stage, based on the concepts laid out in D3.1 and D3.2, the Data Space components will be further developed and rolled out. The project partners will be asked to join the FEDORA Data Space, not only giving them access to the search, exchange and value adding functionalities, but also allowing for a more bottom-up approach to align the data space concepts with the data needs of the pilots and the Network Management Service and Simulation Spaces.

Funded by the European Union

Project funded by Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER State Secretariat for Education, Research and Innovation SERI

7 REFERENCES

Bozzi, G., De Cock, L., Van Gheluwe, C., & Pretzsch, S. (2025). DeployEMDS D2.3 Technical specifications report.

Chevallier, T., Obregon Gonzalez, R., Delmas, V., Randriarison, J., Pham, S., & Cheliotis, K. (2026). D3.2 AI-Enhanced Data Utilization and Smart Contracting.

CKAN. (2026). ckanext-dcat. *Retrieved 08 2026, from https://github.com/ckan/ckanext-dcat*

CKAN. (2026). ckanext-harvest - Remote harvesting extension. *Retrieved 08 2026, from https://github.com/ckan/ckanext-harvest*

CKAN. (2026). CKANEXT-spatial. *Retrieved from https://github.com/ckan/ckanext-spatial*

CKAN. (n.d.). CKAN Organizations and authorization. *Retrieved from https://docs.ckan.org/en/2.11/maintaining/authorization.html*

CKAN. (n.d.). ckanext-harvest-basket . *Retrieved from https://datashades.github.io/ckanext-harvest-basket/*

Commission, T. |. (2025). MobiSpaces - New data spaces for green mobility. *Retrieved from https://trimis.ec.europa.eu/project/new-data-spaces-green-mobility*

COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS A European strategy for data. *(2020, Sep 19). Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0066*

COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Creation of a common European mobility data space. *(2023, Nov 29). Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2023:751:FIN*

CORDIS | European Commission. (n.d.). New data spaces for green mobility. *Retrieved from https://cordis.europa.eu/project/id/101070279*

Daels , N., De Cock , L., Selita , E., & Van Gheluwe , C. (2024). deployEMDS D2.1 Requirements analysis of the technical infrastructure. https://deployemds.eu/wp-content/uploads/2024/05/D2.1-Requirements-analysis-of-the-technical-infrastructure-3.pdf.

Data Space Support Centre. (2026). Data Space Support Centre. *Retrieved 08 2026, from https://dssc.eu/*

Data Spaces Support Centre. (n.d.). DSSC Blueprint. *Retrieved 08 2026, from https://blueprint.dssc.eu/20*

Data Spaces Support Centre. (n.d.). DSSC Blueprint Glossary. *Retrieved 08 2026, from https://blueprint.dssc.eu/?pane=glossary&glossary=1-key-concept-definitions*

DATEX II. (2026). DATEX II - European standard for traffic and travel information. *Retrieved from https://datex2.eu/*

Datos.gob.es. (2026). Technical guidance and model for DCAT-AP-ES Application Profile implementation. *Retrieved from https://datosgobes.github.io/DCAT-AP-ES/en/*

Digitale, A. p. (2022). DCAT-AP_IT v1.1 – Profilo italiano di DCAT-AP. *Retrieved from https://www.dati.gov.it/content/dcat-ap-it-v10-profilo-italiano-dcat-ap-0*

Funded by the European Union

Project funded by Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra Swiss Confederation Federal Department of Economic Affairs, Education and Research EAER State Secretariat for Education, Research and Innovation SERI

Doulkeridis, C., Santipantakis, G. M., Koutroumanis, N., Makridis, G., Koukos, V., Theodoropoulos, G. S., . . . Jimenez-Peris. (2023). MobiSpaces: An Architecture for Energy-Efficient Data Spaces for Mobility Data. IEEE International Conference on Big Data (BigData)*, 1487-1494.*

EC. (2020). A European strategy for data. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS*.*

EC. (2023). Creation of a common European mobility data space. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS*.*

Eclipse Contributors. (2026). Eclipse Dataspace Components*. Retrieved 08 2026, from https://eclipse-edc.github.io/*

Eclipse-tractusx. (2026). PURIS*. Retrieved from https://github.com/eclipse-tractusx/puris*

EP. (2022, May). REGULATION (EU) 2022/868 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. Official Journal of the European Union*. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868*

EP. (2023). REGULATION (EU) 2023/2854 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. Official Journal of the European Union*.*

European Committee for Standardization. (2024). Trusted Data Transaction. CEN Workshop Agreement*.*

Fafoutellis, P., Vakrinou, K., Vlahogianni, E., Stephens, J., Carrapichano, Z., Sydenham, A., . . . Menelaou, C. (2025). D2.1 Best Practices & State-of-the-art.

