Pan-European Railway Data Factory – Infrastruktur und Ökosystem für einen vollautomatisierten Bahnbetrieb

Pan-European Railway Data Factory – infrastructure and ecosystem for fully automated rail operations

Patrick Marsch | Philipp Neumaier | Philippe David | Bart du Chatinier

Viele europäische Eisenbahnen streben einen automatisierten Bahnbetrieb an. Dies erfordert die Sammlung umfangreicher Sensordaten für das KI-Training (KI, Künstliche Intelligenz), was für eine einzelne Bahn oder einen einzelnen Hersteller schwierig sein kann. Eine Pan-European Railway Data Factory (PEDF) als gemeinsames Infrastruktur- und Partner-Ökosystem erscheint hierfür passend. Dieser Beitrag fasst die Highlights und Ergebnisse der CEF2 RailDataFactory Studie von Deutscher Bahn AG (DB), Société nationale des chemins de fer français (SNCF) und Nederlandse Spoorwegen N.V. (NS) zusammen, die von der Europäischen Exekutivagentur für Gesundheit und Digitales, HADEA, kofinanziert wurde und technische, betriebliche, kommerzielle, rechtliche und strategische Perspektiven untersucht hat.

1 Auf dem Weg zum vollautomatisierten Bahnbetrieb

Derzeit gibt es verschiedene Entwicklungen hin zum vollautomatisierten, fahrerlosen Bahnbetrieb (GoA 4), z. B. in der deutschen Sektor-

Many European railways are striving toward automated rail operations. This requires the collection of extensive sensor data for AI (Artificial Intelligence) training, which may be difficult to achieve for individual railways or suppliers. A Pan-European Railway Data Factory (PEDF), as a joint infrastructure and partner ecosystem, is seen as a suitable way forward for the sector. This article summarises the highlights and findings of the CEF2 RailDataFactory study undertaken by Deutsche Bahn AG (DB), Société nationale des chemins de fer français (SNCF) and Nederlandse Spoorwegen N.V. (NS) and co-funded by the European Health and Digital Executive Agency (HADEA), which has investigated the technical, operating, commercial, legal and strategic perspectives.

1 Toward fully automated rail operations

There are currently various developments heading towards fully automated, driverless rail operations (GoA 4), for instance in

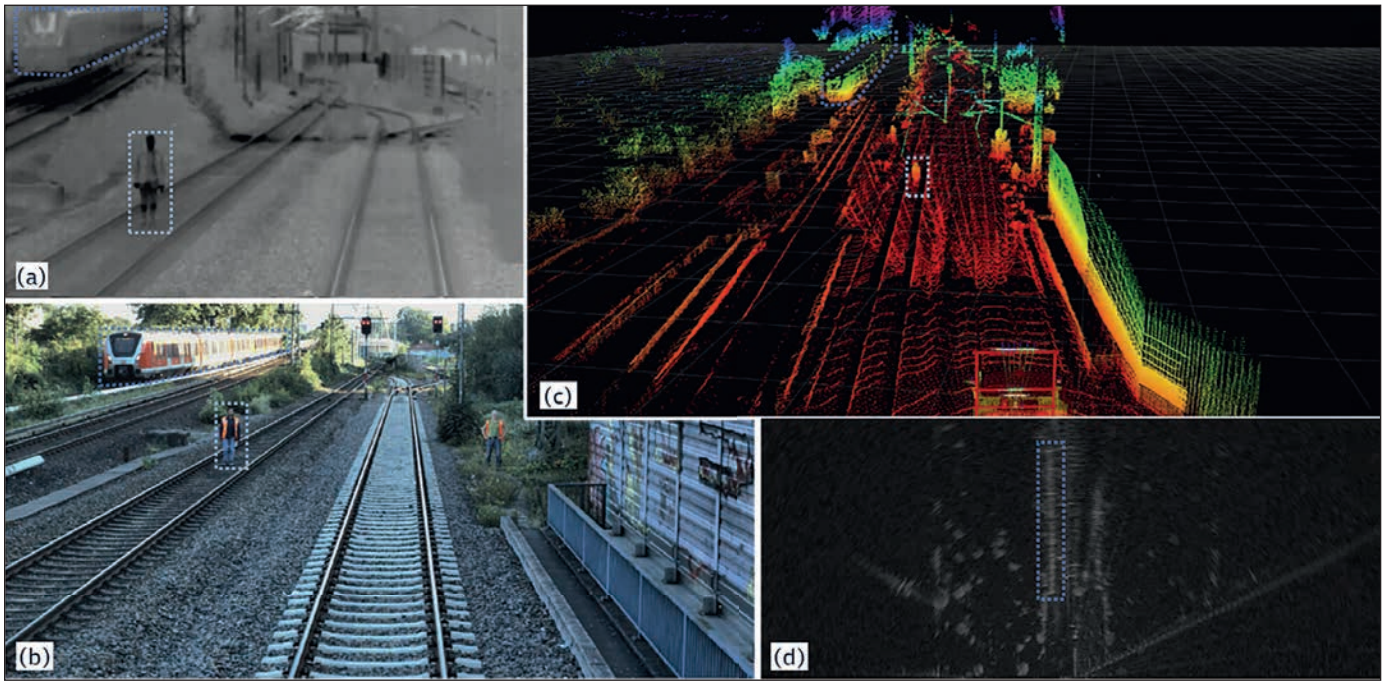

Bild 1: Multimodale Sensordaten von (a) Infrarot-, (b) Kamera-, (c) LiDAR- und (d) Radarsensoren mit beispielhafter Annotation, wie sie für die Entwicklung des vollautomatisierten Fahrens verwendet werden.
Fig. 1: Multimodal sensor data from (a) infrared, (b) camera, (c) LiDAR and (d) radar sensors with exemplary annotation, as utilised for the development of fully automated driving.
Quelle aller Bilder / *Source all figures*: DB AG.

initiative Digitale Schiene Deutschland (DSD) [3] oder dem Programm Europe's Rail [4]. Grade of Automation (GoA) 4 wird von vielen Eisenbahnverkehrsunternehmen (EVU) als Schlüsseltechnologie angesehen, um Antworten auf den Fachkräftemangel zu finden und mit mehr Effizienz und flexiblerer Fahrzeugverfügbarkeit die Kundenzufriedenheit zu steigern.

GoA 4 benötigt sensorbasierte Funktionen zur Wahrnehmung der Umgebung, wie Objekterkennung, Lokalisierung und Störungsmanagement. Diese Funktionen, die größtenteils KI und maschinelles Lernen (ML) beinhalten, erfordern umfangreiche Daten für das Training und die Evaluierung. In Bild 1 sind synchronisierte Sensordaten von (a) Infrarot, (b) Kamera, (c) LiDAR und (d) Radar dargestellt, und Objekte wie Personen und Züge sind markiert. Diese Daten müssen synchron erfasst und anschließend für das KI-Training annotiert werden.

Um seltene Situationen zu erfassen, die im regulären Zugbetrieb kaum vorkommen, ist die Erzeugung künstlicher Sensordaten durch Simulationen erforderlich. Die Kombination realer, annotierter und simulierter Daten ermöglicht das Training von KI-Modellen. Für den späteren Einsatz müssen diese Modelle jedoch zugelassen werden, ein Prozess, der derzeit nicht definiert ist. Eine zertifizierte Toolchain und die Sammlung umfangreicher Testdaten werden als wesentliche Komponenten dafür angesehen.

Für die Durchführung dieser Aufgaben sind Hochleistungsrechenplattformen mit umfangreichen Speicherkapazitäten erforderlich – wir nennen eine solche Plattform „Data Factory". Angesichts der damit verbundenen Komplexität ist es für Hersteller, Eisenbahninfrastrukturunternehmen (EIU) und EVU eine Herausforderung, diese Aufgaben alleine zu bewältigen. Daher scheint eine gemeinschaftliche Pan-European Railway Data Factory (PEDF) eine praktikable Lösung zu sein.

2 Die Vision einer PEDF

Eine PEDF, wie sie im hier zusammengefassten CEF2 RailDataFactory Projekt [1, 2] sowie im Projekt ERJU FP2 R2DATO [5] untersucht wird, wird als Katalysator für die Entwicklung von GoA 4 angenommen. Sie wird es den Bahnen und Herstellern ermöglichen, große Mengen

the German sector initiative Digitale Schiene Deutschland (DSD) [3] or the Europe's Rail program [4]. Many railway undertakings (RUs) see Grade of Automation (GoA) 4 as an attractive option to meet increasing demographic challenges and increase efficiency and flexible vehicle availability, thereby leading to enhanced customer satisfaction.

The development of GoA 4 is reliant on advanced sensor-based functions for perceiving the environment, such as object detection, localisation and incident management. These functions, which largely involve AI and machine learning (ML), demand extensive data for training and evaluation. Fig. 1 shows synchronised sensor data from (a) infrared, (b) camera, (c) LiDAR and (d) radar sources where objects such as people and trains have been marked. This data must be collected synchronously and then subsequently annotated for AI training.

Simulations are necessary to generate artificial sensor data to encompass any rare situations that are hardly ever encountered in regular train operations. Combining real-world sensor data with annotations and simulated data enables AI models to be trained. However, any eventual deployment in trains requires these models to undergo homologation, a process that is currently lacking a defined pathway. Establishing a certified toolchain and accumulating substantial test data are believed to be vital components for this.

Performing these intricate tasks necessitates high performance computing platforms with extensive storage capabilities; we call such platforms "Data Factories". Given the complexity involved, railway suppliers, infrastructure managers (IMs) and RUs might find it challenging to tackle this independently. Therefore, a collaborative Pan-European Railway Data Factory (PEDF) appears to be the most viable solution.

2 The vision of a PEDF

The envisioned PEDF, as studied in the CEF2 RailDataFactory project [1, 2] summarised here and in ERJU FP2 R2DATO [5], is expected to be a key enabler for the development of GoA 4. It

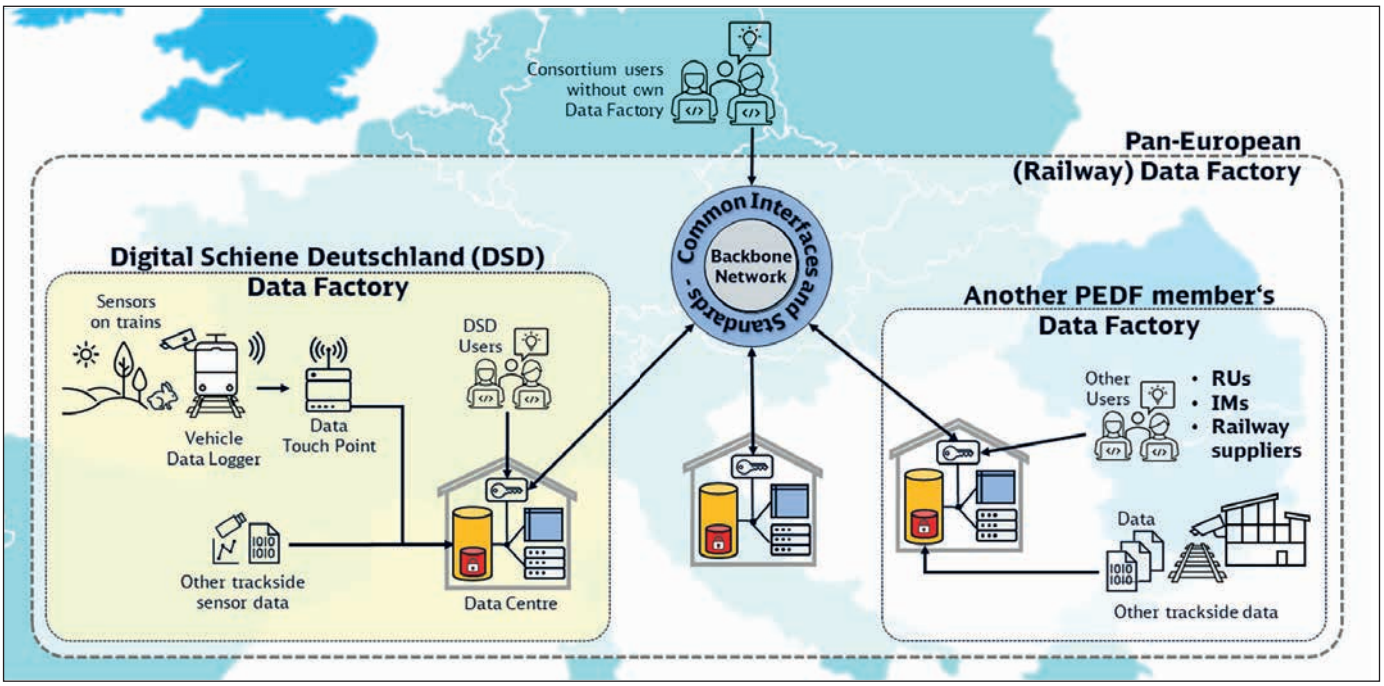

Bild 2: Überblick über die Pan-European Data Factory (PEDF) als Zusammenschluss vieler einzelner Data Factories
Fig. 2: A high-level view of the Pan-European Data Factory (PEDF) as the union of many individual data factories

an Sensordaten und KI-Modellen gemeinsam zu nutzen und Synergien in Bezug auf Prozesse, IT-Infrastruktur und Toolchains für das KI-Training sowie die allgemeine Entwicklung und Zulassung von GoA 4-Lösungen zu schöpfen. Insgesamt wird die PEDF dazu beitragen, wie Eisenbahndaten in ganz Europa erfasst, verarbeitet und genutzt werden, um ein integrierteres, effizienteres und technologisch fortschrittlicheres Eisenbahnnetz zu ermöglichen.

Wie in Bild 2 dargestellt und in [6] vertieft, wird die PEDF ein Verbund einzelner Data Factories in Europa sein, wie z. B. die von DSD eingerichtete Data Factory [7] und ähnliche Infrastrukturen, die von SNCF, NS und anderen aufgebaut werden und welche über ein paneuropäisches Hochgeschwindigkeits-Backbone-Netzwerk miteinander verbunden sind. Data Factories oder Teile davon können von verschiedenen Akteuren wie EIU, EVU, Herstellern usw. betrieben werden.

Die Daten können von verschiedenen Sensorquellen in Zügen oder an der Strecke in die PEDF eingespeist werden, z. B. über sog. Data Touch Points [6], die einen schnellen drahtlosen Datentransfer zwischen den Zügen und der Infrastruktur ermöglichen. Zudem können diese eine Vorauswahl und Vorverarbeitung der Daten vornehmen, bevor sie an die Datenzentren innerhalb jeder Data Factory übertragen werden. Die Touch Points können im Wesentlichen als Edge-Computing-Knoten betrachtet werden.

Das Bild zeigt auch, dass die PEDF-Vision im Wesentlichen zwei Ebenen umfasst:

a) Unterstützung der gemeinsamen Nutzung von Daten in Europa durch gemeinsame Schnittstellen und Datenformate
b) Erschließung weiterer Synergien durch einheitliche Toolchains.

Die in Bild 3 dargestellten Datenzentren sind mit Rechen- und Speicherressourcen ausgestattet und beherbergen eine Reihe von Werkzeugen und Diensten zur Unterstützung der GoA 4-Entwicklung, aber auch des Zugbetriebs im Allgemeinen, des Infrastrukturma-

will allow railways and suppliers to share or jointly work with large amounts of sensor data and AI models and to exploit any synergies concerning the processes, IT infrastructure and toolchains for AI training and the overall development and authorisation of GoA 4 solutions. Overall, the PEDF will contribute to how railway data is collected, processed and utilised across Europe, facilitating a more integrated, efficient, and technologically advanced rail network.

As shown in fig. 2 and detailed in [6], the PEDF will also be an interconnected set of individual Data Factories across Europe, such as the one deployed by DSD [7], and similar infrastructures under construction by SNCF, NS and others, all interlinked through a high-speed pan-European backbone network. Data Factories or parts thereof can be operated by various types of stakeholders such as IMs, RUs, suppliers, etc.

The data can be fed into the PEDF from various onboard or trackside sensor sources, e.g. via so-called data touch points [6] that provide high-speed wireless connectivity between trains and the ground and enable the pre-selection and pre-processing of the data before it is transferred to the data centres within each data factory. The touch points and other parts of the infrastructure can in essence be seen as edge computing nodes.

The figure also indicates that there are basically two levels of ambition in the PEDF vision:

a) support for data sharing across Europe through common interfaces and data formats;
b) exploiting further synergies through uniform toolchains.

The data centres shown in fig. 3 are equipped with computing and storage resources and host an array of tools and services that support not only GoA 4 development, but also train operations in general, infrastructure management, passenger services and predictive maintenance. They also host comprehensive sim-

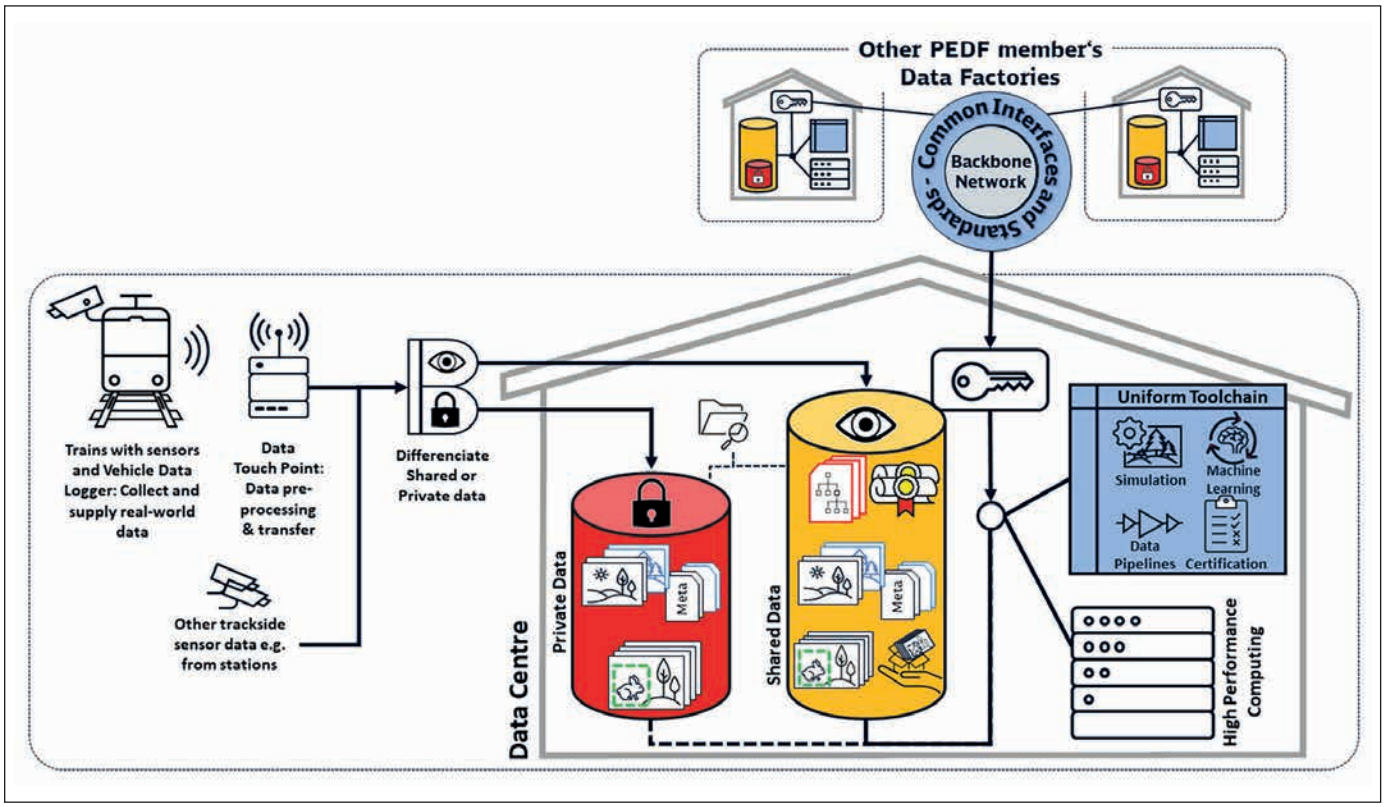

Bild 3: Detailaufnahme eines beispielhaften Datenzentrums in der PEDF
Fig. 3: A detailed view of an exemplary data centre within the PEDF

nagements, der Fahrgastdienste und der vorausschauenden Instandhaltung. Sie verfügen auch über umfassende Simulationsfunktionen zur Erzeugung künstlicher Sensordaten. Wie im Bild dargestellt, wird es geteilte Daten geben, wie z. B. annotierte Sensordaten oder (vor-)trainierte KI-Modelle, die anderen PEDF-Mitgliedern zur Verfügung gestellt werden, sowie private Daten.

Die wichtigsten Nutzer und Begünstigten der PEDF sind:

- Die **EVU** sind die Hauptnutznießer des vollautomatisierten Fahrens und werden dafür verantwortlich sein, dass die eingesetzten Technologien (z. B. die Fahrzeug-Sensorik) alle erforderlichen Betriebsszenarien erfüllen. Die EVU werden daher Data Factories benötigen und von einem entsprechenden europäischen Ökosystem profitieren.
- Die **EIU** werden ein Interesse daran haben, dass die EVU, die das Schienennetz mit einem hohen Automatisierungsgrad befahren wollen, dies auf harmonisierte und koordinierte Weise tun. Vor allem in Ländern mit einem großen EIU kann dieser auch eine naheliegende Entität sein, um einen nationalen Nukleus für die PEDF zu bilden.
- **Eisenbahnhersteller** wie Fahrzeughersteller oder Lieferanten von fahrzeug- oder streckenseitigen Komponenten für den automatisierten Bahnbetrieb werden vom Zugang zu großen Mengen an Sensordaten für die Produktentwicklung, KI-Training usw. profitieren und sicherstellen können, dass ihre Produkte für den europaweiten, grenzüberschreitenden Einsatz geeignet sind.
- **Regulierungs- und Sicherheitsbehörden** werden von genauen und umfassenden Daten profitieren, um bessere politische Entscheidungen zu treffen und die Einhaltung von Vorschriften und Bestimmungen durchzusetzen. Durch die PEDF kann auch die Zulassung von KI-basierten Bahnlösungen europaweit besser koordiniert und harmonisiert werden.
- **Forschungs- und Entwicklungseinrichtungen** wie Universitäten und Innovationszentren können die PEDF für akademische und angewandte Forschung nutzen.

Über diese GoA 4-bezogenen Vorteile hinaus soll die PEDF ein allgemeiner Katalysator für technologischen Fortschritt und Innovation im Eisenbahnsektor sein und Entwicklungen z. B. bei Infrastrukturmanagement, Fahrgastdiensten und vorausschauender Instandhaltung unterstützen. Insgesamt wird erwartet, dass die geplante Infrastruktur und das Ökosystem die betriebliche Effizienz des Eisenbahnsystems verbessern, wirtschaftliche Vorteile bringen und die Kundenzufriedenheit erhöhen.

Die PEDF folgt einer Reihe von zentralen Paradigmen:

- **Datenintegration und Interoperabilität**: Die PEDF ist so konzipiert, dass sie die gemeinsame Erstellung, Verarbeitung und Nutzung von realen und simulierten Eisenbahndaten maximal unterstützt und einen umfassenden Datenspeicher schafft, der allen Beteiligten zugänglich ist.
- **Cybersicherheit und Datenschutz durch Design**: Angesichts der zunehmenden Abhängigkeit des Eisenbahnsektors von Daten und digitalen Technologien ist die Gewährleistung der Sicherheit und Integrität der Daten und ihrer Verarbeitungsinfrastruktur natürlich von größter Bedeutung.
- **Normen und Vorschriften**: Die PEDF soll in ihrem Kerndesign nicht nur die Einhaltung bestehender Normen und Vorschriften gewährleisten, sondern auch die Entwicklung neuer Normen für das digitale Eisenbahnumfeld unterstützen.
- **Datensouveränität und Dezentralisierung**: Trotz der Betonung einer gemeinsamen europäischen Infrastruktur ist es unerlässlich, dass die PEDF den einzelnen Akteuren Autonomie bei der Verwaltung ihrer Daten und der Anpassung von Datenzentren und Toolchains an ihre Bedürfnisse gewährt.

ulation capabilities for the creation of artificial sensor data. As shown in the figure, there will be both shared data (such as annotated sensor data or (semi-)trained AI models that are made available to other PEDF members) and private data.

The main users and beneficiaries of the PEDF are expected to be:

- **RUs** are the main beneficiaries of fully automated driving and will be responsible for ensuring that the deployed technologies (e.g. onboard perception) fulfil all the required operating scenarios. As such, the RUs will need Data Factory facilities and will benefit from a related European ecosystem;
- **IMs** will likely have an interest in ensuring that the RUs running high degrees of automation on their track networks do so in a harmonised and coordinated way. IMs may also be the natural entity to provide the national nucleus for the PEDF, especially in those countries with only one major rail IM;
- **railway suppliers**, such as vehicle manufacturers or suppliers of onboard or trackside components related to automated rail operations, will benefit from access to large amounts of sensor data for product development and AI training etc. and will be able to ensure that their products are fit for crossborder deployment in Europe;
- **regulatory bodies and safety authorities** will benefit from accurate and comprehensive data for better policy-making and regulatory and compliance enforcement. The PEDF may also enable the authorisation of AI-based rail solutions to be better coordinated and harmonised throughout Europe;
- **research and development entities** such as universities and innovation hubs could use the PEDF for academic and applied research.

Beyond these GoA 4-centric benefits, the PEDF is also expected to be an overall catalyst for technological advancement and innovation in the railway sector in that it will also support developments in infrastructure management, passenger services and predictive maintenance, for example. Overall, the envisioned infrastructure and ecosystem are expected to improve operating efficiency in the rail system, to yield economic benefits and to enhance the customer experience.

The PEDF follows a set of key design paradigms:

- **data integration and interoperability**: the PEDF is designed to maximally support the joint creation, processing and use of real-world and simulated rail data, thereby creating a comprehensive data repository that is accessible to all the stakeholders;
- **cybersecurity and data protection by design**: the increasing reliance of the rail sector on data and digital technologies means that ensuring the security and integrity of data and its processing infrastructure is, of course, paramount;
- **standards and regulations**: the core design of the PEDF will not only ensure compliance with the existing standards and regulations, but also support the development of new standards for the digital railway environment;
- **data sovereignty and decentralisation**: despite the emphasis on a joint European infrastructure, it is imperative that the PEDF ensures autonomy for the individual stakeholders in managing their data and customising the data centres and toolchains according to their needs.

3 Architecture and key design paradigms

The architectural design of the PEDF, as pursued in the project, is based on a detailed analysis of the use cases [6] and the

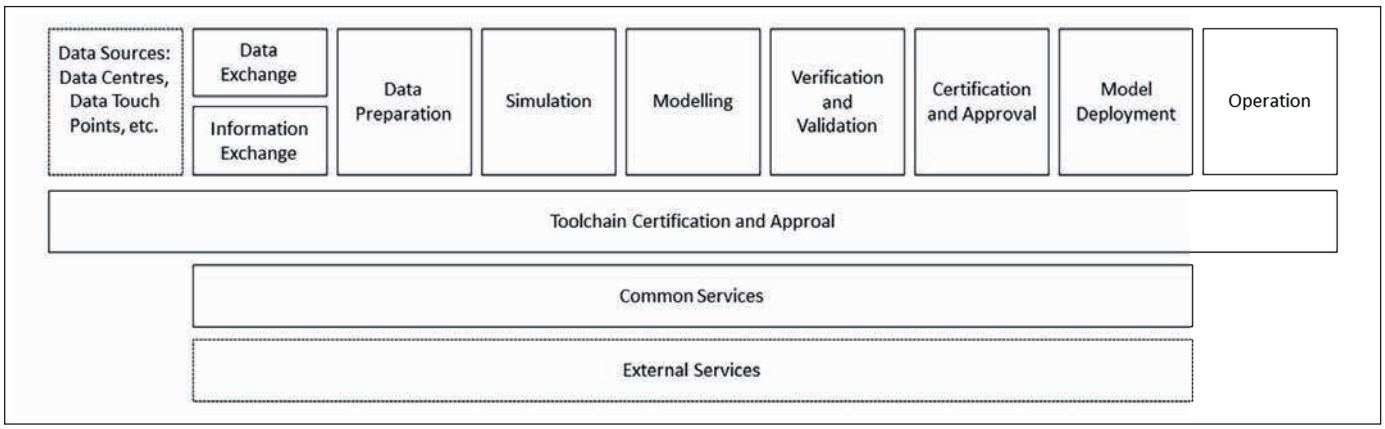

Bild 4: Identifizierte Bausteine der PEDF
Fig. 4: The building blocks identified for the PEDF

3 Architektur und wesentliche Design-Paradigmen

Der im Projekt verfolgte Architekturentwurf der PEDF basiert auf einer detaillierten Analyse von Anwendungsfällen [6] und dem Datenlebenszyklus, der im „Rahmenwerk für Systeme der künstlichen Intelligenz mit maschinellem Lernen" in ISO 23053 [8] angenommen wird. Es wurden die folgenden wesentlichen Bausteine identifiziert, die auch in Bild 4 dargestellt sind:

- **Datenaustausch und Informationsaustausch** beziehen sich auf den Austausch des eigentlichen Dateninhalts (z. B. Sensordaten) und von Metainformationen (z. B. über die Verfügbarkeit neuer Daten) zwischen Datenzentren und Kontaktstellen.
- Die **Datenaufbereitung** umfasst Aspekte der Datenverwaltung wie die Zusammenstellung von Datensätzen und die Datensuche.
- **Modellierung** bezieht sich auf die Modellierung und das Training von KI-Modellen. Dieser Block kann spezifisch für einzelne Data Factories sein, obwohl eine Angleichung innerhalb der PEDF sehr vorteilhaft erscheint.
- **Simulation** wird durchgeführt, um die Datengrundlage zu erhöhen, indem Szenarien simuliert werden, die in realen Tests nicht ohne Weiteres umgesetzt werden können.
- **Verifizierung und Validierung**: Hier werden die Verifizierung (d. h. die Bewertung, ob die Spezifikationen des Systems ordnungsgemäß umgesetzt wurden) und die Validierung (Nachweis, dass das System die betrieblichen Anforderungen erfüllt und seine Funktionen korrekt ausführt) durchgeführt, z. B. gemäß EN 50126 [9].
- **Zertifizierung und Zulassung**: Das KI-System wird für den Einsatz im Eisenbahnbereich zertifiziert. Je nach gefordertem Sicherheitsniveau kann dies verschiedene Bewertungsverfahren und -stellen umfassen.
- **Modellbereitstellung**: Trainierte KI-Modelle werden für den Einsatz im Bahnbetrieb zur Verfügung gestellt.
- **Toolchain-Zertifizierung und -Zulassung**: Zusätzlich kann eine Zertifizierung und Genehmigung für die im gesamten Prozess verwendeten Toolchains und Bausteine erforderlich sein.
- **Gemeinsame Dienste** umfassen z. B. IT-Sicherheit, Netzzugang und Zugangsverwaltung, aber auch Datenimport und -verteilung.
- **Betrieb**: Hier findet der Betrieb der trainierten Modelle statt.

Im Projekt wurde das Design der Bausteine um eine weitere Ebene detailliert, einschließlich Überlegungen zu den Datenflüssen zwischen diesen und einer Analyse geeigneter Implementierungsoptionen [10]. Besonderes Augenmerk wurde auf die folgenden als kritisch angesehenen Aspekte gelegt:

data lifecycle considered in the "framework for artificial intelligence systems using machine learning" from ISO 23053 [8]. The following key building blocks have been identified, as also shown in fig. 4:

- **data exchange and information exchange** refer to the exchange of the actual data content (e.g. sensor data) and of meta information (e.g. on the availability of any new data) among the data centres and touch points;
- **data preparation** covers data management aspects such as dataset composition and data searching;
- **modelling** refers to the actual data modelling and AI model training. This block may be specific to individual data factories, though alignment across the PEDF appears highly beneficial;
- **simulation** is performed to improve the AI system's behaviour through training scenarios that cannot easily be implemented in real world tests. Ideally, training sessions can be executed in parallel, allowing the (re-)training of the AI to be accelerated;
- **verification and validation**: verification (i.e. assessing whether the system specifications have been properly implemented) and validation (demonstrating that the system meets the operating needs and is performing its functions correctly) are performed here, i.e. according to EN 50126 [9];
- **certification and approval**: the AI system is certified for use in the railway environment. This may involve different assessment procedures and entities depending on the required safety level;
- **model deployment**: trained AI models are made available for deployment in rail operations. A key aspect involves the fact that they are exchangeable between individual Data Factories;
- **toolchain certification and approval**: certification and approval may be additionally required for the toolchains and building blocks used in the entire process;
- **common services** cover security, network access and access management, but also data import and distribution;
- **operations**: the trained models are operated here.

The building block design in the project has been taken to a further level of detail, including consideration of the detailed data flows among them and an analysis of the suitable implementation options [10]. Special emphasis has been placed on the following aspects that are considered critical:

- **IT-Organisation und -Betrieb**: Die PEDF sollte als Hybrid aus Cloud- und On-Premise Infrastruktur implementiert werden, wobei das Prinzip der Datengravitation berücksichtigt werden sollte: Große Datenpakete sollten in der Nähe des Ortes verarbeitet werden, an dem sie erzeugt und benötigt werden. Ein intelligenter Workload-Scheduler sollte die Verteilung von Aufgaben über die verfügbare Recheninfrastruktur ermöglichen, und eine Mandantenfähigkeit sollte es verschiedenen Beteiligten erlauben, gemeinsame Infrastrukturen und Toolchains zu nutzen und dabei dennoch eine ausreichende Isolierung zu wahren. Schließlich sollte eine Multi-Site-Fähigkeit Georedundanz und eine schnelle Wiederherstellung im Falle von Katastrophen ermöglichen.
- **Sicherheit**: Die PEDF muss strenge Datenschutzvorschriften wie GDPR einhalten und umfassende Cybersicherheitsmaßnahmen anwenden, einschließlich Verschlüsselung, Erkennung von Angriffen, Protokolle für die Reaktion auf Zwischenfälle und Strategien für die Wiederherstellung im Katastrophenfall. Ein Schwerpunkt im Projekt lag auf der Entwicklung eines föderierten Identitäts- und Zugriffsmanagements (IAM) [11], das es verschiedenen Entitäten ermöglicht, ihre eigenen Benutzeridentitäten zu verwalten und gleichzeitig Interoperabilität und sicheren Datenaustausch im gesamten Netz gewährleistet.
- **Pan-European Backbone-Netzwerk**: Als Kernstück der PEDF muss es eine hohe Bandbreite und geringe Latenzzeiten bieten, nach den Grundsätzen von Zero Trust konzipiert sein und Elastizität im Hinblick auf das erwartete Wachstum der PEDF und sich ändernde Anforderungen bieten [12]. Wichtig ist die Integration bestehender Infrastrukturen einzelner PEDF-Akteure oder bestehender europäischer Infrastrukturen, z. B. von GAIA-X [11].

4 Die PEDF aus betrieblicher, wirtschaftlicher, sicherheitstechnischer und rechtlicher Sicht

Neben der Entwicklung eines technischen Konzepts für die PEDF wurde diese im Rahmen des Projekts aus einer Vielzahl nichttechnischer Perspektiven untersucht. So wurden z. B. in [14] die betrieblichen Herausforderungen untersucht, mit denen EVU konfrontiert sind, z. B. im Zusammenhang mit dem Abruf von Daten an Bord, und

- **IT orchestration and operation:** the PEDF should be implemented in a hybrid consisting of Cloud and on-premise infrastructure, whereby the notion of data gravity is also considered, i.e. large data amounts should be processed near to where they are generated and needed. A smart workload scheduler should allow the tasks to be distributed over the available computer infrastructure and multi-tenancy should allow different stakeholders to share common infrastructure and toolchains while still maintaining sufficient isolation. Finally, multi-site capability should also enable georedundancy and fast recovery in the case of any disasters;
- **security**: the PEDF has to adhere to stringent data protection regulations, such as GDPR, and employ a comprehensive suite of cybersecurity measures, including encryption and intrusion detection, incident response protocols and disaster recovery strategies. Emphasis has been placed on the design of a federated Identity and Access Management (IAM) approach [11], allowing different organisations to manage their own user identities while enabling interoperability and secure data sharing across the network;
- **Pan-European backbone network**: As this lies at the core of the PEDF, it has to provide a high bandwidth and low latency, be designed according to zero trust principles and provide elasticity toward the expected growth of the PEDF and changing requirements[12]. A key aspect involves the integration of the existing infrastructures from the individual PEDF stakeholders or any existing European infrastructure, for instance from GAIA-X [11].

4 The PEDF from an operational, commercial, security and legal perspective

In addition to developing the technical design for the PEDF, the project has also investigated it from a broad range of non-technical perspectives. For instance, the operational challenges that RUs are facing, e.g. related to retrieving onboard data, have been investigated in [14], and the potential economic benefits of an open data infrastructure as envisioned for the PEDF have been studied in [15].

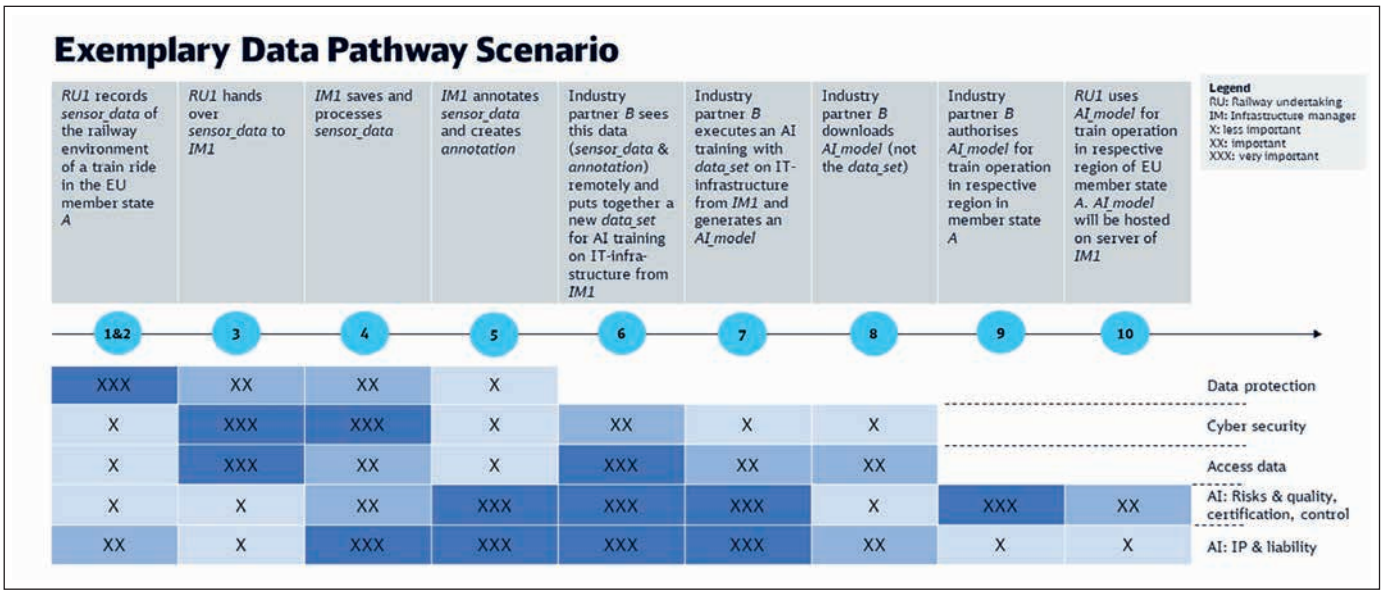

Bild 5: Analyse der rechtlichen und regulatorischen Aspekte, die von der PEDF zu berücksichtigen sind.
Fig. 5: An analysis of the legal and regulatory aspects to be considered by the PEDF.

in [15] der potenzielle wirtschaftliche Nutzen einer offenen Dateninfrastruktur, wie sie für die PEDF vorgesehen ist.
Eine detaillierte Cybersecurity-Risikoanalyse für die PEDF wurde nach den Ansätzen des STRIDE- und Bowtie-Risikomodells durchgeführt [16]. Neben verschiedenen Maßnahmen zur Netz- und Anwendungssicherheit hat sich gezeigt, dass das Datenmanagement für die PEDF von zentraler Bedeutung ist, z. B.

- universelle Standards und Vereinbarungen zu Dateneigentum, Definitionen, Datenformaten und Datenaustauschprotokollen zwischen allen Beteiligten, um eine nahtlose Integration und Zusammenarbeit zu gewährleisten
- Datensicherheit und Risikomanagement, um Bedrohungen, Risiken und Schwachstellen im Zusammenhang mit groß angelegten Datenanwendungen anzugehen und abzumildern.

In einer rechtlichen Studie der PEDF [17] wurden typische Datenpfade von der Erfassung von Sensordaten bis zum Einsatz von KI-Modellen im Bahnbetrieb im Hinblick auf geltende europäische oder nationale Vorschriften und Gesetze analysiert. Wie Bild 5 zeigt, ist Datenschutz in den ersten Schritten der Datenerfassung und -annotierung von Bedeutung, Zertifizierung und Haftung spielen jedoch während des gesamten Datenpfads eine Rolle. Insbesondere die Haftung stellt eine Herausforderung dar: Wenn es im Bahnbetrieb zu einem Unfall kommt, der auf einem KI-Modell basiert, das mit Daten und Toolchains von verschiedenen Akteuren erstellt wurde, wer ist dann haftbar? Diese und andere rechtliche Aspekte werden auch nach dem Projekt weiter untersucht.

5 Eine mögliche Umsetzungsstrategie für die PEDF

Die in [18] beschriebene Einführungsstrategie für die PEDF besteht aus drei Teilen (Bild 6), die jeweils unterschiedliche Aspekte der Entwicklung und Integration behandeln.
Kurzfristig liegt der Schwerpunkt auf maßgeschneiderten technischen und rechtlichen Lösungen für einzelne (FuE-)Projekte innerhalb einzelner (eher „nationaler") Data Factories. Das bedeutet, dass diejenigen, die Bedarf haben, ihre Lösungen entwickeln und gleichzeitig die vertraglichen Verpflichtungen abbilden, um sicherzustellen, dass die Rechte, die Haftung, das geistige Eigentum und das Dateneigentum für die Beteiligten an diesen Projekten abgegrenzt sind.

A detailed cybersecurity risk analysis for the PEDF has been conducted following the STRIDE and Bowtie Risk Model approaches [16]. In addition to various measures related to network and application security, it has also become apparent that data management is crucial for the PEDF, e.g.

- universal standards & agreements concerning the data ownership, definitions, data formats and data exchange protocols among all the stakeholders to ensure seamless integration and cooperation;
- data security & risk management to address and mitigate any data security threats, risks and vulnerabilities associated with the sophisticated models and large-scale data applications.

Typical data pathways from the recording of sensor data through to the deployment of AI models in rail operations have been analysed with regard to the applicable European or national regulations and laws in a legal study of the PEDF [17]. As fig. 5 shows, aspects such as data protection are obviously relevant in the initial steps of data recording and annotation, but certification and liability also play a role throughout the entire data pathway. Liability poses a particular challenge: who is to be held liable if there is an accident in rail operations based on an AI model that has been created with data and toolchains from different stakeholders? These and other legal aspects will continue to be investigated after the project.

5 A possible rollout strategy for the PEDF

The rollout strategy for the PEDF, as detailed in [18], has been divided into three parts (fig. 6), each addressing different aspects of its development and integration.
In the short term, the focus is on customised technical and legal solutions for individual (R&D) projects within the individual (somewhat "national") Data Factories. This means that those who need these will develop their own solutions while mapping the contractual obligations to ensure that any rights, liability, intellectual property and data ownership are delineated for the stakeholders in these projects.
In the medium term, the strategy aims to bring the ongoing activities and developments of the individual Data Factories in line with the standardisation requirements of the PEDF. This will be

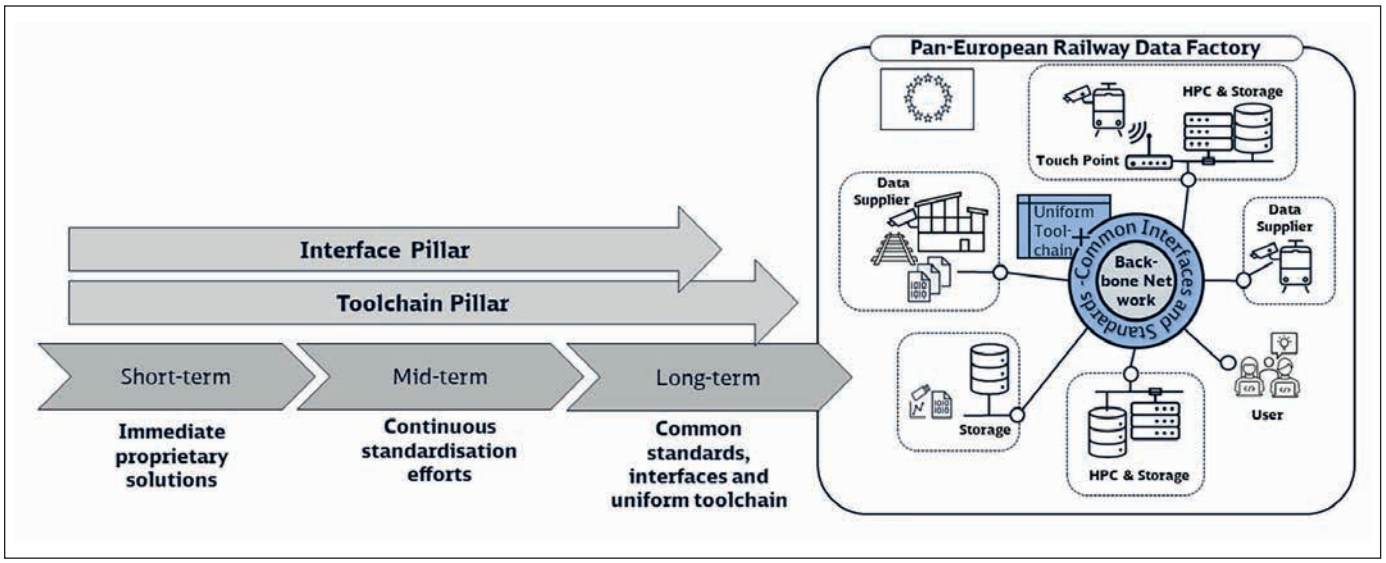

Bild 6: Kurz-, mittel- und langfristige Strategie für die Einrichtung der PEDF, gestützt auf die Schnittstellen- und Toolchain-Säulen
Fig. 6: The short, mid and long-term strategies for setting up the PEDF supported by the interface and toolchain pillars

Mittelfristig zielt die Strategie darauf ab, die laufenden Aktivitäten und Entwicklungen der einzelnen Data Factories mit den Standardisierungsanforderungen der PEDF in Einklang zu bringen. Dies soll durch einen Übergang erreicht werden, bei dem die Einhaltung der PEDF-Standards und -Protokolle schrittweise angeglichen wird, um schließlich die Integration der unabhängigen Data Factories in die PEDF zu gewährleisten.

Teil der langfristigen Strategie ist die umfassende Koordinierung der Standardisierungsbemühungen mit Schwerpunkt auf Datenqualität, -formaten, Schnittstellen und Interkonnektivität über ein europaweites Hochgeschwindigkeits-Backbone sowie eine einheitliche Toolchain. Darüber hinaus sollte eine für diese Aspekte zuständige Organisation geschaffen werden, idealerweise eine europäische, unabhängige und gemeinnützige Einrichtung.

Die folgenden Beteiligungspfade für PEDF-Mitglieder werden vorgeschlagen:

- Die **Schnittstellen-Säule** unterstreicht die für verschiedene Anwendungsfälle erforderliche Flexibilität und erleichtert den Datenaustausch zwischen den Mitgliedern durch gemeinsame Schnittstellen und Standards. Dieser Ansatz stützt sich auf die Koordinierung von Datenformaten, Datenorganisation, Datenqualität, Annotationen, Modellarchitekturen, Datenanonymisierung und Datenschutz, Sensoren und Datenerfassung.
- Darüber hinaus zielt die **Toolchain-Säule** auf eine Harmonisierung der gesamten Toolchain ab. Sie bietet zwar Vielseitigkeit und das Potenzial für vollständige Interoperabilität bei der Datenerfassung, -verarbeitung, -qualitätssicherung, -übertragung, -zugriff und -simulation sowie beim Training und bei der Bewertung von ML-Modellen, doch kann ihre starre Spezifikation die spezifischen Wünsche der einzelnen Mitglieder ggf. nur teilweise erfüllen.

Die Stärke der Strategie liegt in der pragmatischen, schrittweisen Entwicklung einer effizienten Zusammenarbeit, die die PEDF zu einer vielseitigen und effektiven paneuropäischen Initiative macht.

6 Zusammenfassung

Eine PEDF als gemeinsame europäische Infrastruktur und Ökosystem von Eisenbahnen, Herstellern, Behörden, Hochschulen und anderen wird als Schlüssel für die Entwicklung eines vollautomatisierten Bahnbetriebs angesehen und soll dem europäischen Bahnsystem erhebliche Vorteile bringen. Das CEF2-Projekt RailDataFactory [1, 2] hat dies aus vielen Perspektiven untersucht und eine mögliche Einführungsstrategie für die PEDF entworfen. Während viele Fragen beantwortet und die Relevanz der PEDF bestätigt werden konnte, müssen viele Details weiter untersucht werden, z. B. im Rahmen des Projektes ERJU FP2 R2DATO [5] und nachfolgenden Projekten.

Die Autoren danken Alexander Heine, Gertjan Tamis, Guillaume Bussieras, Jan van Gelder, Jens Dalitz, Julian Wissmann, Stéphane Callet, Vanessa Fong Tin Joen-Baarh, Waseem Ul Aslam Peer, Wolfgang Albert und den Beiratsmitgliedern für ihren Beitrag zu diesem Projekt.

AUTOREN | AUTHORS

Dr. Patrick Marsch
Head of Connectivity, IT and Data Platforms and IT/OT Security
Digitale Schiene Deutschland (DSD), DB InfraGO AG
Anschrift / *Address*: Stresemannstraße 123 A, D-10963 Berlin
E-Mail: patrick.marsch@deutschebahn.com

Dr. Philipp Neumaier
Head of Data Factory & Data Acquisition & Engineering
Digitale Schiene Deutschland (DSD), DB InfraGO AG
Anschrift / *Address*: Stresemannstraße 123 A, D-10963 Berlin
E-Mail: philipp.neumaier@deutschebahn.com

Philippe David
Scientific officer
Autonomous train program
SNCF
Anschrift / *Address*: 1-3 avenue François Mitterrand, F-93212 La Plaine Saint Denis
E-Mail: philippe.david@sncf.fr

Bart du Chatinier
IT/OT integration manager
Nederlandse Spoorwegen (NS)
Anschrift / *Address*: Laan van Puntenburg 100, NL-3500 HA Utrecht
E-Mail: bart.duchatinier@ns.nl

LITERATUR | LITERATURE

[1] CEF2 RailDataFactory study. Available: https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/how-to-participate/org-details/999999999/project/101095272/program/43251567/details
[2] CEF2 RailDataFactory Deliverable 5.2, "Final Study Results", Januar 2024. [Online]. Available: https://digitale-schiene-deutschland.de/en/news/2023/Pan-European-Railway-Data-Factory
[3] Digitale Schiene Deutschland, see https://digitale-schiene-deutschland.de
[4] Europe's Rail program, see https://projects.rail-research.europa.eu/
[5] R2DATO project, see https://projects.rail-research.europa.eu/eurail-fp2/
[6] CEF2 RailDataFactory Deliverable 1, "Data Factory Concept, Use Cases and Requirements", Version 1.1, April 2023. [Online]. Available: https://digitale-schiene-deutschland.de/en/news/2023/Pan-European-Railway-Data-Factory
[7] Neumaier, P.: "Data Factory - „Data Production" for the training of AI software," Digitale Schiene Deutschland, 2022. [Online]. Available: https://digitale-schiene-deutschland.de/en/news/2022/Data-Factory
[8] ISO/IEC 23053:2022, "Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)", 2022
[9] EN 50126, "Railway Applications. The Specification and Demonstration of Reliability, Availability, Maintainability and Safety (RAMS) Generic RAMS Process", 2018
[10] CEF2 RailDataFactory D 2.1 – "Technical specifications and available solutions for building blocks, components, Cloud / hybrid-Cloud and Edge-Orchestration & Operational concept", August 2023. [Online]. Siehe Link unter [6]
[11] CEF2 RailDataFactory D 2.2 – "Technical specifications and available solutions for Identity Access Management (IAM), Data Management and Transfer and Cyber-Security", August 2023. [Online]. Siehe Link unter [6]
[12] CEF2 RailDataFactory D 2.3 – "High-speed pan-European Railway Data Factory Backbone Network", August 2023. [Online]. Siehe Link unter [6]
[13] GAIA-X, see https://gaia-x.eu/
[14] CEF2 RailDataFactory D 3.1 – "Report of bottlenecks data application in rolling stock", November 2023. [Online]. Siehe Link unter [6]
[15] CEF2 RailDataFactory D 3.2 – "Business case whether open data infrastructure would be attractive for European rail", November 2023. [Online]. Siehe Link unter [6]
[16] CEF2 RailDataFactory D 3.3 – "Description of cybersecurity vulnerabilities, threat scenario's and usable standards to mitigate associated risks", November 2023. [Online]. Siehe Link unter [6]
[17] CEF2 D 3.4 CEF2 RailDataFactory D 3.4 – "Legal and regulatory assessment catalogue", November 2023. [Online]. Siehe Link unter [6]
[18] CEF2 RailDataFactory, D 4.2 – "Pan-European Railway Data Factory deployment planning and strategy proposal", December 2023. [Online]. Siehe Link unter [6]

achieved by means of a transition that gradually aligns compliance with PEDF standards and protocols, thereby ensuring the eventual integration of the independent data factories into the PEDF.
Part of the long-term strategy involves the comprehensive coordination of the standardisation efforts with a focus on data quality, formats, interfaces and interconnectivity via a pan-European high-speed backbone, as well as a uniform toolchain. In addition, an organisation responsible for these aspects should also be created, ideally a European, independent and non-profit body.
The following participation paths are proposed for future PEDF members:

- the **interface pillar** emphasises the flexibility needed for many different use cases and facilitates the data exchange among the members by means of common interfaces and standards. This approach is based on the coordination of data formats, data organisation, data quality, annotations, model architectures, data anonymisation and data privacy, sensors and data collection;
- in addition, the **toolchain pillar** aims to achieve the harmonisation of the entire toolchain. While it offers versatility and the potential for full interoperability in data collection, processing, quality assurance, transfer, access and simulation, as well as training and the evaluation of ML models, its rigid specification may only partially fulfil the specific wishes of individual members.

The strength of the strategy lies in the pragmatic step-by-step development of efficient cooperation and integration, thereby developing the PEDF into a versatile and effective pan-European initiative.

6 Summary

A PEDF, as a joint European infrastructure and ecosystem for railways, suppliers, authorities, academia and others, is seen as a key enabler for the development of fully automated rail operations and is expected to provide substantial benefits to the evolving European railway system. The CEF2 RailDataFactory project [1, 2] has studied this from various perspectives and laid out a possible deployment strategy for the PEDF. While many questions could be answered and the relevance of the PEDF could be confirmed, many details need further study, for instance in ERJU FP2 R2DATO [5] and subsequent projects.
The authors would like to thank Alexander Heine, Gertjan Tamis, Guillaume Bussieras, Jan van Gelder, Jens Dalitz, Julian Wissmann, Stéphane Callet, Vanessa Fong Tin Joen-Baarh, Waseem Ul Aslam Peer, Wolfgang Albert and the advisory board members for their contributions to the project. ■