CEF2 RailDataFactory

Deliverable 4.1 – Deployment activities description for a pan-European Railway Data Factory

Due date of deliverable: 30/11/2023

Actual submission date: 20/12/2023

Leader/Responsible of this Deliverable: Philippe DAVID (WP 4 lead) / SNCF

Reviewed: **Y**/N

Document status		
Revision	Date	Description
01	09/03/2023	Document template generated
02	15/09/2023	Content transferred from Confluence
03	18/09/2023	First draft complete
04	04/10/2023	Version submitted to advisory board
05	23/11/2023	Final version after addressing all advisory board comments
06	20/12/2023	Version submitted to project officer

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

Start date: 01/01/2023 Duration: 12 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

Name	Company
Philippe David	SNCF
Claire Nicodème	SNCF
Philipp Neumaier	DB
Wolfgang Albert	DB
Patrick Marsch (only editorial)	DB

Note of Thanks

We'd like to thank all Advisory board members who have left valuable comments for this deliverable.

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 detect 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 RailDataFactory is needed, as a part of the overall 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 High-Speed pan-European Railway Data Factory backbone network and data platforms required to realise the vision of the pan-European RailDataFactory.

In this deliverable of the study, the current bottlenecks of transferring & mutating data within a rail network are studied. The report describes the challenges in data connectivity that are currently present while experimenting and deploying (AI) models within the rail industry. A gap between existing rolling stock and future technological advancements is described and proposals are made how the Rail Data Factory can be supplied with a constant flow of data by participating in a pan-European ecosystem.

This report lists the activities needed to organise the process of augmenting the Data Factory activity in Europe on the basis of a core of members.

ABBREVIATIONS, ACRONYMS AND GLOSSARY

Abbreviation	Definition
AI	Artificial Intelligence
ATO	Automatic Train Operation
CPU	Central Processing Unit
Data Augmentation	Adding information in the data. For an image, it can for instance be an incrustation
DF	Data Factory
European organisation	European organisation from which the Pan European Data Factory consortium want approbation. This approbation may be necessary to get European funding if necessary.
(E)EIG	(European) Economic Interest Group
ERA	European Union Agency for Railways
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
HPC	High-Performance Computing: A big set of CPU and GPU resources available for calculation, for instance for Deep Learning
KPI	Key Performance Indicator
ML	Machine Learning
MoU	Memorandum of Understanding
PEDF	Pan-European Data Factory
PEDF consortium	The set of members that constitute the PEDF organisation
PEDF organisation	Legal organisation of the PEDF. It may be an association, a European EIG, a MOU, etc.
GAN	Generative Adversarial Networks
RU	Railway Undertaking
Steering committee	Steering committee or comparable group of members able to take decisions in the PEDF regardless of the chosen organisation
UIC	International Union of Railways

TABLE OF CONTENTS

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 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, an 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.

Cross-border data exchange is crucial for railway undertakings, even if nationally different requirements exist. Through an improved use of technology, for example transfer learning or self-supervision learning with existing data, these national requirements can be partially resolved, and a significant acceleration can be achieved. As an example, transfer learning is a machine learning (ML) technique in which knowledge learned from one task is reused to improve performance on a related task. Among other things, cross-border data exchange enables seamless coordination of the development of fully automated driving and interoperability between different national railway networks and ensures efficient and smooth cross-border operations. The EU Directive (EU) 2016/797 [4] on the interoperability of the rail system provides guidelines and rules to promote such data exchange and ensures a standardised and effective approach across Europe.

1.1 AIM AND SCOPE OF THE CEF2 RAILDATAFACTORY STUDY

The CEF2 RailDataFactory study focuses exactly on vision of a pan-European RailDataFactory (PEDF) for the joint development of fully automated driving. The study, being co-funded through HADEA, aims to assess the feasibility of a pan-European RailDataFactory from technical, economical, legal, regulatory and operational perspectives, and determine key aspects that are required to make a pan-European RailDataFactory a success. For a better understanding of the study's aim and scope, please see Chapter 1.1 in Deliverable 1 [5].