FITKO (Föderale IT-Kooperation). (n.d.). DCAT-AP*. Retrieved from https://www.dcat-ap.de/*

Fragkou, P. (2023). DCAT-AP High Value Datasets*. Retrieved from https://semiceu.github.io/DCAT-AP/releases/2.2.0-hvd/*

GeoPy. (2026). GeoPy*. Retrieved from https://geopy.readthedocs.io/en/stable/#module-geopy.geocoders*

GO FAIR. (n.d.). FAIR Principles*. Retrieved 08 2026, from https://www.go-fair.org/fair-principles/*

Grootendorst, M. (2022). Polyfuzz*. Retrieved from https://maartengr.github.io/PolyFuzz/*

International Data Spaces Association. (2025). IDS RAM 4*. Retrieved from https://github.com/International-Data-Spaces-Association/IDS-RAM_4_0*

International Data Spaces Association. (2026). IDSA Rulebook*. Retrieved 08 2026, from https://kb.internationaldataspaces.org/external/rulebook/001_Introduction/*

International Data Spaces Association. (2026). International Data Spaces Association*. Retrieved 08 2026, from https://internationaldataspaces.org/offers/reference-architecture/*

International Organization for Standardization (ISO). (2015). ISO/IEC DIS 20151. https://www.iso.org/obp/ui/en/#iso:std:iso-iec:20151:dis:ed-1:v1:en.

International Organization for Standardization (ISO). (n.d.). ISO/IEC 10746. https://committee.iso.org/sites/jtc1sc7/home/projects/flagship-standards/isoiec-10746.html.

Interoperable Europe Portal | European Commission. (n.d.). Norway: Implementation of DCAT-AP-NO*. Retrieved from https://interoperable-europe.ec.europa.eu/collection/access-base-registries/solution/abr-bregdcatap-tools/norway-implementation-dcat-ap-no*

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

Lubrich, P. (2025). A set of Quality Frameworks for the assessment of NAP (meta)data quality. Retrieved from https://napcore.eu/wp-content/uploads/2025/06/NAPCORE-M3.8-deliverable_v1.1_250408.pdf

Lubrich, P. (2025). Quality Framework Application Guideline. Retrieved from https://napcore.eu/wp-content/uploads/2025/07/NAPCORE-M3.9-deliverable_v1_250516.pdf

Ministerio para la Transformación Digital y de la Función Pública a través de la Entidad Pública Empresarial Red.es. (2024). DCAT-AP y sus extensiones.

Mobility Data Space. (2026). Mobility Data Space Wiki. *Retrieved from https://github.com/Mobility-Data-Space/mobility-data-space/wiki*

mobilityDCAT-AP. (2026). mobilityDCAT-AP Wiki. *Retrieved from https://github.com/mobilityDCAT-AP/mobilityDCAT-AP/wiki/Chapter-1:-Introduction-and-Overview*

MobiSpaces. (2022). MobiSpaces Technologies. *Retrieved from https://mobispaces.eu/technologies*

MobiSpaces. (2022). MobiSpaces, New Data Spaces for Green Mobility. *Retrieved from https://mobispaces.eu/news/mobispaces-new-data-spaces-green-mobility*

movingpandas. (n.d.). MovingPandas. *Retrieved from https://movingpandas.org/*

NumFOCUS, Inc. (2026, Jul 22). pandas documentation. *Retrieved from https://pandas.pydata.org/docs/*

Obregon, R., Chevallier, T.-X., Cheliotis, K., & Ziogou, M. (2025). D1.2 Data Management Plan and Ethics.

Posit. (2026). Pointblank. *Retrieved from https://github.com/posit-dev/pointblank*

Pretzsch, S., Drees, H., & Rittershaus, L. (2022). Mobility Data Space. In Designing Data Spaces *(pp. 343-361). Springer, Cham. doi:https://doi.org/10.1007/978-3-030-93975-5_21*

Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act). *(2022, May 30). Retrieved from https://eur-lex.europa.eu/eli/reg/2022/868/oj*

Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act) . *(2023, Dec 22). Retrieved from https://eur-lex.europa.eu/eli/reg/2023/2854*

Sáenz de Ugarte, J., Cortés, S., & Cheliotis, K. (2026). D2.2 Living Labs Setup, Spaces Definition, and Evaluation Models. https://fedora-horizon.eu/repository/.

