Rail to Digital automated up to autonomous train operation

D7.6 – Release of the Open-Data-Set

Due date of deliverable: 31/03/2026

Actual submission date: 11/06/2026

Leader/Responsible of this Deliverable: Dr. Philipp Neumaier, DB InfraGO AG

Reviewed: Y

Document status		
Revision	Date	Description
01	06/01/2026	First issue
02	31/03/2026	Final issue
03	11/06/2026	Implemented all recommended editorial fixes, including punctuation, terminology, revision table, heading styles, formatting harmonisation and layout adjustments.

Project funded from the European Union's Horizon Europe research and innovation programme		
Dissemination Level		
PU	Public	x
SEN	Sensitiv – limited under the conditions of the Grant Agreement	

Start date: 01/12/2022 Duration: 42 months

ACKNOWLEDGEMENTS

This project has received funding from the Europe's Rail Joint Undertaking (ERJU) under the Grant Agreement no. 101102001. The JU receives support from the European Union's Horizon Europe research and innovation programme and the Europe's Rail JU members other than the Union.

REPORT CONTRIBUTORS

Name	Company	Details of Contribution
Dr. Philipp Neumaier	DB	Content creator
Dr. Martin Köppel	DB	Content creator
Sebastian Dubiel	DB	Content creator
Philippe David	SNCF	Content creator
Philippe Bourdenet	SNCF	Content creator
Dr. Cedrick Lelionnais	SNCF	Content creator
Louis Delvig	SNCF	Content creator
Anders Ekberg	Chalmers University of Technology	Content creator
Bastian Simoni	Alstom	Expert Review
Dominik Kevicky	AZD	Expert Review
Frederic Antoine	ATSA	Expert Review
Hardik Jain	THD/Thales	Expert Review
Hájek Jiří	AZD	Expert Review
Hjalmar Boulouednine	SMO	Expert Review
Lars-Kristian Vognild	NRD	Expert Review
Maik Baehr	SMO	Expert Review
Michal Novak	AZD	Expert Review
Michele Bruzzo	Hitachi	Expert Review
Saro Thiyagarajan	FT/Wabtec	Expert Review
Sebastiaan Linssen	NS	Expert Review
Tom Jansen	NS	Expert Review

Disclaimer

The information in this document is provided “as is”, and no guarantee or warranty is given that the information is fit for any particular purpose. The content of this document reflects only the author’s view – the Joint Undertaking is not responsible for any use that may be made of the information it contains. The users use the information at their sole risk and liability.

EXECUTIVE SUMMARY

The transition towards higher levels of railway automation (GoA3/GoA4) requires robust, high-quality datasets to train, validate, and benchmark AI-based perception systems. However, the rail domain currently lacks openly available, standardised datasets that adequately capture both operational conditions and rare, safety-critical events. This gap limits reproducibility, comparability of results, and the broader adoption of AI solutions across Europe.

This deliverable addresses this challenge by presenting the **R2DATO Open Dataset**, developed within WP7 of the R2DATO project. The dataset aims to support the European railway sector by providing a structured, multimodal data foundation for AI development, while also contributing to the broader objectives of interoperability, safety, and innovation within the European rail ecosystem.

From a technical perspective, the dataset follows a **dual approach**, combining:

- **Real-world annotated railway data** from Deutsche Bahn (DB) and SNCF, and
- **Synthetic simulation data** generated using a digital twin environment

This combination allows the dataset to cover both realistic operational scenarios and controlled **Non-Regular Situations (NRS)** that are difficult or unsafe to capture in real-world settings. The work builds upon prior activities within WP7, including data generation (D7.3), annotation workflows (D7.4), and AI-based validation methodologies (D7.5), thereby ensuring methodological consistency across the Data Factory toolchain.

The real-world subset contributes annotated scenes reflecting diverse railway environments, including platform areas, trackside infrastructure, varying lighting conditions, and weather-related visual challenges. The data integrates contributions from multiple partners and follows a multimodal structure, incorporating camera and, where available, LiDAR-based representations. At the same time, the synthetic subset introduces controlled obstacle scenarios and simulation-based sensor outputs, enabling systematic testing and reproducibility.

A key outcome of this deliverable is the **structured integration of heterogeneous data sources** into a unified dataset concept. While the dataset does not enforce a fully standardised format across all contributions, it demonstrates how different annotation approaches (e.g. sequence-based vs. frame-based) can coexist within a common framework. This reflects current realities in the rail sector and provides valuable insights for future harmonisation efforts.

The dataset further establishes a **baseline for validation and benchmarking**, building on the methodology defined in D7.5. By enabling AI-based performance evaluation (e.g. object detection tasks), it provides a reference point for future research and development activities. This contributes to increased transparency, reproducibility, and comparability of AI models in railway applications.

From a European perspective, the dataset delivers clear added value by:

- Facilitating cross-partner collaboration and data sharing
- Supporting the development of interoperable AI solutions
- Providing a foundation for future standardisation efforts in railway data formats and annotation practices

At the same time, certain limitations must be acknowledged. The dataset represents a **prototype-scale release**, with a limited number of annotated scenes and a focus on static simulation scenarios. Furthermore, differences in data structure and annotation approaches across contributors highlight the need for further harmonisation.

Overall, the work presented in this deliverable can be considered a **significant step towards a scalable railway data ecosystem**, but not yet a complete solution. Further research and innovation activities are required to expand dataset coverage, standardise data formats, and integrate more advanced simulation and annotation capabilities.

Future work should focus on:

- Scaling the dataset in volume and diversity
- Harmonising annotation standards and metadata structures
- Expanding simulation capabilities to dynamic and complex scenarios
- Supporting industrial deployment and alignment with emerging standards

The R2DATO Open Dataset thus provides both a **practical resource** for current AI development and a **strategic foundation** for future advancements in railway automation and data-driven innovation.

ABBREVIATIONS AND ACRONYMS

Acronym/ Abbreviation	Full English Expansion
2D / 3D	Two-Dimensional / Three-Dimensional
6-DOF	Six Degrees of Freedom
ADIF	Administrador de Infraestructuras Ferroviarias
AI	Artificial Intelligence
AICat	AI Cataloguing Approach
API	Application Programming Interface
AP	Average Precision
ASAM	Association for Standardization of Automation and Measuring Systems
ATO	Automatic Train Operation
ATSA	Alstom Transport SA
AZD	AŽD Praha s.r.o.
BAT	Bounding Box Annotation Tool
BG	Background (Error)
CCO	Creative Commons Zero
CIDOC	International Committee for Documentation
CKAN	Comprehensive Knowledge Archive Network
CMOS	Complementary Metal-Oxide-Semiconductor
COCO	Common Objects in Context
CV	Computer Vision
CVAT	Computer Vision Annotation Tool
D7.x	Deliverable 7.x
DAE	Data Annotation Engine (Requirements)
DB	Deutsche Bahn (DB InfraGO AG)
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DCAT	Data Catalog Vocabulary
DCAT-AP	Data Catalog Vocabulary - Application Profile
DCAT-AP-HVD	DCAT-AP for High-Value Datasets
DEP	Digital Europe Programme
DGA	Data Governance Act
DL	Deep Learning

DMU	Diesel Multiple Unit
DTP	Data Touchpoint
DZSF	Deutsches Zentrum für Schienenverkehrsforschung
ELI	European Legislation Identifier
EMU	Electric Multiple Unit
ERJU	Europe’s Rail Joint Undertaking
FAIR	Findable, Accessible, Interoperable, Reusable
FMCW	Frequency-Modulated Continuous-Wave
FN	False Negative
FP	False Positive
FT	Faiveley Transport (Wabtec)
GDPR	General Data Protection Regulation
GNSS	Global Navigation Satellite System
GoA3	Grade of Automation 3
GoA4	Grade of Automation 4
GPS	Global Positioning System
GTD	Global Trade Digitization
GTSD	GTS Deutschland GmbH
HDR	High Dynamic Range
HIL	Hardware in the Loop
HPC	High-Performance Computing
HVD	High-Value Datasets
Hz	Hertz
ICAO	International Civil Aviation Organisation
IMU	Inertial Measurement Unit
IoU	Intersection over Union
IP	Intellectual Property
IR	Infrared
ISA²	Interoperability solutions for public administrations
JSON	JavaScript Object Notation
JU	Joint Undertaking
KPI	Key Performance Indicator
LAN	Local Area Network