1.2 DELINEATION FROM AND RELATION TO OTHER WORKS

The Shift2Rail project **TAURO** [6] also looks into the development of fully automated rail operation, for instance focusing on developing

- a common database for 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;
- environnement perception technologies (e.g., artificial vision).

The difference of the CEF2 RailDataFactory project is that this puts special emphasis on the **pan-European Railway 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 pan-European RailDataFactory can be realised.

DB Netz AG and the German Centre for Rail Traffic Research (DZSF) have released OSDaR23, the first publicly available multi-sensor data set for the rail sector [7][8]. The data set is aimed at training AI models for fully automated driving and route monitoring in the railway industry. It includes sensor data from various cameras, infrared cameras, LiDARs, radars, and other sensors, recorded in different environments and operating situations, and annotated with labels for different objects and situations. The data set will be utilized in the Data Factory of Digitale Schiene Deutschland to train AI software for environment perception, and more annotated multi-sensor data sets will be created in the future.

The Europe's Rail Innovation Pillar **FP2 R2DATO project** [9], overall focusing on the further development of automated rail operations, also has a work package dedicated to the pan-European RailDataFactory. Here, however, the main focus is on creating first implementations of individual data centres 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 pan-European RailDataFactory activities is ensured through an alignment on use cases and operational scenarios, though the actual focus of the projects is then different.

EU-wide research programs are being carried out on Flagship Project 2: "Digital & Automated up to Autonomous Train Operations" and in this context the European perspective is discussed. In addition, each country and each railway infrastructure provider have its own programs, where there is usually also an exchange within the Innovation and System Pillar in the R2DATO. The participants in this study also work in these bodies and try to reflect the European picture. Within the sector initiative "Digitale Schiene Deutschland", Deutsche Bahn already started to set up some components of the data centre in Germany [10].

1.3 AIM AND STRUCTURE OF THIS DELIVERABLE

This current document is the deliverable D4.1 of the WP4 from the CEF 2 RailDataFactory project.

WP 4 has derived a detailed deployment strategy for the first implementation of the PEDF, starting with the required backbone network and data platform, as a prerequisite to the further development of automated rail operations.

The document is structured as follows:

- In Section 2, possible member categories are described to address the different motivations to participate to the PEDF. These categories express the diversity of services proposed by the Data Factory and the diversity of needs among the members;
- In Section 3, basic principles for the construction of the PEDF are listed, for instance various hypotheses assumed in the remainder of this document, a list of the possible member statuses, from a complete external membership to a full membership, and considerations on different possible governance organisations;
- In Section 4, activities are listed that are needed for the construction of the PEDF in general;
- In Section 5, the activities are detailed that should be taken for a candidate member to join the PDEF;
- and finally, in Section 6, the deliverable is concluded.

2 MEMBER CATEGORIES

A Data Factory may offer numerous services linked to data. It can give access to downloadable data but is also open to new data. Some stakeholders may want to access the services related to Artificial Intelligence (AI) algorithms (e.g., building a data set, training or testing a model). In particular, manufacturers will likely be willing to access sensor data collected by railway companies, as they need this for product development and validation, and would likely not be able to collect a sufficient amount of data themselves. Other members may generate synthetic data, and others are more confident in the role of computer vision algorithms. Overall, members of the Data Factory may have different roles and needs. It is considered essential that the PEDF is open to all these actors.

Therefore, an entity may access different “functionalities” offered by the Data Factory. Depending on the set of wished functions to address, some technical element describing the relation between a user of the Data Factory and the Data Factory itself may vary a lot. The need for network bandwidth, for example, is dependent on the quantity of data accessed.

Table 1 shows possible roles and activities of the members of the PEDF, building upon the roles already defined in Deliverable D 1 [5]. The table does not provide a fixed categorisation of the Data Factory members, but a way to have a general panel of the different use cases. The real profile of each member will be seen through the list of services it may be able to access.