Speer, R. (2024). ftfy: fixes text for you. *Retrieved from https://ftfy.readthedocs.io/en/latest/*

Unit8 SA. (2026). Darts - Time Series Made Easy in Python. *Retrieved from https://unit8co.github.io/darts/index.html*

W3C. (2024, August 22). Data Catalog Vocabulary (DCAT) - Version 3. *Retrieved from https://www.w3.org/TR/vocab-dcat-3/*

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

ANNEX

SOFTWARE REPOSITORY REFERENCES

IDENTIFIER	DESCRIPTION	REPOSITORY	VERSION / COMMIT
SW-01	Project source code	https://github.com/fedora-horizon	Release v1.0
SW-02	Data Space documentation (README)	https://fedora-horizon.github.io/docs/	Main branch (accessed version)

DATA SOURCES INVENTORY

Table 17: Sources inventory extract

ORGANIZATION NAME / DATASET PUBLISHER	SOURCE TYPE	SPATIAL COVERAGE	FORMATS	STANDARD / SPECIFICATION FOR METADATA SHARING	ACCESS POLICY	TYPE OF CATALOGUE
NAP Austria - MobilityData AustriaTech	NAP ITS	Vienna, Austria	XML, RSS, XLSX, GTFS, NeTEx, mobilityDCAT-AP, DATEX II	mobilityDCAT-AP, DATEX II	open access	(meta)data catalogue
Open Data Portal - Austria (Datagovat)	NAP	Vienna, Austria	CSV, GML, GIF, JPEG, SVG, PNG, GIF, JSON, LIDAR, GIS, WFS (SHP, GeoRSS), WMS (KML, KMZ)		open access	CKAN
ÖBB Open data	Transport/logistics operator	Austria	TXT, GTFS		open access	CKAN / API
Statistik Austria INSPIRE search service	Statistical Institute/portal	Vienna, Austria	GIS	INSPIRE, OGC	open access	Geoportal: GeoNetwork
Open Data Euskadi	Data portal	Regional, Basque Country, Spain	CSV, XLSX, XML, JSON, TSV, RDF, ZIP, PDF,	REST, SPARQL, DCAT-RDF	open access	(meta)data catalogue

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

			ODS, ATOM, REST, SPARQL, DCAT-RDF			
Open Data Bizkaia	Data portal	Biscay (Basque Country), Spain	ATOM, CSV, JSON, TSV, XLSX, XML, ZIP, WFS, WMS, WMTS, KML		open access	(meta)data catalogue
Gipuzkoa Irekia	Government portal	Gipuzkoa (Basque Country), Spain	CSV, DOC, DOCX, ODT, ODS, PDF, TXT, XLS, XLSX	DCAT, RDF	open access	(meta)data catalogue
OTEUS – statistical information	Statistical Institute/portal	Basque Country, Spain	CSV, PDF, EXCEL		open access	catalogue
Cyprus National Access Point (CyNAP)	NAP ITS	Cyprus	mobilityDCAT-AP, XML, GTFS, GTFS-RT, NeTEx, SIRI, XML, DATEX II V3, TN-ITS	XML, DATEX II V3, TN-ITS	open access	CKAN
National Open Data Portal Cyprus	NAP	Nicosia, Cyprus	CSV, XLSX	DCAT-AP	open access	DKAN
Cyprus Department of Civil Aviation (DCA)	Aviation authority	Cyprus			open access	website
NextBike	Micromobility service provider	International (Cyprus)	XML		open access (partially)	system data
NAP Denmark (Dataudvekslere n)	NAP ITS	Denmark	CSV, EXCEL, TXT, PNG, JSON, GML2, GML, Retsakt-B, SIRI, INSPIRE, DATEX II and DATEX II with	INSPIRE, DATEX II and DATEX II with definitions	open access	catalogue