LiDAR	Light Detection and Ranging
Loc	Localization (Error)
LWIR	Long-Wave Infrared
mAP	mean Average Precision
mAP_50 / mAP_75	mAP at IoU 0.50 / 0.75
mAP_l / mAP_m / mAP_s	mAP for Large/Medium/Small objects
MCAP	MCAP (Foxglove Format)
ML	Machine Learning
NRD	Norwegian Railway Directorate
NRS	Non-Regular Situations
NS	Nederlandse Spoorwegen
ODD	Operational Design Domain
OSDaR23	Open Sensor Data for Rail 2023
Oth	Other (Error)
PCD	Point Cloud Data
PEDF	Pan-European Data Factory
PIMS	Personal Information Management Service
PNG	Portable Network Graphics
PR	Precision-Recall (Curve)
PSI	Public Sector Information
PTP	Precision Time Protocol
PU	Public
px	Pixels
R2DATO	Rail to Digital automated up to autonomous train operation
RADAR	Radio Detection and Ranging
RailGoerl24	Görlitz Rail Test Center Dataset
RBAC	Role-Based Access Control
RDF	Resource Description Framework
RGB	Red-Green-Blue
RoI	Region of Interest
ROS	Robot Operating System

RPN	Region Proposal Network
S3	Simple Storage Service
SASP	System Access & Security Platform
SEN	Sensitive
Sim	Similar (Error)
SMO	Siemens Mobility GmbH
SNCF	Société Nationale des Chemins de fer Français
SSD	Solid-State Drive
THD	Thales
TN	True Negative
TP	True Positive
TRL	Technology Readiness Level
TSI	Technical Specification for Interoperability
UI	User Interface
UUID	Universally Unique Identifier
VDL	Vehicle Data Logger
WP7	Work Package 7
YOLOv8	You Only Look Once version 8

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION TO THE R2DATO OPEN DATASET

This deliverable marks the final step of Work Package 7. Following the definition of requirements (D7.1), the legal assessment (D7.2), the simulation setup (D7.3), and the completion of data annotation (D7.4/D7.5), D7.6 provides the public access to the resulting high-quality datasets.

1.1 RELATION TO OTHER DELIVERABLES

- **D7.1:** Provided the technical requirements for the Data Factory.
- **D7.2:** Established the legal foundation for data ownership and IP.
- **D7.3:** Performed simulations with the initial implementation of the Data Factory toolchain to validate the data pipeline.
- **D7.4:** Defined the annotation ontology and finalized the labels.
- **D7.5:** Validated the dataset quality through initial ML model training.

1.2 MOTIVATION AND CONTRIBUTION TO AUTONOMOUS RAIL RESEARCH

The advancement of autonomous railway operations, particularly towards higher Grades of Automation (GoA3 and GoA4), relies heavily on the availability of high-quality, representative, and well-annotated datasets. In contrast to domains such as automotive or robotics, the railway sector currently faces a significant shortage of openly accessible datasets tailored to its specific operational and safety requirements.

This limitation is particularly critical in the context of **AI-based perception systems**, which form a core component of autonomous rail solutions. These systems require large volumes of diverse and reliable data to ensure robust performance under varying environmental conditions, infrastructure layouts, and operational contexts. At present, the lack of standardised and shareable railway datasets constrains:

- the development and validation of AI models,
- the reproducibility of research results, and
- the comparability of approaches across organisations.

Moreover, railway environments pose unique challenges that are not sufficiently covered by existing datasets from other domains. These include:

- highly structured infrastructure (e.g. tracks, signals, catenary systems),
- safety-critical operational constraints, and
- rare but impactful **Non-Regular Situations (NRS)** such as unexpected obstacles or degraded visibility conditions.

Capturing such scenarios in real-world settings is often impractical, costly, or unsafe, further reinforcing the need for complementary data generation approaches.

In this context, the R2DATO Open Dataset contributes to autonomous rail research by providing a **multimodal and domain-specific data resource** that addresses these challenges. The dataset introduces a combined approach that integrates:

- real-world annotated railway scenes reflecting operational complexity, and
- synthetic simulation data enabling controlled and reproducible representation of rare scenarios.

This combination allows researchers and developers to work with both realistic and systematically generated data, thereby improving the robustness and generalisability of AI models.

Beyond the dataset itself, the deliverable contributes to the research landscape by:

- demonstrating practical approaches to handling heterogeneous data sources,
- providing insights into annotation strategies and data structuring in the rail domain, and
- supporting the establishment of benchmarking practices for railway perception tasks.

From a European perspective, this work supports the broader objectives of strengthening innovation capacity and fostering collaboration across stakeholders in the rail sector. By making structured data available and promoting shared methodologies, the dataset helps to reduce fragmentation and enables a more coordinated development of AI-based railway technologies.

In summary, the motivation of this deliverable lies in addressing a fundamental bottleneck in autonomous rail research: the lack of accessible, structured, and representative data. Its contribution is not only the provision of a dataset, but also the advancement of **data-driven methodologies and collaboration practices** that are essential for the future of autonomous railway systems.

1.3 Relation to the R2DATO Data Factory

The R2DATO Open Dataset is a direct outcome of the **Data Factory concept** developed within WP7 and represents its most tangible and integrative result. The Data Factory aims to establish a structured, end-to-end framework for the generation, annotation, management, validation, and dissemination of data for AI-based railway applications.

Within this framework, the present deliverable (D7.6) constitutes the **final aggregation and release stage**, consolidating outputs from preceding tasks and deliverables into a coherent and usable dataset.

The relationship between this deliverable and the Data Factory can be understood along the following stages:

Data Generation (D7.3)

The synthetic component of the dataset originates from the data generation processes defined in D7.3. These include:

- the creation of a digital twin environment,
- the simulation of railway scenarios, and

- the generation of multimodal sensor data.

This stage enables the systematic production of controlled scenarios, particularly for Non-Regular Situations, which are difficult to obtain from real-world operations.

Data Annotation (D7.4)

The real-world component of the dataset is based on the annotation workflows and methodologies developed in D7.4. This includes:

- the definition of annotation schemas,
- the use of multimodal labelling approaches, and
- the creation of structured annotated datasets by project partners.

The differences observed between partner contributions (e.g. sequence-based vs. frame-based structures) reflect the current diversity of annotation practices and provide valuable input for future harmonisation.

Data Validation and AI Benchmarking (D7.5)

The dataset is further contextualised by the validation methodologies introduced in D7.5. These include:

- the use of AI models (e.g. object detection architectures)
- the application of performance metrics such as mean Average Precision (mAP), and
- the establishment of reproducible evaluation procedures.

While D7.6 itself focuses on data provision, it is closely linked to these validation concepts, as the dataset is intended to serve as a basis for such evaluations.

Dataset Integration and Release (D7.6)

Building upon these prior stages, D7.6 performs the **integration of heterogeneous data sources** into a unified dataset concept. This includes:

- combining real-world and synthetic data,
- structuring the dataset for external use, and
- preparing it for dissemination within the constraints of project governance and data-sharing policies.

Rather than enforcing full standardisation, the deliverable adopts a pragmatic approach that preserves the structure of partner contributions while ensuring overall coherence.

Role within the Data Factory

Within the Data Factory lifecycle, this deliverable serves as:

- the **output interface** for external stakeholders,
- a **reference implementation** of the Data Factory concept, and
- a **baseline for future extensions and scaling**.

It demonstrates how data can flow through the different stages of the Data Factory, from generation to annotation and validation, culminating in a usable dataset.

2 DATASET COMPOSITION AND SPECIFICATION:

2.1 Data Acquisition

The real-world portion of the dataset consists of diverse scenes captured across the SNCF (French) and DB (German) rail networks, as defined in D7.4. These scenes were selected to represent a comprehensive range of operational environments across two of Europe's most significant rail infrastructures:

Locations and Infrastructure: Recordings include mainline tracks (high-speed and regional), complex station environments with multiple platforms, and shunting yards. Specific focus was placed on critical infrastructure elements such as switches, buffer stops, and signaling systems characteristic of both French and German networks.

Temporal and Environmental Conditions: To ensure model robustness, data was acquired during different times of day (daylight, twilight, and night) and under varying weather conditions (clear sky, rain, and light snow).

Scene Content: Beyond standard track-following, the dataset includes interactions with other trains (EMUs, freight), maintenance vehicles, and human activity (passengers on platforms, authorized personnel near tracks).

2.2 Data Generation

In addition to real-world recordings, the dataset incorporates high-fidelity synthetic data generated during the D7.3 simulation phase. Synthetic data is vital for training safety-critical models because:

Edge Case Coverage: It allows for the safe generation of Non-Regular Situations (NRS) that are too dangerous or rare to record in real-world traffic, such as people or large obstacles (fallen trees, vehicles) directly on the tracks.

Perfect Ground Truth: Synthetic environments provide pixel-perfect labels and depth maps, serving as a baseline for validating the performance of models trained on noisy real-world data.

Domain Adaptation: By mixing real and synthetic data, the Data Factory enables "augmentation" of real scenes with simulated hazards, enhancing the system's ability to react to unforeseen events.

2.3 Sensor Suite and Calibration Details

The R2DATO dataset leverages a multi-modal sensor pod designed to provide redundant and complementary environmental perception.

Sensor Specifications:

RGB Cameras: Industrial-grade CMOS sensors with high dynamic range (HDR) to handle tunnel exits and direct sunlight. Configurations include a long-range narrow-angle camera for signal detection and wide-angle cameras for perimeter monitoring.

Infrared (IR): Long-wave infrared (LWIR) sensors used primarily for detecting thermal signatures of persons and animals in low-visibility or nighttime conditions.

LiDAR: High-resolution 3D LiDAR (e.g., 128-beam) providing precise spatial data and depth perception up to 200m+.

Radar: Frequency-Modulated Continuous-Wave (FMCW) radar for robust velocity estimation and detection through adverse weather (heavy rain/fog).

Calibration:

Full calibration parameters are provided for every sequence to support sensor fusion:

Intrinsic Calibration: Lens distortion parameters and focal lengths for all optical sensors.

Extrinsic Calibration: Precise 6-DOF (translation and rotation) relative to the vehicle coordinate system, enabling the projection of LiDAR point clouds into camera frames.

Time Synchronization: All sensors are synchronized via PTP (Precision Time Protocol) to ensure temporal alignment across modalities.

2.4 Maintenance and Operational Support Use Case

This section integrates the contribution provided Trafikverket focusing on how sensor data from the R2DATO Data Factory can be leveraged for predictive maintenance, asset monitoring, and operational decision support in autonomous railway environments.

This part of the Work Package focuses on how sensor data from the R2DATO Data Factory and additional sensor data can be leveraged for predictive maintenance, asset monitoring, and operational decision support in autonomous railway environments.

The main motivation for the investigation is that with increased automation, the possibility for human observations and inspections of the track status decreases. Assessment of operational and safety related issues must then to a larger degree be based on sensor data from (autonomous) vehicles.

To this end, the current work investigates possibilities of inspecting components and influential factors from (autonomous) trains. This assessment describes the scope of inspection (e.g., track component, influential factors), and the feasibility to employ train-based sensor data to assess the status. The scope includes monitoring and management of track and vehicle health, interaction with wayside monitoring equipment, operation focused decision support and risk analysis.

Identified sensor data may be a part of the Data Factory data, in which case the use for maintenance and operational support will be an added benefit. For sensor data not part of the Data Factory, the feasibility assessment will provide a basis to decide whether they should be included in future extensions.

The research is carried out in conjunction with Iam4Rail WP18 where the feasibility of assessment based on sensor data from an (autonomous) robot is scrutinized. Although some uses overlap, the capabilities of commercial trains and track-based robots in general differ.

To measure the different parameters will in some cases pose demands on the vehicle related to weight, speed, equipment, traction etc. In addition, there will be demands on the measurement abilities relating to speed of measurements, resolution, sampling rate, positioning of the measurement. Regarding the measurements themselves, some relevant aspects are

- Costs – including both measurement and analyses – in relation to benefits
- Precision – what is the (average) accuracy of the measured quantity?
- Reliability – are the data reliable or are there (systematic or random) deviations that may vary over time
- Analysis abilities – how well can the collected data be related to status quantifications and predictions of deterioration
- Track occupancy – how will the data collection affect revenue traffic
- Safety – is the data collection (and use) causing any safety concerns?
- Security – can the data be employed for malicious purposes?
- Legal issues – who owns the data and how can they be shared and used?

  For monitoring tasks deemed interesting, relevant aspects from the lists above are discussed in the report.

  Some overall conclusions from the research (presented in detail in the report) are:

- For track health monitoring, vehicle-based sensors can be. employed if the inspection target is accessible from the vehicle (e.g., true for monitoring of sleepers, but not for bridges) and can be inspected at high speed (e.g., true for water pooling, but not in general for cracks in sleepers).
- For vehicle health monitoring, much of the monitoring is already employed in high-end trains. So, in general, sensor solutions exist. The main research challenge is to enhance the value of the data by better assessments including measurement strategies. With that handled, the use of sensor data will essentially be a cost/benefit assessment.
- Vehicle monitoring can be enhanced by use of data from wayside monitoring equipment. This requires collection, interpretation and action on these data. The feasibility of using different types of detector data is investigated.
- The study has also considered decision support and risk analyses. Here, some phenomena that may pose potential safety issues are scrutinized. Essentially, the common challenge is to ensure safe operations while causing a minimum of traffic disruptions. Some possibilities and recommendations are presented. It should be noted that for many of these phenomena operational practices are currently prescribed in regulations. These regulations should of

course be followed. The presented recommendations should here be seen as input in future updates of the regulations.

- Much assessment can be made using video data, which provides an added benefit to the use of video recording on autonomous trains. Other types of monitoring can employ sensors that are often standard in modern trains, e.g., accelerometers. This would provide a possibility for railway operators to increase the benefits if these data. There are also some monitoring. tasks (e.g., rail crack detection) that require additional sensors. For these a commercial arrangement for data purchase would be required if train operators should equip (some of) their trains. (Anders Ekberg, 2026)

3 ANNOTATION STANDARD AND TAXONOMY

3.1 THE RAILLABEL SCHEMA

The R2DATO dataset utilizes **RailLabel**, a specialized JSON-based schema derived from the **ASAM OpenLABEL** standard. RailLabel extends the generic OpenLABEL structure to meet the unique requirements of the rail domain, ensuring that metadata, sensor configurations, and object annotations are stored in a machine-readable and highly interoperable format.

3.1.1 Architectural Foundation and OpenLABEL Conformance

RailLabel is designed to be fully compliant with the ASAM OpenLABEL 1.0.0 standard. This architectural choice ensures that the dataset can be ingested by any toolchain that supports the standard automotive labeling ecosystem. The JSON structure is partitioned into logical blocks:

- **Metadata:** This section provides provenance for the file, including the schema version, tagger information, and dataset UUIDs.
- **Ontology:** Instead of external documentation, the label file itself contains the semantic definitions (types and attributes) used within the sequence.

3.1.2 Semantic Extensions for Rail Operations

While standard OpenLABEL is optimized for road traffic (Cars, Pedestrians, Bicycles), RailLabel introduces high-granularity rail-specific classes. This includes the distinction between different types of **Signals** (shape vs. light signals), **Infrastructure** (switches, buffer stops), and **Catenary Systems**. These extensions are crucial for GoA4, where the perception system must interpret track-side signaling and infrastructure clearance.

3.1.3 Geometric and Temporal Framework

A key strength of the RailLabel schema is its integrated handling of sensor-to-world transformations:

- **Coordinate Systems:** RailLabel explicitly defines "Parent-Child" relationships between coordinate systems (e.g., Sensor -> Vehicle -> World).
- **Calibration Integration:** Extrinsic and intrinsic parameters are embedded within the JSON, allowing for the direct projection of 3D LiDAR cuboids into 2D camera frames without requiring external calibration files.
- **Streams:** The schema supports multiple streams (e.g., camera_front_center, lidar_roof). Each frame in the sequence references these streams, ensuring that multi-modal data is perfectly aligned in time.

3.1.4 Object Representation and Attributes

RailLabel supports a variety of 2D and 3D primitives:

- **2D Bounding Boxes:** Primarily used for optical camera streams.
- **3D Cuboids:** Defined in the LiDAR coordinate system with rotation (yaw, pitch, roll).

- **Polylines & Polygons:** Used for track centerline estimation and area-of-interest (AOI) definitions.
- **Attributes:** Every object can carry complex attributes such as "Occupancy State" for tracks or "Signal Aspect" (e.g., Green/Red) for signaling equipment. (DB InfraGO AG, 2026)

3.2 Object Class Definitions and Attributes

The taxonomy used for annotation in D7.4 was designed to be both comprehensive and safety-oriented. It distinguishes between dynamic participants, infrastructure elements, and environmental conditions.

1. **Core Object Classes:**

- **Persons:** Includes authorized personnel (track workers), passengers on platforms, and unauthorized trespassers. Attributes include orientation and activity.
- **Trains:** Classified into sub-types such as Passenger (EMU/DMU), Freight, and Maintenance vehicles. Attributes include state (moving/static) and light configuration.
- **Infrastructure Elements:**
    - **Signals:** Light and shape signals, annotated with their current aspect or state.
    - **Track Elements:** Switches, buffer stops, and crossing infrastructure.
    - **Catenary System:** Catenary poles and overhead lines, critical for shunting and clearance monitoring.
- **Obstacles:** Any objects on or near the track that do not fit into standard classes (e.g., fallen trees, animals, debris).

2. **Attribute Scopes:**

    As defined in the DAE requirements, attributes are categorized by their scope:

- **Object-Level Attributes:** Constant properties of a real-world object (e.g., "Color: Red" for a van) that persist across time and sensor modalities.
- **Frame-Level Attributes:** Environmental or situational data that applies to an entire scene at a specific point in time (e.g., "Weather: Rain", "TimeOfDay: Night").
- **Sensor-Specific Attributes:** Data relevant only to a particular sensor view, such as occlusion levels or truncation status in a specific camera frame.

This granular taxonomy allows ML models to not only detect objects but also understand their operational context, which is fundamental for GoA4 decision-making logic.

4 MAIN EUROPEAN STANDARDS FRAMEWORKS FOR AI DATASETS

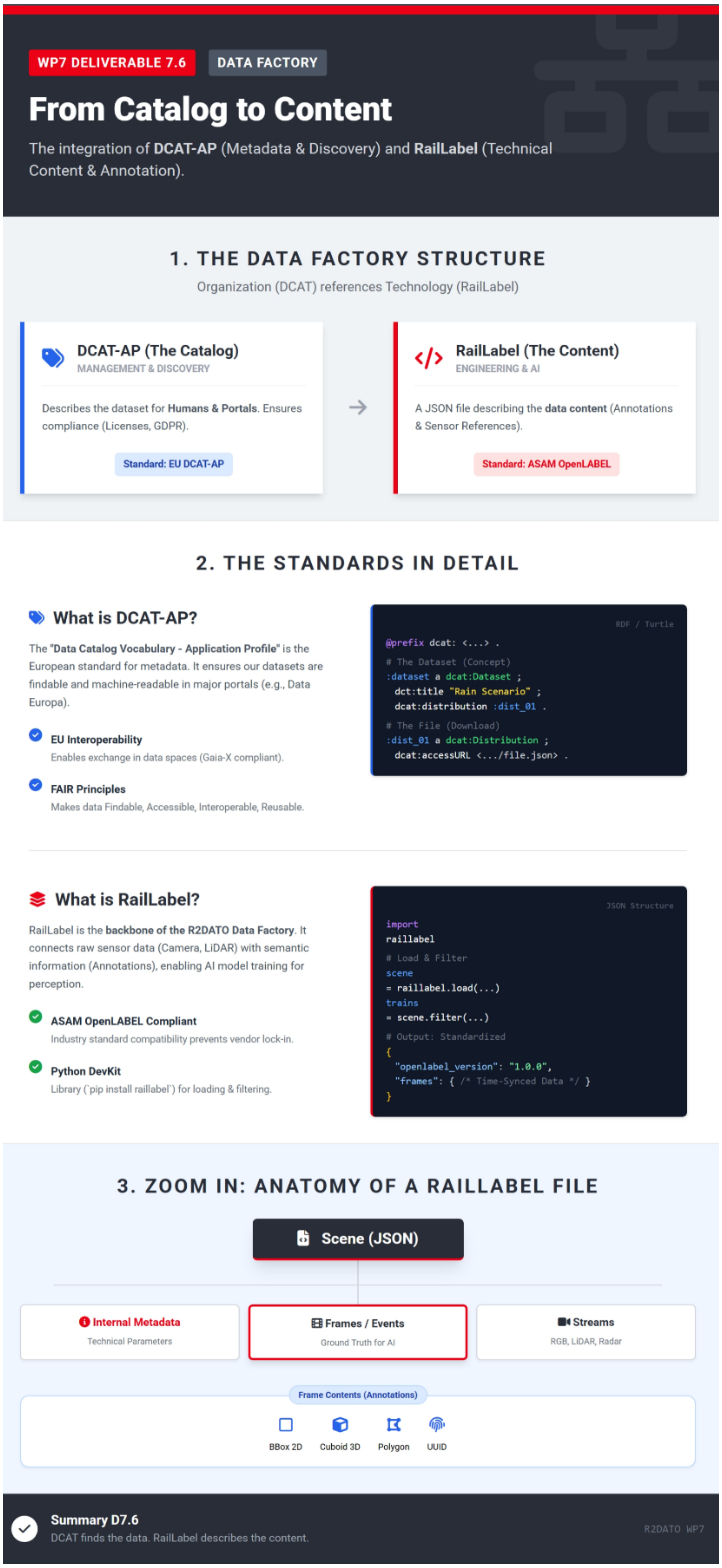

Figure 1 DCAT and RailLabel

This section provides a detailed explanation of the attached Figure 1, which illustrates the idea of the integration of metadata management and technical data content within the R2DATO Data Factory.

1. The Two-Layer Architecture: "Catalog vs. Content"
The infographic visualizes the clear separation and interaction between the organizational layer and the technical engineering layer.

- **Layer 1: The Catalog (Management & Discovery)**
    - **Standard: DCAT-AP** (Data Catalog Vocabulary - Application Profile).
    - **Purpose:** This layer serves **Humans and Portals**. It ensures that datasets are findable (FAIR principles), legally compliant (licenses, GDPR), and interoperable within European Data Spaces (e.g., Mobility Data Space).
    - **Key Function:** It acts as the "wrapper" or "packaging" for the data, providing high-level information such as the publisher, release date, and access rights.
- **Layer 2: The Content (Engineering & AI)**
    - **Standard: RailLabel** (based on ASAM OpenLABEL).
    - **Purpose:** This layer serves **AI Engineers and Algorithms**. It describes the actual content of the data, such as sensor configurations, synchronized frames, and ground truth annotations.
    - **Key Function:** It provides the deep technical "truth" required to train perception models.

2. The Link: From Metadata to Physical Files
The arrow between the two blocks represents the reference link. In DCAT-AP terms, the **Dataset** (Abstract Concept) points to a **Distribution** (Physical File).

- **The AccessURL** in DCAT points directly to the **RailLabel JSON file**.
- This architecture allows us to keep the catalog lightweight while storing complex, high-frequency sensor data references in the specialized OpenLABEL format.

3. Zoom In: Anatomy of a RailLabel File
The lower part of the infographic breaks down the structure of the RailLabel JSON file, which is the backbone of our training data:

- **Root (Scene):** The entry point, containing a synchronized sequence of data.
- **Metadata:** Technical parameters necessary for processing (e.g., coordinate system definitions, ontology versions).
- **Sensors (Streams):** References to the raw data streams (Camera, LiDAR, Radar). RailLabel does not store the image bytes itself but links to them, keeping the annotation file lightweight.
- **Frames (Events):** The core of the dataset. This section organizes data into time-synchronized snapshots.
- **Annotations (Ground Truth):** Inside each frame, specific objects are defined using the standardized ASAM OpenLABEL schema:
    - **2D Bounding Boxes:** For image-based detection.
    - **3D Cuboids:** For LiDAR/Radar spatial perception.
    - **Polylines:** For rail track definition.
    - **UUIDs:** Unique identifiers to track objects across multiple frames (crucial for tracking algorithms).

Conclusion
This integrated approach combines the best of both worlds:

1. **High Visibility** through DCAT-AP compliance (EU standard).
2. **High Technical Usability** through ASAM OpenLABEL (Industry standard).

4.1 THE USER-CENTRIC "CONSUMABLE LAYER"

SNCF's contribution to the Data Factory framework shifts the focus from a purely archival perspective to a **user-centric "consumable layer"**. The primary objective was to validate whether standard DCAT-AP metadata is sufficient for an external AI developer to discover, assess, and reuse datasets that were not originally designed for AI (a process termed "despecialization"). This approach simulates the perspective of an end-user—specifically a developer—seeking to exploit datasets for tasks like lateral signal detection or passenger counting, without prior knowledge of the dataset's internal structure.

4.1.1 Interoperability Profiles & DCAT-AP Adoption

To ensure seamless cross-system compatibility between France and Germany, SNCF defined **Interoperability Profiles** grounded in the **DCAT-AP** standard. This work was based on an inventory of existing datasets and a collaborative review with DB to harmonize metadata fields.

- **Core Standard (The "Minimal Profile"):** DCAT-AP serves as the backbone for describing dataset governance, provenance, and access. A "minimal core" of mandatory fields was defined (e.g., Title, License, Access URL, Publisher) to ensure that any dataset exposed through the Data Factory meets basic discoverability requirements.
- **Domain Extensions (The "AI Gap"):** The analysis of a minimal proof-of-concept dataset (a small multimodal dataset with ~100 entries) revealed that while DCAT handles general metadata well, it lacks the semantic precision required for AI engineering. Consequently, SNCF proposed lightweight, domain-specific extensions to capture AI-critical information. These include:
  - **Annotation Characteristics:** Type (bounding box, segmentation), format (OpenLABEL, COCO), and object classes (e.g., "Signal", "Pedestrian").
  - **Sensor Modalities:** Explicit tags for "LiDAR", "RGB Camera", or "Radar" to allow filtering by sensor type.
  - **Operational Context:** Metadata describing capture conditions (e.g., "tunnel", "night", "rain") which are crucial for defining the Operational Design Domain (ODD).

4.1.2 Synchronization and the "AI Readiness" Audit

A key innovation in the SNCF approach is the development of a **Metadata Audit Tool** and a synchronization API. Instead of simply cataloging data passively, this tool actively retrieves metadata from public repositories (tested against the OSDaR23 dataset via CKAN and DCAT endpoints) to evaluate its quality.

This process introduces the concept of an **"AI Readiness Score"** (scale 1-5), which serves as a quality indicator for developers. The score is calculated based on five key dimensions:

1. **Metadata Completeness:** Verifying the presence of all mandatory fields defined in the interoperability profiles.
2. **Annotation Coverage & Quality:** Checking for the existence and consistency of labels required for supervised learning.
3. **Modality Diversity & Volume:** Assessing if the dataset contains sufficient samples and diverse sensor data (e.g., multimodal alignment).
4. **Legal & Usage Compliance:** Confirming that licenses (e.g., CC-BY-4.0) are compatible with the intended AI application (research vs. commercial).
5. **Operational Representativeness:** Evaluating how well the dataset reflects real-world scenarios (e.g., diverse weather conditions).

Result: In the case study, the OSDaR23 dataset achieved a score of **5/5**, validating its high maturity for reuse. This transforms the catalog into an active quality assurance mechanism.

4.1.3 Despecialization of Legacy Data

The SNCF contribution highlights the strategic and technical challenge of "despecialization"—the process of repurposing highly specific operational data for broader AI applications. Historically, vast amounts of railway data have been collected for siloed objectives: infrastructure maintenance (e.g., measuring rail wear), legal archival (e.g., event recorder data), or regulatory certification (e.g., homologation tests). These "legacy" assets were often stored in proprietary formats with minimal metadata, designed only to answer the specific question asked at the time of collection, without considering future machine learning requirements (e.g., lack of frame-by-frame annotations or environmental tags).

Transforming these assets into "consumable" AI resources requires more than just access; it demands a rigorous, multidimensional assessment of their metadata and content. The Data Factory acts as a translation layer, enforcing strict interoperability profiles that map legacy attributes to standardized DCAT-AP fields. Through the automated "readiness audits" described earlier, the system can determine if a dataset collected for one purpose (e.g., detecting cracks in sleepers) holds sufficient semantic richness to serve another (e.g., training a general-purpose ballast segmentation model).

This mechanism ensures that a dataset originally collected for a niche task, such as verifying overhead line geometry, can be effectively "despecialized." It becomes discoverable to data scientists, legally cleared for secondary use through explicit licensing metadata, and technically validated for entirely new use cases, such as training a generic obstacle detection algorithm. By systematically unlocking these dormant reserves, the Data Factory maximizes the return on investment of past data acquisition campaigns, turning static historical archives into dynamic fuel for the next generation of railway AI.

4.1.4 What This Means Practically for the R2DATO Data Factory

- Adopt **DCAT-AP** (and possibly the HVD extension) to describe datasets, ensuring strong interoperability with European data assets.
- Use **controlled vocabularies** for licences, formats, provenance, versions, which enables automatic sharing and federation of catalogues—fundamental to what the Data Factory is meant to provide.
- Establish clear **metamodels** (ontologies, JSON or RDF schemas) to make metadata understandable both by machines and humans.
- If providing AI models (or handling highly sensitive datasets), consider using **AICat** to catalogue models and data according to future transparency requirements under the AI Act.
- Apply **FAIR principles** to ensure data is findable, accessible, interoperable and reusable, including within European collaboration contexts.
- If the Data Factory is or may become part of a data space, consider adopting the **Gaia-X framework** to ensure trust, security and sovereignty in data sharing, helping guarantee long-term viability.

5 DATA FORMAT AND ACCESS

5.1 OVERVIEW OF THE OPEN DATASET

The **R2DATO open dataset** combines two complementary data sources:

1. **real-world annotated railway sequences** from DB and SNCF, and
2. **synthetic simulation outputs** generated within WP7 as part of the Data Factory prototype.

This combined structure broadens the dataset beyond ordinary operational scenes and adds controlled, safety-relevant **Non-Regular Situations (NRS)** that are difficult, rare, or undesirable to capture in real railway operations. D7.3 explicitly positions the synthetic outputs as an input to the D7.6 open dataset and as a basis for later AI training, testing, and validation activities.

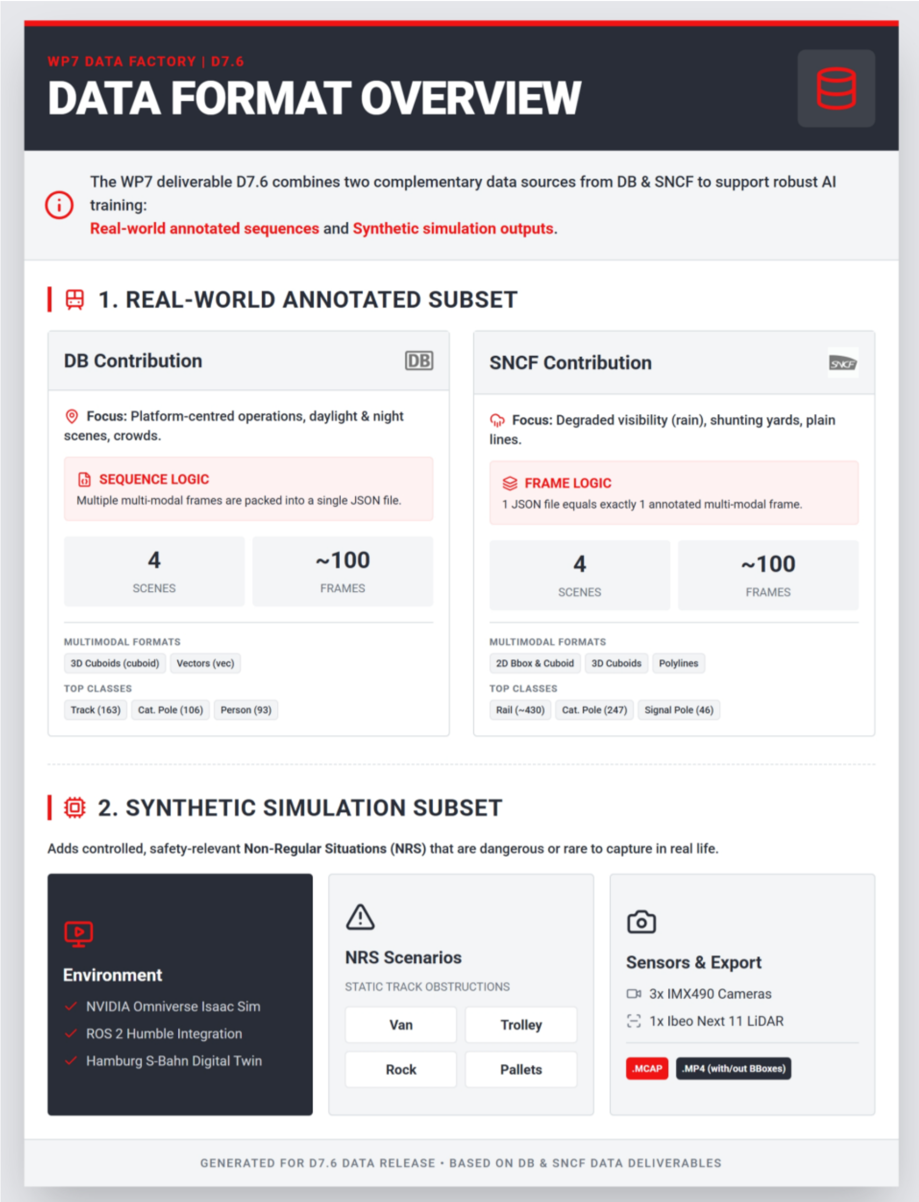

Figure 2 Data Format Overview of the Open Dataset

5.1.1 Real-world annotated subset

The real-world annotated subset consists of **eight selected scenes** contributed by **DB** and **SNCF**. It covers platform and line-side contexts, including **daylight and night-time scenes**, **crowded and sparse passenger environments**, **moving trains**, and **weather-affected recordings** with degraded visibility such as **raindrops on the lens** and reflections from wet surfaces. Based on the scene inventory you shared, this subset contains **approximately 207 annotated multimodal frames** in total and therefore provides a compact but diverse benchmark for railway perception tasks.

The DB scenes focus on platform-centred operational situations such as **Morning Platform View**, **Night Station Scene**, **Crowd at Platform**, and **Incoming Train with Passengers at Platform**. The SNCF scenes complement this with recordings such as **Platform with Customers**, **Shunting Yard Arrival**, **Crossing Train**, and **Plain Line**, which introduce rain-related optical artefacts and more degraded visual conditions.

It is important to distinguish between a **multi-modal frame** and an **image**. In D7.4, the term *frame* refers to one synchronised timestamp across the sensor setup. Such a frame may contain multiple sensor-specific outputs recorded at the same instant. An *image*, in contrast, refers to one individual visual file produced by a single camera sensor. Consequently, the number of images can be higher than the number of frames. For the DB subset, **240 multi-modal frames** recorded across **2 camera sensors** correspond to **480 individual images**.

The four DB annotation files are:

- 2022-11-30_23-29-58__2022-11-30_23-52-39+13.json
- 2022-12-09_09-48-40__2022-12-09_09-50-39+12.json
- 2022-11-23_07-46-50__2022-11-23_08-27-46+16.json
- 2022-12-09_09-48-40__2022-12-09_09-51-15+12.json

This means that the DB contribution to the annotated dataset is structured as **4 sequences**, **2 sensors**, and **480 individual annotated images**.

Object Class	Total Count	2022-11-23_07-46-50__...	2022-11-30_23-29-58__...	2022-12-09_09-48-40__...	2022-12-09_09-48-40__...
Track	163	31	16	58	58
Catenary Pole	106	23	3	40	40
Person	93	32	4	7	50
Signal	61	4	7	25	25
Road Vehicle	59	55	0	0	4
Signal Pole	52	4	8	20	20
IgnoreTracks	15	0	0	8	7
Train	13	0	4	4	5
Switch	10	0	2	4	4
Train Front	9	0	4	2	3
Crowd	3	0	0	1	2
Personal Item	2	2	0	0	0
Bicycle	1	1	0	0	0
Pram	1	0	0	0	1
Total Objects	588	152	48	169	219

Table 1 Distribution of object classes across annotated DB Scenes

For better readability in the table above, the filenames are truncated. Here are the full references:

- Col 3: 2022-11-23_07-46-50__2022-11-23_08-27-46+16.json
- Col 4: 2022-11-30_23-29-58__2022-11-30_23-52-39+13.json
- Col 5: 2022-12-09_09-48-40__2022-12-09_09-50-39+12.json
- Col 6: 2022-12-09_09-48-40__2022-12-09_09-51-15+12.json

Data Notes:

- **Unique IDs:** Counts are based on the unique UUIDs found in the objects section of the OpenLabel files.
- **Temporal Annotation:** Most objects are annotated across multiple frames within their respective sequences.

SNCF annotated data structure

The SNCF contribution follows a frame-based structure. The annotation package is organised around:

- a data folder
- a result folder
- a statistic.json file

According to statistic.json, the SNCF subset contains 101 total frames, all of which are marked as annotated and valid. The same file reports a total of 2,378 annotation elements across the set.
Unlike the DB contribution, the SNCF structure stores one annotated frame per JSON file. In this case:

1 JSON file = 1 annotated multi-modal frame

A sample SNCF frame file shows that one such frame may contain a combination of:

- 2D image-based annotations
- 3D LiDAR-based annotations
- rail polylines
- multiple railway object classes such as PLATFORM, SIGNAL, SIGNAL POLE, CATENARY POLE, and RAIL.

This means that an SNCF frame JSON should not be understood as a simple single-image annotation file. Rather, it acts as a multi-modal annotation container for one synchronised timestamp. The statistical summary further shows that the SNCF annotations are distributed across several annotation types:

- 2D Bounding Box: 504
- 2D Cuboid: 505
- 2D Polyline: 433
- 3D Cuboid: 504
- 3D Polyline: 432

The label distribution includes, among others:

- Catenary Pole: 247
- Rail: 433 / 432
- Signal Pole: 46
- Signal: 38
- Platform: 35
- Level Crossing: 22
- Person: 22
- Train: 21
- Wagons: 9
- Buffer Stop: 4
- Road Vehicle: 3.

The SNCF contribution is therefore best described as a frame-level annotation export, in which each timestamp is stored separately and already contains its corresponding multi-modal annotation content.

Key structural difference

The central difference between the two partner contributions can be summarised as follows:

- DB uses a sequence-based packaging logic: a small number of JSON files, each containing multiple annotated frames.
- SNCF uses a frame-based packaging logic: one JSON file per annotated frame.

This means that the number of JSON files is not directly comparable between the two partner contributions. A higher number of JSON files on the SNCF side does not in itself indicate a larger annotation scope; it primarily reflects a different export structure.

5.1.2 Synthetic simulation subset

The synthetic subset adds a second dataset layer based on controlled simulation runs. According to D7.3, the simulation environment was built around **NVIDIA Omniverse Isaac Sim 2022.1** and integrated with **ROS 2 Humble**, **Foxglove**, **Blender**, and **Unity**. The simulations were executed in a **digital twin of the Hamburg S-Bahn corridor from Rothenburgsort to Bergedorf**, while integrating **SNCF train and asset models** into the DB simulation environment.

For the initial simulation phase, the consortium focused on **static NRS obstacles**, as dynamic scenarios would have introduced higher computational load and additional modelling effort. The four implemented simulation scenarios are:

- **Van**
- **Trolley**
- **Rock**
- **Pallets**

These scenarios were selected to represent realistic track obstructions and to support the generation of synthetic sensor data for rare but safety-relevant situations. D7.3 states that each simulation run lasted **30 seconds**, matching both the final rendered videos and the recorded ROS bag sessions.

Sensor basis of the simulation data

The D7.3 deliverable describes the simulated sensor twin setup as consisting of **three IMX490-based cameras** and **one Ibeo Next 11 LiDAR**. The cameras were configured to provide **centre, left, and right** image streams, while the LiDAR generated point-cloud data. D7.3 also explains that the output was intentionally reduced in quality compared with the physical target setup in order to keep the simulation stable and computationally manageableral.

The supplementary SNCF sensor material you shared adds useful technical context for D7.6. The sensor folder includes:

- a **sensor positioning presentation**,
- the **Sony IMX490 datasheet**,
- a **sensor specification document**,
- and a **Tele-15 LiDAR field-of-view STEP model**.

These files show that SNCF provided supporting information on recommended sensor placement, including mounting the camera and LiDAR as high as possible for maximum track coverage, spacing stereo cameras at least two metres apart, and considering driver-eye-level views for lateral signalling. They also document the **IMX490** as a **5.40 MP automotive CMOS image sensor** with **2896 × 1876 recommended recording pixels**, HDR capability, and automotive-oriented output characteristics.

One point worth handling carefully in D7.6: the **supporting SNCF sensor documents** refer to a **Tele-15 / Livox-related LiDAR reference**, while the **final D7.3 simulation description** names **Ibeo Next 11** as the LiDAR twin used in the simulation runs. I would therefore present the SNCF folder as **supporting technical reference material**, while describing the actual released simulation outputs according to the final D7.3 wording.

Formats and released simulation outputs

D7.3 states that simulation data was recorded from the following main ROS topics:

- /camera/CENTER/image_raw
- /camera/LEFT/image_raw
- /camera/RIGHT/image_raw
- /etb/lidar/pointcloud

The recordings were saved in **ROS 2 Humble** using the **.mcap** format. In addition, final rendered videos were produced at **2896 × 1876 px**, and for each scenario two render variants were created:

- **with bounding boxes**, generated directly within Omniverse Isaac Sim,
- **without bounding boxes**, showing the plain rendered scene.

The synthetic part of the dataset includes:

3. **DB Simulations**

- **ROS Bags**
    - PALETTE.zip
    - ROCK.zip
    - TROLLEY.zip
    - VAN.zip
- **Rendered videos**
    - SNCF_Palette.mp4
    - SNCF_Palette_boundingbox.mp4
    - SNCF_Rock.mp4
    - SNCF_Rock_Boundingbox.mp4
    - SNCF_Trolley.mp4
    - SNCF_Trolley_Boundingbox.mp4
    - SNCF_Van.mp4
    - SNCF_Van_Boundingbox.mp4
- **Additional supporting media**
    - Camera Position.mp4
    - Recording Rosbag.mp4
    - one example ROS2 recording folder containing an .mcap file

6 LICENSING AND TERMS OF USE

6.1 DUAL-LICENSING MODEL

The licensing approach for the R2DATO open dataset follows the dual-licensing model recommended in Deliverable D7.2. Under this approach, the dataset is intended to be made available free of charge for non-commercial use, in particular for academic research and related research activities, under the licence specified in the release package. Any commercial use is not covered by this non-commercial release and requires a separate commercial agreement with the relevant publishing Partner Organization(s).

This approach reflects the objective identified in D7.2 to balance broad accessibility for research and innovation with the need for control and possible monetisation by the Partner Organizations. D7.2 further states that licensing shall be managed by the Partner Organizations, either jointly or individually, on the basis of centrally defined conditions by the Pan-European Entity.

For the avoidance of doubt, this D7.6 release concerns the dataset elements explicitly described in this deliverable. No rights are granted beyond the released dataset content itself unless expressly stated in the applicable licence or release documentation. In particular, any commercial exploitation, redistribution for commercial purposes, or use outside the scope of the applicable non-commercial licence requires prior authorisation by the relevant rights holder(s). This formulation is consistent with the D7.2 recommendation that commercial use should be subject to a separate licence and fee-based conditions.

Where users create derivative works or further developments based on released dataset elements, such use remains subject to the applicable licence terms. In line with D7.2, any broader licence-back mechanism for commercially relevant further developments may be addressed separately in commercial licensing arrangements, where applicable.

6.2 CITATION AND ATTRIBUTION REQUIREMENTS

Users of the R2DATO open dataset shall give appropriate credit when the dataset is used in publications, scientific papers, reports, presentations, demonstrators, or other public outputs.

As a minimum, users should reference:

- R2DATO – Rail to Digital Automated up to autonomous Train Operation
- the title of the dataset or dataset release
- the publishing Partner Organization(s), where indicated in the release package
- the year of release
- the applicable licence

A recommended citation format may be provided together with the dataset release metadata. If such a citation format is provided, users should use that format in order to ensure consistent attribution across publications and dissemination activities.

7 DATASET VALIDATION AND QUALITY ASSURANCE

The R2DATO open dataset is intended to provide not only curated data, but also a transparent basis for its later use in machine learning development and comparative evaluation. In this context, two complementary perspectives are relevant: first, the quality assurance of the annotation and curation process as described in Deliverable D7.4; second, the quantitative benchmarking of a baseline object detector as demonstrated in Deliverable D7.5.

In line with the current status of the project deliverables, it is important to distinguish clearly between these two perspectives. Deliverable D7.4 defines a structured framework for annotation quality assurance and validation, including schema checks, geometric and semantic consistency checks, cross-modal validation, temporal coherence analysis, ML-assisted review, and human-in-the-loop correction. However, D7.4 explicitly presents this as a conceptual and recommended framework for operational implementation rather than as a fully operationalised and empirically validated QA pipeline within the project. Accordingly, the D7.6 release follows the principles and recommendations laid out in D7.4, without claiming that the full proposed validation pipeline has already been executed end-to-end on all released data.

Complementing this, Deliverable D7.5 provides the currently available quantitative validation evidence by evaluating a baseline object detection workflow within the Data Factory context. The benchmark uses a Faster R-CNN R50 model pre-trained on COCO as generic baseline and compares it with models fine-tuned on rail-specific datasets. In its post-review update, D7.5 extends this evaluation to newly acquired and annotated data originating from D7.4, thereby establishing an initial bridge between the annotated project data and reproducible model-based performance assessment.

7.1 ANNOTATION QUALITY ASSURANCE FRAMEWORK AND CURATION PRINCIPLES (CF. D7.4)

The annotation quality assurance approach for the D7.6 dataset is based on the validation logic described in D7.4. This framework defines annotation quality along several dimensions, including formal schema integrity, geometric validity, semantic correctness, distributional regularity, redundancy detection, and temporal continuity. It further proposes a multi-stage validation pipeline consisting of:

(1) heuristic and rule-based checks,

(2) statistical and distributional analysis,

(3) cross-modal consistency validation between RGB, IR, and LiDAR annotations,

(4) temporal consistency and tracking validation across sequences,

(5) ML-assisted identification of potentially problematic annotations, and

(6) manual review with corrective feedback under a four-eyes principle.

Examples of the proposed checks include validation of bounding-box dimensions and coordinate ranges, detection of duplicate boxes with high IoU overlap, consistency of class labels across modalities, persistence of object identities across frames, and prioritised manual review of automatically flagged samples. D7.4 also stresses that standardised schemas, persistent object IDs, temporal alignment, and structured feedback loops are essential preconditions for robust downstream AI use.

For D7.6, this means that the released dataset is curated in accordance with the annotation standards and QA principles established in D7.4, especially with regard to structured annotation formats, traceability, consistency, and suitability for later ML use. At the same time, and in order to remain fully aligned with the wording of D7.4, this deliverable does not claim that all elements of the proposed D7.4 validation pipeline were already operationalised as a complete automated QA system across the entire released dataset. Rather, D7.6 adopts D7.4 as the methodological reference for annotation quality assurance and dataset curation.

7.2 INITIAL BASELINE BENCHMARK RESULTS ON THE DB D7.4 SUBSET (CF. D7.5)

The currently available benchmark evidence for project-specific data is reported in D7.5. In that deliverable, a Faster R-CNN R50 detector pre-trained on COCO is used as a generic baseline, and its performance is compared with rail-specific fine-tuned models. In the post-review extension, these models were evaluated on a newly compiled DB D7.4 test set in order to assess how well the trained detectors generalise to project-generated annotated data.

A relevant limitation must be stated explicitly for consistency with D7.5: the quantitative evaluation on D7.4 data was carried out only on the DB-generated and DB-annotated subset. SNCF data was excluded from this metric calculation, because it had been annotated using a different methodology and schema, and would have required additional mapping and harmonisation for a fair common benchmark. This is an important lesson for D7.6, as it underlines the role of common annotation standards for future cross-partner benchmarking.

For the OSDaR-related class set evaluated on the DB D7.4 test data, the pre-trained baseline achieved an overall bbox_mAP of 0.173, while the OSDaR-fine-tuned model reached 0.154 overall. For the class person, the pre-trained model scored 0.345, compared with 0.281 for the fine-tuned model. For the class train, however, the pre-trained model scored 0.000, whereas the fine-tuned model achieved 0.027, meaning that only the fine-tuned model showed any train detection capability on this D7.4 slice. D7.5 interprets this as evidence that domain adaptation is necessary for rail-specific classes, while also showing that train appearance in the D7.4 test set differs visibly from that in the OSDaR training data.

For the RailGoerl-related class set evaluated on the same DB D7.4 test data, the fine-tuned model slightly outperformed the pre-trained baseline overall, reaching bbox_mAP 0.178 compared with 0.173. For the class person, the RailGoerl-fine-tuned model achieved 0.356 versus 0.345 for the pre-trained model. For the class bicycle, both models scored 0.000. D7.5 explicitly states that the information value for bicycle is very limited because only two bicycle instances were present in the evaluated test data, making reliable statistical validation impossible for that class.

Taken together, these results provide a first reproducible benchmark on project-specific annotated data and show two points clearly. First, fine-tuning on rail-domain data can introduce or improve detection capabilities for railway-relevant object classes, especially where generic models fail completely, as seen for the train class. Second, benchmark validity remains highly dependent on class coverage, annotation harmonisation, and sufficient sample size. For this reason, the D7.6 release should be understood as providing an initial benchmark reference point rather than a fully comprehensive pan-European leaderboard. Future extensions of the dataset, especially with harmonised annotations across partners and broader class balance, are expected to improve both comparability and statistical robustness.

8 CONCLUSION AND FUTURE WORK

This deliverable has presented the **R2DATO Open Dataset** as a key outcome of WP7, consolidating the results of data generation, annotation, and validation activities into a coherent and accessible dataset for the railway domain.

The dataset demonstrates the feasibility of integrating **heterogeneous data sources**, combining real-world annotated railway scenes with synthetic simulation outputs. This dual approach enables coverage of both operational conditions and safety-critical scenarios, thereby addressing a central challenge in AI-based railway perception: the scarcity of representative and diverse training and validation data.

A major contribution of this work lies in the **practical implementation of the Data Factory concept**, showing how data from different partners, formats, and methodologies can be brought together within a common framework. In doing so, the deliverable highlights both the potential and the current limitations of data interoperability in the rail sector.

Furthermore, the dataset builds upon and extends the validation methodology defined in earlier work, providing a **baseline for reproducible AI benchmarking**. This supports the development of comparable and transparent performance metrics, which are essential for the adoption of AI systems in safety-critical railway environments.

From a sectoral perspective, the dataset contributes to:

- Advancing AI readiness in the European railway ecosystem
- Supporting collaborative research and innovation
- Providing a foundation for future standardisation efforts

However, the dataset also reflects the current maturity level of the domain. Differences in annotation structures, limited dataset size, and the focus on static simulation scenarios underline that further work is required to achieve a fully scalable and standardised data ecosystem.

Building on the results of this deliverable, several directions for future research and development can be identified.

Dataset Expansion and Diversity

Future efforts should focus on significantly increasing the dataset size and diversity, including:

- Additional real-world recordings across different countries and infrastructure types
- Broader coverage of environmental conditions (weather, lighting, seasonal effects)
- Inclusion of more complex and rare operational scenarios

Standardisation and Interoperability

The observed heterogeneity in data structures highlights the need for:

- Harmonised annotation standards
- Consistent metadata schemas
- Alignment with emerging initiatives such as OpenLABEL and related standards

This will be essential for enabling cross-dataset compatibility and large-scale data sharing.

Advanced Simulation Capabilities

The synthetic dataset can be further enhanced by:

- Introducing dynamic and interactive scenarios
- Modelling complex traffic situations and human behaviour
- Improving realism through advanced rendering and sensor modelling

Such developments will increase the relevance of simulation data for training and validation.

Integration into Industrial Pipelines

To support real-world deployment, future work should address:

- Integration of the dataset into industrial AI development workflows
- Continuous data updates and lifecycle management
- Validation in operational environments

Standardisation and Regulatory Impact

The dataset and associated methodologies can contribute to:

- The development of common benchmarking practices
- Input to standardisation bodies
- Supporting regulatory frameworks for AI in rail

Towards a Scalable Data Factory

Finally, the long-term vision is the evolution towards a **fully operational Data Factory**, capable of:

- Continuous data generation and curation
- Automated annotation and validation pipelines
- Scalable data sharing across stakeholders

REFERENCES

Anders Ekberg, E. K. (2026). *Maintenance and operation support using (autonomous) railway vehicles.* Gothenburg, SWEDEN: CHARMEC, Chalmers University of Technology.

DB InfraGO AG. (2026). *https://github.com/dbinfrago/raillabel*. Von https://github.com/dbinfrago/raillabel abgerufen

ANNEX

8.1 Main European Standards / Frameworks for AI Datasets

DCAT-AP (Data Catalogue Vocabulary – Application Profile)

This is the most widely used European application profile for describing metadata in data catalogues.

Data Europa: https://data.europa.eu/en/publications/datastories/linking-data-data-catalogue-vocabulary-application-profile

Source: https://data.europa.eu/en/news-events/events/webinar-unlocking-data-interoperability-exploring-ways-standardisation-real-use

See also the Data Catalog Vocabulary (DCAT-W3C): https://www.w3.org/TR/vocab-dcat/

It defines mandatory classes and properties for catalogues, datasets and distributions (e.g., title, description, access URL, publisher, etc.).

DCAT-AP facilitates the interoperability of data catalogues within European data spaces.

[2025] Interoperable Europe Portal: https://interoperable-europe.ec.europa.eu/collection/rolling-plan-ict-standardisation/data-interoperability-rp-2025

There are extensions such as **DCAT-AP-HVD** for *High-Value Datasets*, which may be relevant depending on the nature of AI datasets.

datos.gob.es: https://datos.gob.es/en/etiquetas/estandares

European Regulation – Data Governance Act (DGA)

The Data Governance Act (Regulation (EU) 2022/868) establishes essential interoperability requirements for European data spaces.

European Parliament: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0069_EN.html

[2024] Interoperable Europe Portal: https://interoperable-europe.ec.europa.eu/collection/rolling-plan-ict-standardisation/rolling-plan-2024

Under this framework, data structures (schemas, vocabularies, taxonomies, ontologies…) must be publicly and consistently described.

The DGA encourages neutral, open and inclusive technological solutions, and the possibility of using already existing harmonised specifications.

See the European Parliament website.

Licensing and Data Provenance

For “open” datasets, the DGA encourages the use of open, machine-readable licences.

Interoperable Europe Portal (ibid.)

The DCAT-AP standard provides for the use of controlled vocabularies for licences, formats, languages, etc.

Data Europa (ibid.)

It is also necessary to account for **traceability (provenance)**: where the data comes from, how it has been cleaned/annotated, versions, etc.

Cataloguing AI Systems – AICat

Recent research presents **AICat**, an extension of DCAT for describing AI systems (models, datasets, algorithms) in a catalogue aligned with the requirements of the European AI Act.

Ceur Workshops: https://ceur-ws.org/Vol-3910/aics2024_p38.pdf

AICat aims to make AI model metadata more machine-readable, traceable, interoperable, and compliant with transparency obligations.

Same source (Ceur Workshops).

FAIR Principles / FAIR 2.0

The FAIR principles (Findable, Accessible, Interoperable, Reusable) are widely adopted to structure data and metadata in a way that promotes reuse.

There are also proposals to evolve FAIR towards **FAIR 2.0**, especially to better integrate semantic interoperability (ontologies, schemas, mappings).

arXiv: https://arxiv.org/abs/2405.03345

The concept of **FAIRer Digital Objects** emphasises semantic modularity and metadata "navigability."

arXiv: https://arxiv.org/abs/2301.04202

Data Sovereignty / Trusted Infrastructure

Gaia-X is a European initiative aimed at creating federated, trusted data infrastructures.

Wikipedia: https://fr.wikipedia.org/wiki/Gaia-X

If the Data Factory operates within a data space or shared infrastructure, complying with Gaia-X principles can ensure alignment with European values (data sovereignty, security, transparency).