Open Data is a key objective of the PEDF. Please note, however, that external entities solely making use of the Open Data provided by the PEDF, but otherwise not contributing to or using services of the PEDF, in the following referred to as **Open Data users**, are not considered as members of the

PEDF, and hence not covered in Table 1. Nevertheless, Open Data users will, of course, have a substantial impact on the network bandwidth, which has to be considered in the design of the overall system. Technically, we may consider all Open Data users as a unique one with the feature to be located everywhere, with a statistical behaviour to be defined. The Open Data user may access any of the services open to this community according to the decision of the PEDF. The services and data accessible by this community will be chosen depending on the collective interests of the members but it could also be the choice of single member alone to provide a specific (additional) form of Open Data. Among these Open Data users are the technological new-born company, the start-ups, who may find the usage of Open Data an attractive option with a low entrance barrier, compared to a more complicated full membership also involving more commitment and administrative effort.

Table 1. Possible roles and activities of members of the Pan-European Data Factory

	Role of contribution (see D1 [5])						Activity						
Categories	**User**	**Financial Contr.**	**Data Provider**	**Service Provider**	**Instance Prov.**	**Node Provider**	**Produce data**	**Use data (dataset construction, Deep Learning, tests)**	**Add algorithm**	**Add value to the data (annotation, etc.)**	**Use HPC**	**sell data**	**buy data**
Railway company	X	X	X	X	X	X	Real data, data from simulation, data augmentation, Generative Adversarial Networks (GAN), etc.	X	X	Anno-tation, create new KPI	X	X	X
Tech. provider (e.g., for obstacle detection system)	X			X	X	X		X	X	X	X	X	X
Simulation provider	X		X				Scenarios, artificial / synthetic data	X		X	X	X	
Academic institution	X	X		X			Data from simulation, augmentation, GAN, etc.	X	X	X	X		X
Annotation provider			X	X			Annotations	Acce ss only for anno tation		X		X	

Railway manufacturer	X	X		X		X	data from simulation mainly	build dataset for training and testing etc.	(In a private zone only)		X		X
Certification organisation	X						May specify a dataset, generate and issue a certificate	Access for control?					

3 GENERAL PRINCIPLES FOR THE CONSTRUCTION OF THE PEDF

In this chapter, general assumptions and principles related to the construction of the PEDF are listed, which then serve as a basis for the activity descriptions in the subsequent chapters.

3.1 WORKING HYPOTHESIS

In this document, we assume that the following elements are known or given:

- **technical specifications for storing data** or at least the way to store or access the data;
- **technical specifications for working on the data** within the data factory;
- **legal aspects of data**, especially regarding ownership in case of data modification (annotation, augmentation, modification of the data in a general way);
- **technical compatibility** between the different data centre locations and the network performances;
- a **somewhat regular or recurring flow of data** from a member and that an episodic acquisition is not considered (in the process). It means that we may be confident in the fact that automatic processes of taking the data from the train to the ground are available.

It is to be noted that the operational distribution of trained AI models (coming from the PEDF and to be used in operational conditions in the train) are not studied here, as the general volume of these models seems to be small in the first stages (by lack of massive fleet of autonomous train) but also as a model is negligible compared to the volume of data used for its production. Consequently, the general loop of updating the model in the train and getting the "results" of the PEDF is not considered here even if we see an imbrication exposed in our vision of the future.

Data Factories are already (or will be soon) in activity. The DB Data Factory is under construction and operational [7]; the SNCF one is partially operational. These Data Factories are, first, an answer to the need of each company to address the technological issues faced in the different national research initiatives. The PEDF construction (from a technical point of view) should then more being concerned by building the relation between these existing Data Factories (according to the agreement on common standards, interfaces, and toolchains) than creating the whole PEDF

including all national Data Centres from scratch. This means that we have chosen here the enrolment of a new member among an already constituted community to describe the PEDF as it seems to give a general view which encompasses the PEDF operation. Depending on the member profile, the enrolment scenario may differ from something very simple (e.g., grant licenses for downloading Open Data) to something more complex (e.g., adding a partner owning its own data centre with all the interoperability and all the sharing capabilities permitted by the platform). This point of view is an intellectual exercise which may not be applicable strictly speaking for the first members as all the technical specifications are not, at this point, known in detail, the reality may differ from this description.

The network consideration and especially the way toward interconnection of national Data Factories is detailed in Deliverable D 4.2 [11]. Shortly, a PEDF may be seen also as a way to distribute the data among Europe, and the flow of data is not only to be seen as a connection from a data provider to a member, but possibly also as a flow of data through the network of Data Factories, depending on the distribution of services that the members request. Similarly, the flow of data may also encompass the distribution of operational models to the fleet of autonomous trains.

Furthermore, the PEDF is not a complete system that must implement directly all the functionalities (use cases) evoked in our documents to be valid, but, one by one, functions may be switched on and step by step become a complete Data Factory. For example, it may be possible that data augmentation using Generative Adversarial Networks (GANs) or Transformers models will not be available for the PEDF at the beginning. Building a Data Factory is a continuous process of adapting the proposed services to the users, depending on their needs. These needs are also related to the technical adaptations imposed by the use of the produced AI models in real conditions.

3.2 STATUS OF A MEMBER

To become a member, the four membership statuses as shown in Table 2 are considered.

Table 2. Considered membership statuses.

Status	Rank	Comment
Non-member	0	The PEDF is closed to non-members. Even the Open Data part needs a registration.
External member	1	May access the “open data” part of the Data Factory after registration. At this stage, the stakeholder cannot access the other parts of the PEDF. It may also be a way to become a candidate member but is not a mandatory stage.
Candidate member	2	It is a **non-member** or an **external member** which is in the preliminary tasks to specify to the DF steering committee the way it will conform to the “full membership” obligations. At this stage, the duties and obligation of the candidate are known, and the DF steering committee endorse its obligation to welcome the new

		members once it will be able to conform to the listed specifications. The specifications are those which are expected to be able to use a specific service. It may be to be compatible with some kind of file format or to conform to a cybersecurity specification attached to a service. This is mandatory for answering to the investment the future member is doing to reach the following rank.
Full member	3	A full member may access the Data Factory according to the agreed functions with the steering committee of the PEDF. Of course, among the members, some may own a data centre and others may just be users of the DF. For each additional function, a full member wants to access, the process of technical conformity must be performed. A full member contributes to the follow-up of the candidate member and may help it.
Decision maker member	4	It is a **full member** with voting power (on new candidates, etc.). This category is in relation with the chosen organisation (association, MoU, etc.) and is not treated here, given its purely political aspect.

3.3 POSSIBLE GOVERNANCE ORGANISATIONS

Based on the work of this project and on the work of the R2DATO project [9], a governance organisation must be chosen. In general, the following options appear viable:

- **European association** [12]. The European association status has been created. It must imply at least 2 countries in Europe. This could be a very symmetric way to organise the PEDF. This is compatible with Data Factories owned by members as the association is a way to get a governance, not to own by itself the PEDF. However, this may not be the easiest choice, as it needs a critical number of members and a mature consortium in the sense that usages are already known, each national Data Factory is well organised around a data flow etc.
- **National association**. While it is well known and easy to manage, it is not symmetric regarding the different nationalities. This is not a real problem but rather of symbolic nature, and consequently not as good as a European association.
- **Memorandum Of Understanding (MoU)** [13]. An MoU is a type of agreement between two (bilateral) or more (multilateral) parties. It expresses a convergence of will between the parties, indicating an intended common line of action. It is often used either in cases where parties do not imply a legal commitment, or in situations where the parties cannot create a legally enforceable agreement.
- **European Economic Interest Group (EEIG)** [14]. This is a type of legal entity of the European corporate law created on 1985-07-25 under European Community (EC) Council Regulation 2137/85. It is designed to make it easier for companies in different countries to do business together, or to form consortia to take part in EU programs.
- **Peer to peer organisation**. This could be a good solution at low complexity especially when the number of members is still low. It leaves the control and the freedom to each partner regarding the Data Factory organisation. At a moment and with an increasing number of members, another solution must be chosen to avoid complexity.

3.4 CATEGORIES OF ACTIVITIES COVERED IN THIS DOCUMENT

This document lists two categories of activities:

- How to build a kernel for the PEDF,

and

- What must be done by a new candidate member to become a full member.

The first part is detailed in Deliverable D 4.2 [11], as it is more a strategic consideration than an operational study. Nevertheless, a general overview must be presented as this aspect is at the basis of the PEDF construction. The second part is the main objective of this document.

4 LIST OF ACTIVITIES FOR THE CONSTRUCTION OF THE PEDF

The following activities are considered necessary to setup the PEDF. In line with the previous governance considerations, it would be assumed that a European organisation which the PEDF would be associated to should endorse these steps.

- CEF2 Rail Data Factory project (ongoing)
- R2DATO WP7 work package (ongoing) – complementary to the CEF2 RailDataFactory study, the R2DATO WP7 aims to derive the requirements of the Data Factory and outlines the legal implications of defining the Data Factory as a legal entity. A prototype of the Data Factory is also being developed at DB Netz AG, which will be used to carry out simulations and model training on an ML basis. Finally, an Open Data set with sensor data, simulations, annotations, and the digital map will be published
- Choosing a governance organisation
    - the following elements must influence the decision:
        - business model of the PEDF, (part of self-financing, ...)
        - technical model
            - Is there a central server (one of the existing one or another) playing a regulatory role. This element is important regarding the ownership, the confidentiality between partner and some other transversal elements. See Deliverable D 2.2 [15] for more information on this
        - legal model
            - membership regulation
            - conflict resolution rules
            - competent jurisdiction
            - ...
        - commitment type

        - non-European membership. (Permitted or not) This should also address the open data access to non-European company
- Funding (depending on the partner, a national funding may be looked for)
    - self-financing appears to be mandatory
    - how to make the user pay for the data or the services if they do not contribute to provide data
- Specification of the way to extend the PEDF with new members
- Preparation of a step-by-step plan for the construction, expansion and further development of the PEDF
- How to add new functionalities to the PEDF to keep the compatibility with the existing network and to keep the commitment we may have with the others
- Building the core technical specification to make the PEDF working.
    - Possible harmonisation of some of the sensor specifications (sensor, optics, extrinsic parameters, triggering, frequency, etc.)
    - Agreement on data quality requirements and rules
    - Harmonisation of the annotation process
    - Harmonisation of a scenario based (sensor) data simulation process
    - Data formats and interfaces
    - …
- Approbation by the European organisation of a construction strategy and a proposed project
- Building the first version of the PEDF with few members on core services

5 LIST OF ACTIVITIES TO BECOME A MEMBER

In this section, we now focus on the steps that a candidate member should take to join the PEDF and become a full member. The listed steps are detailed in the subsequent sections.

- **Self-evaluation by the candidate**
    - general needs and objectives
    - needs of data and processes
    - member profile
    - data production capacity (if relevant)
    - network connectivity evaluation
    - touch point evaluation (if the candidate provides own Touch Points)

- **Identification of technical adaptation needs**
    - location of the data centre(s) (if any)
    - network adaptation
    - High Performance Computing (HPC)

  - software interfaces conformity
  - data format to implement or to be able to manage
  - cyber-security conformity to the PEDF

- **Data evaluation (if data is present)**
  - quality evaluation
  - value of this data
  - validation of the data annotation
  - validation of the meta data quality (for instance presence of the intrinsic and extrinsic parameters for the camera)

- **Agreement on the list of tasks to perform for the PEDF enrolment**

- **Enrolment validation and agreement**
  - Validation for each requested function
  - Financial and legal agreements
  - Agreement and endorsement from the full members, membership signature

- **Onboarding of the new member**
  - follow-up and assistance to the candidate by the PEDF
  - technical adaptation
  - integration into the PEDF

5.1 SELF-EVALUATION BY THE CANDIDATE

General needs and objectives:

The new member must have an idea of what it seeks in its way to become a candidate. An academic organisation may look for data to work on new models, or a start-up may look for data and HPC to be able to build a model attached to a new device. The request to be candidate will be seen through the external member motivations.

Needs of data and processes:

The candidate member must address its needs in terms of data, algorithms and HPC needs to be able to list the functionalities requested as a member.

A member must create a list of the needed data, so that these needs can be compared with the data already available, so that other members can provide it if possible or get the opportunity to plan it for the future.

This need is expressed in terms of data volumes, HPC power, timing, or period when this data is needed.

Member profile:

A Data Factory is basically a set of tools, data, and functionalities from which a candidate must choose to define its profile. This list requested by the candidate is regarded by the steering committee (or a technical committee which gets this delegation) for approval.

This list constitutes the **member profile**. It can evolve over time, depending on the needs of the obsolete ones.

Data production capacity (if relevant):

Producing data is not mandatory to become a member. If relevant, the production of data is evaluated in term of quantity and quality if the candidate wants to offer it to the PEDF. capacity to be an active member is seen through the amount and the quality of data the partner can share with the other.

Data must also be considered in the way it is recorded or located. It can be stored in a local data centre and offer to the others as coming from this data centre or stored in the cloud in which case the access to the data is different.

Network connectivity evaluation:

A potential member must specify what network connectivity bandwidth is possible to the pan-European Data Factory backbone.

A potential member, in consultation with the consortium, must develop a connectivity plan that includes the implementation timeline and technical specifications. The technical specifications include, among other things, the target bandwidth of the network connection to the pan-European Data Factory backbone.

In this context, the network capacity is seen through 2 aspects:

1) The real need of network bandwidth to share data. This could be described through a temporal diagram in phase with the partner projects. Anticipation may not be an easy task but an evaluation is needed.
2) The available bandwidth given to the partner project regarding the other needs not related to the data factory. As a provider of data, a minimal capacity is required to be able to open the data to the others.

The second aspect is to be declared by the partner to the PEDF.

Connectivity of the candidate member must be at the same level as the functions the new member wants to use. The PEDF must provide a way to evaluate the network connectivity performance depending on the services addressed.

Touch Point evaluation (if the candidate provides own Touch Points):

A key issue of data collection is at the Touch Point where the recorded data from the train must be available on the ground in a constrained time.

Touch Points are assumed to be part of some member's infrastructure (though not every member must necessarily bring in their own Touch Point infrastructure, see also below).

Overall, data is a means of payment within the Data Factory and a data production capacity must be agreed upon with a new candidate member.

Therefore, the data injection capabilities of the Touch Points and the connected network should be outlined and estimated.

A Touch Point may or may not belong to the member. This situation is possible for different reasons. It may be because the Touch Point is a part of the infrastructure or a part of a technical centre belonging to the railway company. In any case, a candidate must validate its access to the network of Touch Point needed for uploading the onboard data.

5.2 IDENTIFICATION OF TECHNICAL ADAPTATION NEEDS

Location of the Data Centre(s):

Depending on the network topology of the candidate (DB or SNCF have a nationwide internal network), there may be multiple options for the location of data centre(s) of new members. Both the distance between data providers and consumers and the international connectivity should be taken into account for this choice. For a cloud only member, this kind of consideration may not be relevant.

This must be weighted by the required function the new member is asking for.

Network adaptation:

In case of an undersized network connectivity, and if modification may improve the connectivity, the candidate member must list the necessary works to comply to the bandwidth specification

High Performance Computing (HPC):

Any HPC needs have to be evaluated to request the corresponding PEDF's IT assets especially for ML training.

Software interfaces conformity:

Depending on the listed needs (user profile) in terms of data, algorithms and HPC, the candidate member must list the software interfaces required by the requested resources.

The basic idea of the Data Factory involves the creation of a uniform toolchain. This toolchain, in particular the functionalities and interfaces, has been commonly defined by the consortium. These software interfaces to existing software and hardware of candidates must be checked for conformity and compatibility for

- data format to implement or to be able to manage: To ease the data sharing, compatibility in the data formats managed by the members is mandatory;
- cyber-security conformity to the PEDF, as covered in Deliverable D 2.2 [15].

5.3 DATA EVALUATION (IF DATA IS PRESENT)

The Data Factory stores a range of data, such as multimodal sensor data and the associated metadata file (file size, creation date, change history, etc.). Furthermore, it encompasses extended metadata like weather, scene information and annotations. The annotations have the railroad relevant object classes like: Trains, tracks, catenary poles, and persons etc. where unique object IDs are assigned. In addition, the objects have attributes such as pose, state, object expression, etc.

This data can come from the train or from trackside.

In the future, other data, not necessarily sensor data, can also be made available in the Data Factory.

Quality evaluation:

To be able to share the data among the members, a good knowledge, and a way to evaluate the quality of the data is essential. Among the metrics available are the meta data joined which may give an idea of the compatibility between what a “client” is looking for and what it may already own. The quality of the data and the way to express it is part of the work in the ERJU R2DATO WP 7 [9].

An open question is how candidates could, before they join the PEDF, be able to assess the quality of their data. An option could be that the PEDF provides algorithms to candidates to do so.

Value of the data:

The value of the data is related to some financial aspect of the data factory whereas the value in term of quality is already seen in the previous paragraph. The value is more related to the expectation by the community of the offered data and may be more seen as a market logic.

The value of the data, either real-world data or simulated data is difficult to measure but it could be useful to define a financial equivalent. This evaluation could be different function of the partners and cannot be an absolute value. To quantify the value of the data, a systematic valuation concept must be created. Subsequently listed are some related attributes:

Data quality and data content attributes can be:

- Sensor data quality in terms of time synchronisation, frame drops, acquisition frequency, etc.;
- Availability of associated metadata like scene descriptions, weather information, annotations, etc.;
- Metadata quality in terms of accuracy and completeness;
- Data content in terms of the amount of railway relevant object classes, railway scenarios, non-regular situations, etc.;
- Data generation attributes:
    - Technical complexity of sensor setups and sensor to train integration;
    - Technical and computational effort and complexity of generated synthetic data;
- Furthermore, the value of the data is strongly impacted by the stakeholder’s needs in terms of developing the use cases and the necessary data for development. It might be beneficial to address certain use cases to the data or data sets.

Validation of the data annotation:

Validation of the annotation is a specific way to deal with the quality of the data seen through the metadata.

Ensuring the quality of the annotations requires a dedicated quality management and annotation quality process. Such a process needs to be established, so that the requirements for the annotations are checked in an automated, semi-automated and manual way. In case the quality is not met, a subsequent rework process must be derived and integrated to ensure the annotation quality, or these data and associated annotation may be stored in a specify category as the quality may reach some quality level required for the owner use case.

Note: Annotations are also some types of metadata, but are discussed here separately, because of their special role.

Validation of the meta data quality:

For example: Presence of the intrinsic and extrinsic parameters for the camera and of the description and usage of it for letting users understand the way to use it.

Other types of meta data, such as data file information, sensor parameters, contextual information, scenario and scene information, intrinsic and extrinsic calibrations and more needs to have a high quality, that needs to be assessed in a meta data quality process.

5.4 Agreement on the List of Tasks to perform for the PEDF Enrolment

In this step, both the existing PEDF members and the candidate member agree on the list of tasks that the candidate (and possibly also existing members) should take to prepare the enrolment of the candidate.

5.5 Enrolment Validation and Agreement

In this phase, the exact enrolment of the member candidate in the PEDF is agreed upon, including and financial and legal aspects that this involves.

Validation for each requested function

To ensure a smooth and consistent operation of the PEDF for the largest possible number of users, it is necessary to regulate the functionalities of the PEDF. A candidate should therefore apply for individual functionalities that it desires (such as specific types of data it would like to access, process or provide, or specific toolchains and other infrastructure it would like to use). It is assumed that the membership of a candidate should be validated individually for each requested functionality.

Financial and legal agreements:

Some membership may include a financial part in the case of an unbalanced situation between the data a candidate can provide and the notion and amount of data or services it would like to obtain from the PEDF.

It is recommended that there is a forum in which these financial and legal aspects are discussed. In particular, the case must be discussed and assessed in which a potential new member may contribute with less data than expected, or data of a different or lower quality than expected.

In general, the monetary valuation of the contributed data is a difficult undertaking, so a policy needs to be established for this. In addition, special usage rights may need to be discussed and defined.

Agreement and endorsement from the full members, membership signature:

Once some verification has been made, the new member may acquire the new candidate status which opens the possibility to launch all the technical tasks to connect the member to the PEDF.

The acceptance of a new member by the Steering Committee requires the signing of the contract, which both the Steering Committee and the candidate must sign.

By signing, a new member is committed to respect the terms of the contract passed with the PEDF and consequently with the other members.

5.6 ONBOARDING OF THE NEW MEMBER

At this stage, the candidate is ready to perform all the required tasks to connect its own Data Factory infrastructure to the PEDF, and can, in accordance with the PEDF steering committee, launch the process.

Follow-up and assistance to the candidate by the PEDF:

In the target image of the PEDF, there is a uniform tool chain and standardized interfaces and formats that must be served. In case of technical questions regarding the interfaces to be served, the candidate should be supported with technical discussions.

Technical adaptation

Execution of all technical adaptations required for the new member to fulfil the previously agreed points.

Integration into the PEDF

This step finally reflects the actual integration of the candidate member into the PEDF, involving tasks such as:

- granting of acces to the physical persons from the new member (creation of users and roles with dedicated access rights);
- opening of the different services the new member is registered to;
- initialisation of the new data into the PEDF;

- initialisation of the new IT assets of the PEDF;
- provision of the documentation regarding the data and IT assets by the new member.

6 CONCLUSION

The general strategy toward the establishment of a Pan-European Data Factory is depicted in Deliverable D 4.2 [11].

In this present document, we have elaborated on different possible categories of Data Factory members, possible membership statuses, and possible governance organisations. We have then detailed the activities expected to be needed to setup the Pan-European Data Factory, and in particular for adding new members to the ecosystem.

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] DIRECTIVE (EU) 2016/797 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL, see https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016L0797

[5] CEF2 RailDataFactory Deliverable 1, "Data Factory Concept, Use Cases and Requirements", Version 1.1, May 2023. [Online]. Available: https://digitale-schiene-deutschland.de/Downloads/2023-04-24_RailDataFactory_CEFII_Deliverable1_published.pdf

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

[7] P. Neumaier, "First freely available multi-sensor data set for machine learning for the development of fully automated driving: OSDaR23", 2023. [Online]. Available: https://digitale-schiene-deutschland.de/en/news/OSDaR23-multi-sensor-data-set-for-machine-learning

[8] Open Sensor Data for Rail 2023, 2023. [Online]. Available: https://data.fid-move.de/dataset/osdar23

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

[10] 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

[11] CEF2 RailDataFactory D 4.2 – "Pan-European Railway Data Factory deployment planning and strategy proposal", 2023. [Online]. Available: https://digitale-schiene-deutschland.de/news/en/Data-Factory

[12] "The meaning of association under EU law", see https://www.europarl.europa.eu/RegData/etudes/STUD/2019/608861/IPOL_STU(2019)608861_EN.pdf

[13] MoU, see https://www.investopedia.com/terms/m/mou.asp

[14] EEIG, see: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31985R2137:en:HTML

[15] CEF-2 RailDataFactory D2.2 – "Technical specifications and available solutions for Identity Access Management (IAM), Data Management and Transfer and Cyber-Security", 2023. [Online]. Available: https://digitale-schiene-deutschland.de/en/news/Pan-European-Railway-Data-Factory