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

			definitions, SHP, WMS, WFS			
Open data DK - City of Copenhagen	Data portal	Copenhagen, Capital Region of Denmark, Denmark	CSV, HTML, XYZ, MTS, RAR, ZIP, SHP, WFS, WMS, GeoJSON		open access	CKAN
Road Directorate (vejdirektoratet)	Road authority	Denmark	PDF	REST	open access (partially)	API, website (partially catalogue
Technical University of Denmark (DTU)	University research group/ institution/department	Copenhagen, National, Denmark				
Reggio Emilia Opendata	Data portal	Reggio Emilia, Emilia-Romagna, Italy	CSV, JSON, ZIP, XLSX, TXT, GML, CSW, KML, GeoJSON, WMS	CSW	open access	CKAN
Agency for Digital Italy	Digital government authority	Italy	CSV, JSON, PDF, DCAT-AP, RDF, GeoJSON, SHP, TIF, KMZ	DCAT-AP, RDF	open access (mostly open)	CKAN 2.6.8
ISPRA	Institute for Environmental Protection & Research	Italy	XLSX, LOD (SPARQL queries, RDF dumps), GIS	LOD (SPARQL queries, RDF dumps)	open access (mostly open)	website
Arpea	Environment Energy Prevention Agency	Reggio Emilia, Emilia-Romagna, Italy	ARC, CSV, PDF, ZIP, GRIB, REST, WFS, WMS, WebGIS	REST	open access (partially)	CKAN
magyar Nemzeti Adathozzáférési Pont (NA) - Hungarian National Data Access Point (NAP)	NAP ITS	Budapest, Central Hungary, Hungary	XML, JSON, DATEX, FWS, TN-ITS, WFS, ShapeZIP, GML2, GML3, GeoJSON, KML and	DATEX, FWS, TN-ITS	open access	catalogue

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

			GeoPackaga e f			
OpenData Portal Centre for Budapest Transport (BKK)	Data portal	Budapest, Central Hungary, Hungary	TEXT, GTFS, GTFS, GTFS-RT	GTFS, GTFS-RT	open access	catalogue / API
National Public Data Portal /Nemzeti Közadatportál	NAP	Hungary	CSV, XLS, XLSX, PDF, ZIP, DATEX II, RDF, JSON, TTL	DATEX II, RDF, JSON, TTL	open access	CKAN
EuRIS	River Information Services platform	Europe	GeoJSON		open access (mostly open)	catalogue / API
Cities - Waze Connected Citizens Program (CCP)	Traffic data sharing program	Internation al	XML or JSON		open access (partially: API partnership s)	catalogue
TomTom	Navigation & traffic data provider	Internation al (Hungary, Cyprus)	DATEX II	DATEX II	commercial licensing	data services / API
HERE	Geospatial & navigation data provider	Internation al	XML, JSON		commercial licensing	data platform / API
Octo Telematics	Telematics & mobility data/analytics company				commercial data platform	
Carto	Geospatial analytics/mapping company				premium (commercia l)	(meta)data catalogue, service
OpenWeather	Weather data provider				fremium	catalogue / API

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

DATA AND METADATA QUALITY IN THE CONTEXT OF THE GROWING NUMBER OF NATIONAL ACCESS POINTS IN EUROPE

Context and challenges

The gradual implementation of National Access Points (NAPs), in accordance with the Intelligent Transport Systems (ITS) Directive and its delegated regulations, has led to a significant increase in the number of mobility data platforms and catalogues within the European Union.

In this context, the mere availability of data is not sufficient to guarantee interoperability, cross-border reuse, or regulatory compliance. The quality of data and metadata is therefore a key factor in ensuring the consistency of mobility data and supporting the deployment of reliable, interoperable, and automated digital services on a European scale.

NAPCORE's role in quality harmonization

To address these challenges, NAPCORE serves as a European coordinating organisation, supporting and harmonising the implementation of NAPs by Member States. In this context, NAPCORE published the "Quality Framework Application Guideline" in May 2025, which establishes a common reference framework for assessing the quality of data, metadata, and services accessible via NAPs.

This framework aims to:

- Make quality **measurable, transparent, and comparable** across Member States;
- Enhance **transparency for re-users;**
- Support **compliance with regulatory requirements** arising from ITS delegated regulations;
- Hold **all stakeholders involved** in the data value chain **accountable;**
- Encourage **continuous quality improvement**, rather than a strictly ad hoc approach to monitoring.

Shared responsibility among stakeholders

Building on this common foundation, NAPCORE promotes a quality approach based on shared responsibility among all stakeholders in the mobility data value chain.

This collective responsibility involves, in particular:

- **NAP operators**, responsible for defining common rules, providing validation tools, feedback mechanisms, and control processes;
- Data providers, responsible for the accuracy, completeness, timeliness, and compliance of published data and metadata;
- **Data users and re-users**, who contribute to identifying quality issues and articulating functional requirements;
- **National agencies and independent bodies**, responsible for oversight, conformity assessment, and, where applicable, regulatory oversight. This shared governance is essential in an environment characterized by the continuous flow of data between multiple national catalogues and platforms, where responsibilities must be clearly identified to ensure trust and traceability.

