CEF2 RailDataFactory

D 1.3 – Pan-European Railway Data Factory Concept

(as part of an overarching deliverable
D 1 - Data Factory Concept, Use Cases and Requirements)

Version 1.3

Due date of deliverable: 31/03/2023

Actual submission date: 17/05/2023

Resubmission date: 11/08/2023

Responsible of this Deliverable: Philipp Neumaier (WP 1 lead, DB), Patrick Marsch (editor, DB)

Document status			
Revision		Date	Description
Referring to overall D 1 - Data Factory Concept, Use Cases and Requirements	0.1	09/03/2023	Document template generated
	0.2	24/03/2023	Major parts of content transferred from Confluence
	0.3	29/03/2023	First complete draft
	0.4	04/04/2023	Draft version submitted to advisory board
	0.5	11/04/2023	Use case and requirements sections merged
	0.6	19/04/2023	First review and commenting the advisory board comments
	0.7	24/04/2023	Final version after addressing of all advisory board comments, sent for final consortium approval
	1.0	28/04/2023	Version submitted to the project officer
	1.1	03/05/2023	Correction on formatting and spelling errors
	1.2	17/05/2023	Extract of deliverable D 1 created for D 1.3
	1.3	11/08/2023	Disclaimer updated based on the feedback of the granting authority

Project funded by the European Health and Digital Executive Agency, HADEA, under Connecting Europe Facilities Digital Grant Agreement 101095272		
Dissemination Level		
PU	Public	
SEN	Sensitive – limited under the conditions of the Grant Agreement	X

Start date: 01/01/2023 Duration: 9 months

ACKNOWLEDGEMENTS

This project has received funding from the European Health and Digital Executive Agency, HADEA, under Connecting Europe Facilities Digital Grant Agreement 101095272.

REPORT CONTRIBUTORS (IN ALPHABETICAL ORDER)

Name	Company
Bart du Chatinier	NS
Julian Wissmann	DB
Mayank Singh	DB
Patrick Marsch	DB
Philipp Neumaier	DB
Philippe David	SNCF
Wolfgang Albert	DB

Note of Thanks

We would like to thank our Advisory Board Members Maria Aguado, Saro Thiyagarajan and Manuel Kolly for the valuable discussion and in particular Xiaolu Rao, Janneke Tax and Oliver Lehmann for their thorough reviews of deliverables D 1.1, D 1.2 and D 1.3 and input to this work!

Disclaimer

Licensing

EXECUTIVE SUMMARY

The European rail sector is currently on the verge to the strongest technology leap in its history, with many railway infrastructure managers and railway undertakings striving toward large degrees of automation in rail operation, and mechanisms to increase the capacity and quality of rail operation.

In particular in the pursuit of fully automated driving (so-called Grade of Automation 4, GoA4), where sensors and cameras on trains will be used to automatically react to hazards in rail operation, it is commonly understood that an individual railway company or railway vendor would not be able to collect enough sensor data to sufficiently train the artificial intelligence (AI) eventually deployed in the rail system.

For this reason, it is commonly assumed that a form of pan-European Railway Data Factory is needed, as an infrastructure and ecosystem that allows various railway players and suppliers to collect and process sensor data, perform simulations, develop AI models, certify models, and ultimately deploy the models in the automated railway system.

In close sync with related activities listed in Section 1.2, the **CEF2 RailDataFactory** study focuses in particular on the pan-European Data Factory backbone network and data platforms required to realize the vision of the Data Factory.

In a first set of deliverables of the study comprising D 1.1, D 1.2 and D 1.3, the high-level vision of the pan-European Data Factory is introduced, key operational scenarios and use cases are defined, and related requirements in particular on the pan-European Data Factory backbone network and data platforms are derived and complemented with legal, regulatory and Cyber-security related aspects to be considered. Altogether, these requirements serve as a basis for the further work in this study.

ABBREVIATIONS AND ACRONYMS

Abbreviation	Definition
AI	Artificial Intelligence
CEF	Connecting Europe Facilities
ERA	European Union Agency for Railways
GoA4	Grade of Automation 4
HADEA	European Health and Digital Executive Agency
IAM	Identity Access Management
IM	Infrastructure Manager
ISMS	Information Security Management System
ML	Machine Learning
PII	Personally Identifiable Information
RU	Railway Undertaking
TLS	Transport Layer Security

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The European railway sector is on the verge to the strongest technology leap in its history, with many railway infrastructure managers (IMs) and railway undertakings (RUs) striving toward large degrees of automation in rail operation, and mechanisms to increase the capacity and quality of rail operation.

In particular, various railway companies - both IMs and RUs - and railway suppliers are currently working toward fully automated rail operation (so-called Grade of Automation 4, GoA4), for instance in the context of the Shift2Rail [1] and Europe's Rail [2] programs, in which sophisticated lidar and radar sensors as well as cameras are used to automatically detect and respond to hazards in rail operation, such as objects on the track or passengers in stations in dangerous proximity of the track. Another important use case is high-precision train localization by detecting static infrastructure elements and locating them on a digital map, as for instance covered in the Sensors4Rail project [3]. While the rail system has various properties that render fully automated driving principally easier than, e.g., in the automotive sector (for instance, railway motion is only one-dimensional, scenarios are typically much less complex than automotive scenarios, etc.), key challenges on the way to fully automated driving in the rail sector are that hazardous situations have to be detected much earlier due to long braking distances, and it is very challenging to collect and annotate sufficient amounts of sensor data with sufficient occurrences of relevant incidences to perform the required artificial intelligence (AI) training and to be able to prove that the trained AI meets the safety needs.

For this, it is expected that single railway suppliers, IMs and RUs will not be able by themselves to collect and annotate sufficient amounts of sensor data for AI training purposes - but instead, a European data platform and ecosystem is required into which railway stakeholders (suppliers, IMs, RUs, railway undertakings, safety authorities, and others) can feed, process and extract sensor data, as well as simulate artificial sensor data, and through which the stakeholders can jointly develop and assess the AI models needed for fully automated driving.

1.1 AIM AND SCOPE OF THE CEF2 RAILDATAFACTORY STUDY

The CEF2 Rail Data Factory study focuses exactly on aforementioned vision of a Pan-European Data Factory for the joint development of fully automated driving. The study, being co-funded through HADEA, aims to assess the feasibility of a Pan-European Data Factory from technical, economical, legal, regulatory and operational perspectives, and determine key aspects that are required to make a pan-European Data Factory a success. In particular, the study aims to:

- clarify the key operational scenarios and use cases to be covered by a Data Factory;
- determine the requirements of these use cases and scenarios on the Data Factory infrastructure (in particular w.r.t. the Pan-European Railway Data Factory Backbone Network, security, data and IT platforms, etc.);
- determine legal and regulatory aspects to be considered as well as a possible economic incentive model for the Data Factory;
- determine potential show-stoppers toward a pan-European Data Factory and related mitigation means; and

- speak out specific recommendations on how a pan-European Data Factory should be setup, incl. a detailed deployment strategy, or elaborate on the advantages and disadvantages of different options, where it is not possible to speak out a single recommendation.

For clarity, Table 1 lists which exact aspects are in the scope of this study, and which are not.

Table 1. Delineation of what is in scope and out of scope of this study.

In scope of this study	NOT in scope
• Description of the vision of a Pan-European Data Factory incl. definition of key terminology; • Definition of roles and users of the Data Factory and derivation of use cases related to the Pan-European Data Factory; • Derivation of requirements in particular related to a Pan-European Data Factory Backbone Network and required data and compute platforms; • Development of an architecture of the Pan-European Data Factory, with a particular emphasis on the platform architecture of data centers, pan-European usage of tools and services, and their connection through a Pan-European Data Factory Backbone Network, incl. elements required for security such as Identity Access Management (IAM); • Assessment of a Pan-European Data Factory from legal, regulatory, economic and operational perspectives, and derivation of key points that have to be addressed to make the Data Factory a success; • Development of specific recommendations how to realize a Pan-European Rail Data Factory, including a specific deployment strategy.	• Details on sensor data sources (on train or trackside) or specific sensor types; • Details on the data structure, format and quality requirements, etc., of the data being fed into, stored and processed in the Pan-European Data Factory; • Details on the AI algorithms, AI training, simulations, and the forms of fully automated driving (GoA4) the Pan-European Data Factory would be used for; • Ethical aspects related to the usage of AI in fully automated driving (GoA4); • Details on billing aspects; • Details on the management of individual data centers or tools, etc. (beyond the notion of aspects that appear necessary to be harmonized across data centers); • Implementation activities.

1.2 DELINEATION FROM AND RELATION TO OTHER WORKS

The Shift2Rail project **TAURO** [4] also looks into the development of fully automated rail operation, for instance focusing on developing

- a common database for artificial intelligence (AI) training;
- a certification concept for the artificial sense when applied to safety related functions;
- track digital maps with the integration of visual landmarks and radar signatures to support enhanced positioning and autonomous operation;
- environment perception technologies (e.g., artificial vision).

The difference of the CEF2 RailDataFactory project is that this puts special emphasis on the **pan-European Data Factory backbone network and data platform** (located on the infrastructure side, but used for sensor data collected through both onboard and infrastructure side sensors) required for the Data Factory, and also investigates **commercial, legal and operational aspects** that have to be addressed to ensure that the vision of the Data Factory can be realized.

The input from the TAURO project is, however, taken into consideration in particular in the derivation of use cases for the Data Factory, as covered in D 1.1

The Europe's Rail Innovation Pillar **FP2 R2DATO project** [5], overall focusing on the further development of automated rail operations, also has a work package dedicated to the Data Factory. Here, however, the main focus is on creating first implementations of individual data centers and toolchains as required for specific other activities and demonstrators in the FP2 R2DATO project, and on developing an **Open Data Set**. A strong alignment between the CEF2 RailDataFactory study and the FP2 R2DATO Data Factory activities is ensured through an alignment on use cases and operational scenarios, though the actual focus of the projects is then different.

Within the sector initiative "Digitale Schiene Deutschland", Deutsche Bahn already started to set up some components of the Data Factory [6].

1.3 AIM AND STRUCTURE OF THIS DELIVERABLE

This current document represents D 1.3 of the CEF 2 RailDataFactory project, describing the basic concept of the envisioned pan-European Data Factory, including the envisioned contributor concept and roles in the RailDataFactory.

The aim of the document is to obtain early feedback and possible additions from the sector on the described Data Factory concept, in order to update the work accordingly and consider the obtained input in the subsequent phases of the project, in which the detailed Data Factory architecture, legal and business aspects will be developed.

Note: This deliverable D 1.3 is part of an overarching deliverable D 1 Data Factory Concept, Use Cases and Requirements. As the deliverables D 1.1, D 1.2 and D 1.3 comprised in D 1 are strongly related and build upon each other, the reader is here pointed to the overarching D 1.

The remainder of this document is structured as follows:

- In Chapter 2, the vision of the pan-European Data Factory is shortly introduced, together with key terminology used in the remainder of this document;
- In Chapter 3, the envisioned contributor concept of the Data Factory is introduced, and roles are defined;
- Finally, in Chapter 4, this document is concluded with a summary and the expected next steps in the study.

2 VISION AND DEFINITION OF THE DATA FACTORY

In this section, the broader vision behind the pan-European Data Factory is introduced, and key aspects and terminology are defined.

The **pan-European Data Factory** is a set of interconnected **Data Centers** - operated by individual IMs, RUs, railway suppliers and others - comprising **Computing Resources** and **Data Storage Resources** and hosting **Tools** and **Services**.

Key to the Data Factory is a **uniform and consistent Tool Chain** that connects all Data Centers in their functionality so that **data can be jointly stored, processed, annotated, simulated and managed** across Europe. This forms the basis of a joint development, training and evaluation of AI functionalities - at the end of which is the approval to use trained AI models for fully automated rail operation.

On one hand, the main data source are environment sensors such as lidar or radar sensors, cameras, load or chemical/fire detectors, etc. (on the trains or on the trackside, for instance at level crossings or in stations) that record real-world data. This data is fed into the Data Centers via Data Entry Points (so called **Touch Points**) and stored there.

On the other hand, simulations of virtual rail environments are performed in the Data Centers using digital sensor twins, thus generating **Artificial Sensor Data**.

Both types of data constitute the data basis for training and evaluating AI functions.

Figure 1 shows a high-level illustration of the pan-European Data Factory, with the aforementioned Data Centers and Data entry points (Touch Points).

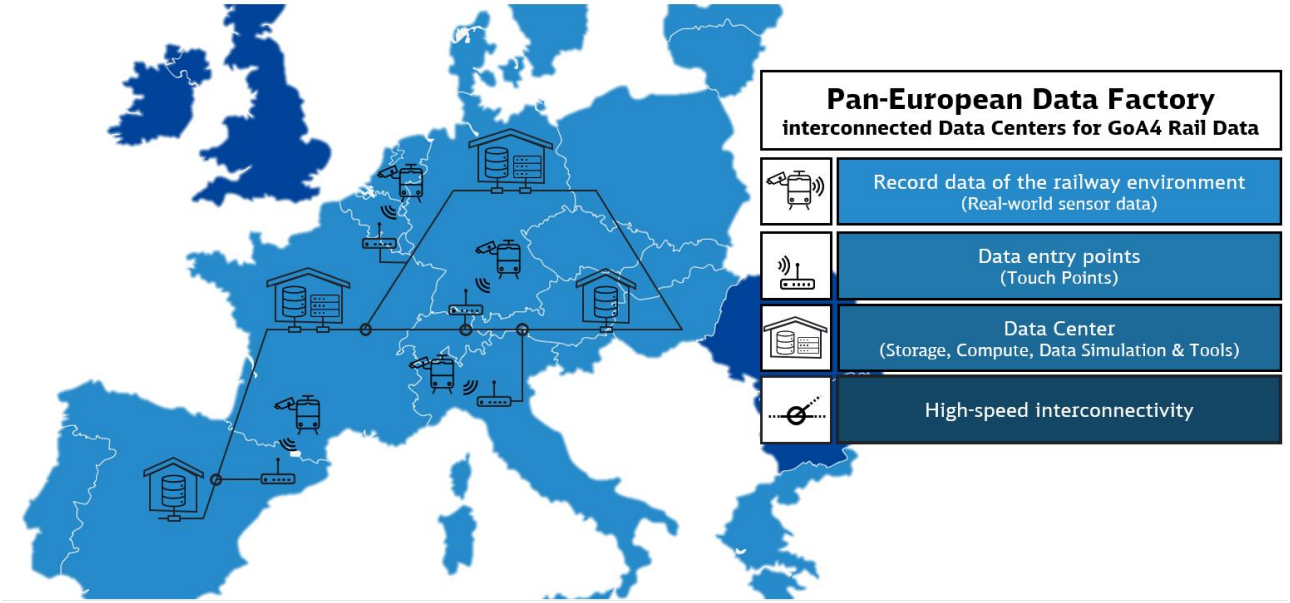

Figure 1. High-level illustration of the Pan-European Data Factory.

Figure 2 shows an exemplary Data Center in more detail. As shown in the figure, it is expected that both private data of specific users as well as shared data could be stored in a Data Center.

An example for **private data** could for instance be sensor data that has been collected by one RU, but which has not yet been validated for its suitability for AI training, and which the RU would hence (for instance for liability reasons) not yet want other users of the Data Factory to access. Another example could be highly user- or supplier-specific data which has to be treated confidentially.

Examples for **shared data** could be (possibly processed and/or annotated) real-world sensor data, simulated sensor data, trained AI models, or validation certificates for certain data or models, which are made available to other users of the Data Factory.

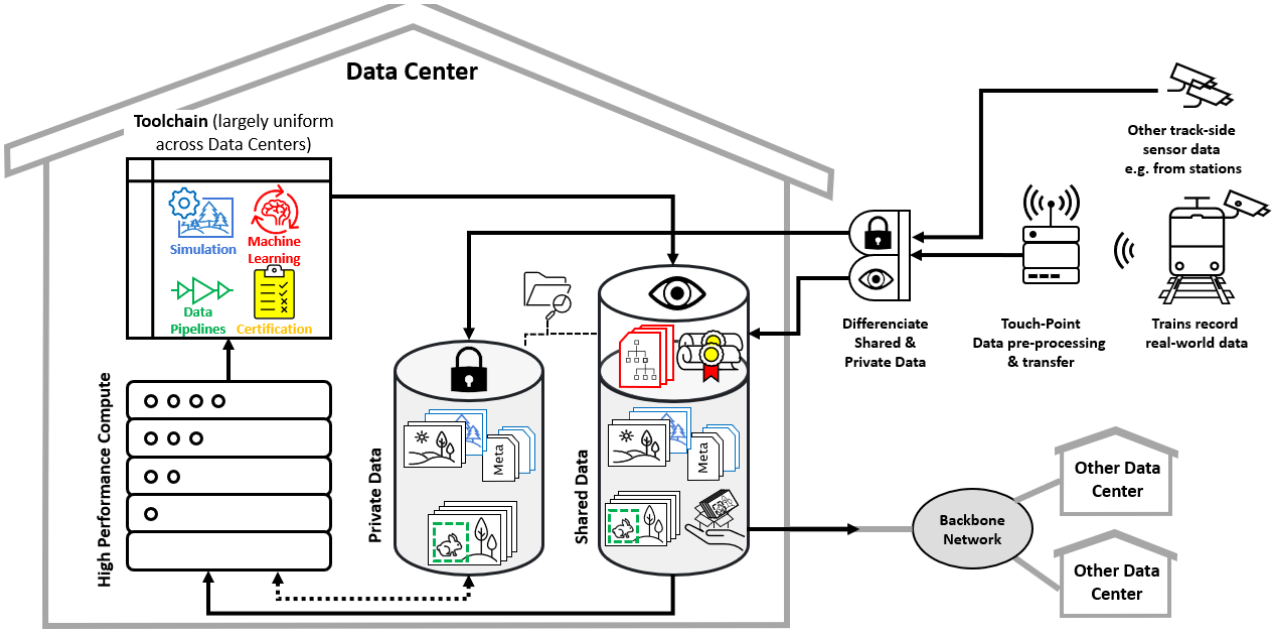

Figure 2. More detailed view of an exemplary data center.

As mentioned before, a key paradigm of the Data Factory is that the Data Centers deploy **Tool Chains** that are largely uniform across all Data Centers and will be managed within a data center itself. In terms of acceptance of a trained AI model, a unified and harmonized tool chain is advisable,

which would simplify approval, as well as when users collaborate across multiple, federated data centers. Nevertheless, it is important to have a common understanding of the data quality and data ontology as well using the same technologies for a better and faster data processing. This is for instance important to enable the creation of AI models based on sensor data from different countries, as is required for cross-border train operation: As the raw sensor data required for AI training may easily be on the order of multiple Petabytes for a single country, it may not be feasible to transfer this among countries, but the data would rather permanently reside in a single Data Center. The initial AI training would then be performed in this Data Center, based on the local sensor data stored there, and then the resulting AI model would be transferred to another Data Center (e.g., in another country), to be further evolved with sensor data stored locally there, to obtain the final AI model. This so-called notion of **Transfer Learning** is only possible if the Tool Chains in the involved Data Centers is largely uniform.

Another benefit of **uniform Tool Chains** is that synergies across Data Centers can be better exploited. For instance, a user needing to do simulations could in principle do this on any Data Center in the Data Factory where compute resources are currently available - and need not rely on a specific local Data Center.

Key overall paradigms that are considered essential for the vision of a Pan-European Data Factory to succeed are:

- Individual stakeholders must be able to maintain **data sovereignty**;
- The overall Pan-European Data Factory must be **decentralized** in the sense that it does not belong to or is controlled by a single entity, and individual stakeholders are always able to expand the Data Factory through further data centers, tools, etc.

  Note: It may of course be that specific entities take a special (e.g., governing) role in a Pan-European Data Factory, such as the European Union Agency for Railways (ERA)
- Individual stakeholders must be able to setup data centers, establish toolchains etc. **customized** to their specific needs, and with elements, tools etc. that are only used **privately** by these stakeholders;
- At the same time, however, it must be possible that entities that do not have own data centers, or do not have much experience in data processing, AI training, etc., have a low entrance barrier to joining and using the Pan-European Data Factory – which of course requires a decent level of infrastructure and tool harmonization across the Data Factory.

3 DATA FACTORY CONTRIBUTION CONCEPT AND ROLES

The concept of a Pan-European Railway Data Factory is also based on the fact that a consortium (i.e., a group of stakeholders) or individual consortium participants (contributors) can participate in it. Furthermore, there are also possibilities to participate in the data and services within a data center by acquiring access through a contribution. This can be done in monetary form, as well as by contributing data and information and also by contributing resources (hardware/software) and further tools. As soon as a participant or a consortium joins, access to the collaborative Data Factory is released accordingly. A role concept and multi-tenancy ensures that access and resources are available.

This approach ends in a **federated European Eco-system** consisting of data and resource sharing among all participants (contributors and consumers). It is assumed that there will basically be two Eco-systems in the end. One Eco-system concerning data and data management and another which deals with the infrastructural system parts.

The means of contribution of a consortium or a contributor can be as follows:

- Financial contribution;
- Providing high-quality data;
- Connecting or contributing resources through hardware;
- Contributing tools;
- Providing external computing power.

In the remainder of this document, roles as defined in Table 2 and illustrated in Figure 3 are used.

Table 2. Roles defined for the Pan-European Data Factory.

Role of contribution	Description
User	The role of a user is, when authorized, to log in into a interconnected facility of the pan-European Data Factory. Note: A user can also be a contributor
Financial Contributor	A user of the system who did financial contribution and can log in into the Data Factory to use the services and tools which are provided.
Data-Provider	A Data-Provider is the role that stores its own high-quality data in the Data Factory. This role also has data sovereignty over this data and can release it to other participants for further processing.
Service-Provider	A Service-Provider define and provide services which consumer of the system can use to access and process data. Also it is possible a Service-Provider connect existing services to a more complex service.
Instance-Provider	An Instance-Provider define where and how a service runs, they take care of pipelines and orchestration of processes.
Node-Provider	A Node-Provider support the Data-Factory with infrastructure and compute power. A Node-Provider provides information where to run services best.

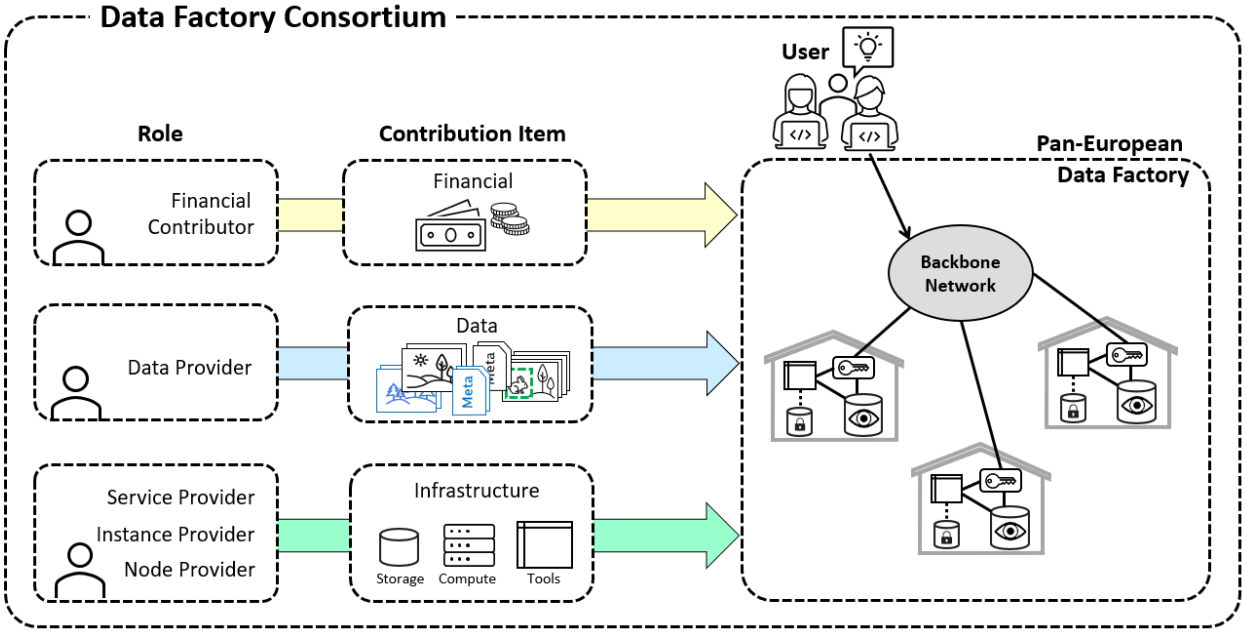

Figure 3. Illustration of the roles involved in the Pan-European Data Factory.

4 SUMMARY AND NEXT STEPS

In the first deliverables D 1.1, D 1.2 and D 1.3 of the CEF2 Railway Data Factory study, the vision and concept of a pan-European Data Factory has been introduced, including the definition of terminology and of roles. Further, representative operational scenarios and use cases have been introduced, and requirements in particular on the underlying connectivity and computing infrastructure have been derived. The work has then further been complemented by considerations related to legal, regulatory and Cyber-security aspects that have to be addressed in the context of a pan-European Data Factory.

This work will serve as an input to the further work in this study, in particular:

- The development of an overall architecture for the pan-European Data Factory, with a particular emphasis on the required pan-European backbone network and edge Cloud facilities, as well as a Cyber-security concept, multi-tenancy support and data management concept;
- A profound commercial and operational assessment of the pan-European Data Factory, including a study on legal and regulatory aspects to be considered.

REFERENCES

[1] Shift2Rail program, see https://rail-research.europa.eu/about-shift2rail/

[2] Europe's Rail program, see https://projects.rail-research.europa.eu/

[3] Sensors4Rail project, see "Sensors4Rail tests sensor-based perception systems in rail operations for the first time," Digitale Schiene Deutschland, 2021. [Online]. Available: https://digitale-schiene-deutschland.de/en/Sensors4Rail

[4] Shift2Rail TAURO project, Horizon 2020 GA 101014984, see https://projects.shift2rail.org/s2r_ipx_n.aspx?p=tauro

[5] R2DATO project, see https://projects.rail-research.europa.eu/eurail-fp2/

[6] P. Neumaier, “Data Factory - "Data Production" for the training of AI software,” Digitale Schiene Deutschland, 2022. [Online]. Available: https://digitale-schiene-deutschland.de/news/en/Data-Factory