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

An approach based on continuous improvement

In addition to this multi-stakeholder governance model, NAPCORE recommends adopting an approach based on continuous quality improvement. This approach relies on successive assessments throughout the data lifecycle, covering:

- **The collection phase;**
- **The processing and aggregation phase;**
- **The phase of making data and services available via the NAPs.**

The framework distinguishes between several levels of anomalies (warning, non-blocking error, critical error), enabling proportionate management of non-conformities and ensuring that published data remains usable while complying with applicable regulatory requirements.

Clear definition of the evaluation object

To ensure that these evaluations are fully usable, NAPCORE emphasizes the need to clearly and explicitly define the object of the quality evaluation, to avoid any ambiguity in the interpretation of results and to promote their comparability.

The NAPCORE framework thus distinguishes several levels of evaluation, applicable to both data and metadata:

- **The evaluation of an isolated data element across a set of entities** (for example, a specific metadata field such as spatial or temporal coverage);
- **The evaluation of all data associated with a given entity;**
- **The evaluation of the dataset in its entirety**, including its compliance with a given schema or standard;
- **The specific evaluation of the associated metadata,** as a separate entity, focusing on its completeness, accuracy, and consistency with the data provided.

This clarification is essential to ensure a consistent interpretation of the results by re-users and to enable coherent comparisons between NAPs at the European level.

Assessment methods recommended by NAPCORE

Once the scope of the assessment has been clearly defined, NAPCORE recommends using a combined methodological approach based on several complementary methods tailored to different types of data, metadata, and services. The main methods identified include:

- **Automated syntactic and logical validation**, aimed at verifying compliance with schemas, applicable profiles, and business rules;
- **In-depth analysis of metadata**, particularly in an ex-ante logic, before the reuse of datasets;
- **Comparative analysis with other data sources**, when available, to detect potential inconsistencies;
- **Self-assessment by data providers**, particularly in the absence of external reference sources;
- **Consideration of feedback from users and re-users**, reflecting perceived quality and actual usability;
- The **contextual use of complementary data sources** to assess specific indicators such as coverage or representativeness.

NAPCORE recommends that these methods be implemented in a proportionate and complementary manner, depending on the type of data, the maturity level of the providers, and applicable regulatory requirements.

Funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

Common Quality Criteria

To support this harmonized approach, NAPCORE defines a common set of quality criteria applicable to data and metadata, including in particular:

- **Completeness;**
- **Compliance** with business schemas and rules;
- **Accuracy and internal consistency**;
- **Timeliness and up-to-date status;**
- Service **availability and performance**;
- Traceability of data and responsibilities.

These criteria constitute a common framework, essential for ensuring a high level of interoperability between NAPs and mobility data catalogues.

Importance of metadata quality

In this framework, metadata quality plays a central role, as it directly determines:

- The discoverability of datasets;
- Their comprehensibility to re-users;
- Their automated integration into digital services.

NAPCORE therefore recommends an explicit and systematic evaluation of metadata, based on complementary methods (automated validation, manual analysis, self-assessment, and user feedback), to ensure its completeness, accuracy, and consistency.

Transparency of results and integration into metadata

To maximize the value of the evaluations conducted, NAPCORE recommends that results be documented in a structured manner, including all contextual elements necessary for their interpretation (methods, scope, assumptions). This approach is aligned with the MobilityDCAT-AP metadata model and includes a field dedicated to quality description. This allows for the integration of a Quality Statement directly into the metadata of the relevant dataset.

Example of a metadata standard – MobilityDCAT-AP

As part of the European standardization efforts, NAPCORE promotes the use of the MobilityDCAT-AP metadata specification (designed to provide a harmonized description of mobility datasets within the European Union). This specification includes a precise field for data “Quality Description” that allows a Quality Statement to be associated with the metadata, documenting the criteria evaluated, the methods used, and the results obtained. This approach enhances transparency, comparability, and re-users’ confidence in a constantly evolving European mobility data ecosystem.

Conclusion

In a context marked by the proliferation of NAPs and data catalogues, the quality framework proposed by NAPCORE serves as a key tool for supporting the implementation of European policies on mobility data. By

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

integrating regulatory requirements, metadata standards, and common quality management practices, it contributes to the creation of an interoperable, reliable, and reuse-oriented European mobility data space.

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI