CEF2 RailDataFactory

Deliverable 2.2 - Technical specifications and available solutions for Identity Access Management (IAM), Data Management and Transfer and Cyber-Security

Due date of deliverable: 30/04/2023

Actual submission date: 17/07/2023

Leader/Responsible of this Deliverable: Julian Wissmann (WP 2 lead) / DB Netz AG

Reviewed: Y

Document status		
Revision	Date	Description
01	09/03/2023	Document template generated
02	26/05/2023	Content transferred from Confluence
03	26/05/2023	First draft complete
04	06/06/2023	Version submitted to advisory board
05	30/06/2023	Final version after addressing of all advisory board comments
06	17/07/2023	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: 9 months

KNOWLEDGEMENTS

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
Alexander Heine	DB
Guillaume Bussieras	Capgemini
Jens Dalitz	DB
Julian Wissmann	DB
Mayank Singh	DB
Patrick Denzler	DB
Philipp Neumaier	DB
Waseem Ul Aslam Peer	DB
Wolfgang Albert	DB
Patrick Marsch (only editorial)	DB

Note of Thanks

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

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 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 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 pan-European Data Factory backbone network and data platforms required to realize the vision of the Data Factory.

In this deliverable of the study, the IAM and data management concepts for the pan-European Data Factory are introduced, key concepts are defined, and related requirements in particular on the federated identity management and different aspects of data management are provided. Altogether, these requirements serve as a basis for the further work in this study.

ABBREVIATIONS AND ACRONYMS

Abbreviation	Definition
AD	Active Directory
B2C	Business to Customer
AI	Artificial Intelligence
AM	Access Management
CEF	Connecting Europe Facilities
DGA	European Data Governance Act
ERA	European Union Agency for Railways
GDPR	General Data Protection Regulation
GoA4	Grade of Automation 4
HADEA	European Health and Digital Executive Agency
IAM	Identity and Access Management
IdP	Identity Provider
IM	Infrastructure Manager
ISMS	Information Security Management System
LDAP	Lightweight Directory Access Protocol
MFA	Multi-Factor-Authentication
ML	Machine Learning
MPLS	Multiprotocol Label Switching
OIDC	OpenID Connect
OSM	Open Street Map
PII	Personally Identifiable Information
PKI	Public Key Infrastructure
RBAC	Role Based Access Control
ROS	Robot Operating System
RU	Railway Undertaking
SAML	Security Assertion Markup Language
SSO	Single Sign On
TLS	Transport Layer Security
VPN	Virtual Private Network
XML	eXtensible Markup language

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. For a better understanding of the studys aim and scope, please see Chapter 1.1 in Deliverable 1 [4].

1.2 DELINEATION FROM AND RELATION TO OTHER WORKS

The Shift2Rail project **TAURO** [5] 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 Europe's rail Innovation Pillar **FP2 R2DATO project** [6], 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 [7].

1.3 Aim and Structure of this Deliverable

This current document is the deliverable D 2.2 of the CEF2 RailDataFactory project, covering the IAM security concept of the envisioned pan-European Data Factory specifically aimed at providing a concept, requirements and available solutions in the context of a federated, distributed pan-European Data Factory. Additionally, this deliverable will cover a data management concept aimed at providing a data architecture, input on data classification and formats, data quality, data governance and data security.

The aim of the document is to obtain early feedback and possible additions from the sector on the high level IAM concept, the requirements identified within and the data management 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 management concept, legal and business aspects will be developed.

The remainder of this document is structured as follows:

- In Chapter 2, the overarching security framework is described;
- Chapter 3 describes the IAM concept including its purpose, term definitions, user types, roles, context and lifecycle, as well as requirements, and market available solutions
- In Chapter 4, the data management concept is introduced;
- In Chapter 5, the data transfer concept is introduced;
- In Chapter 6, a summary and outlook are provided.

2 OVERARCHING SECURITY FRAMEWORK

As part of the concept, design, testing, operation, maintenance and decommissioning of the pan-European data factory, an overarching cyber security framework must be defined and implemented as required by the EU regulation (e.g., NIS 2 Directive).
The cyber security framework for the pan-European Data Factory is made-up of three pillars, as shown in Figure 1:

- **Organisational framework** defining security roles, processes, guidelines, management of cybersecurity risks, compliance and legal aspects;
- **Engineering process framework** defining a security process over the whole life cycle of the systems / products and components. As part of this framework, the connection and relationships between safety and security should also be addressed;
- **Continuous cyber security activities** such as security monitoring, continuous risk identification / mitigation and incident response / Business Continuity.

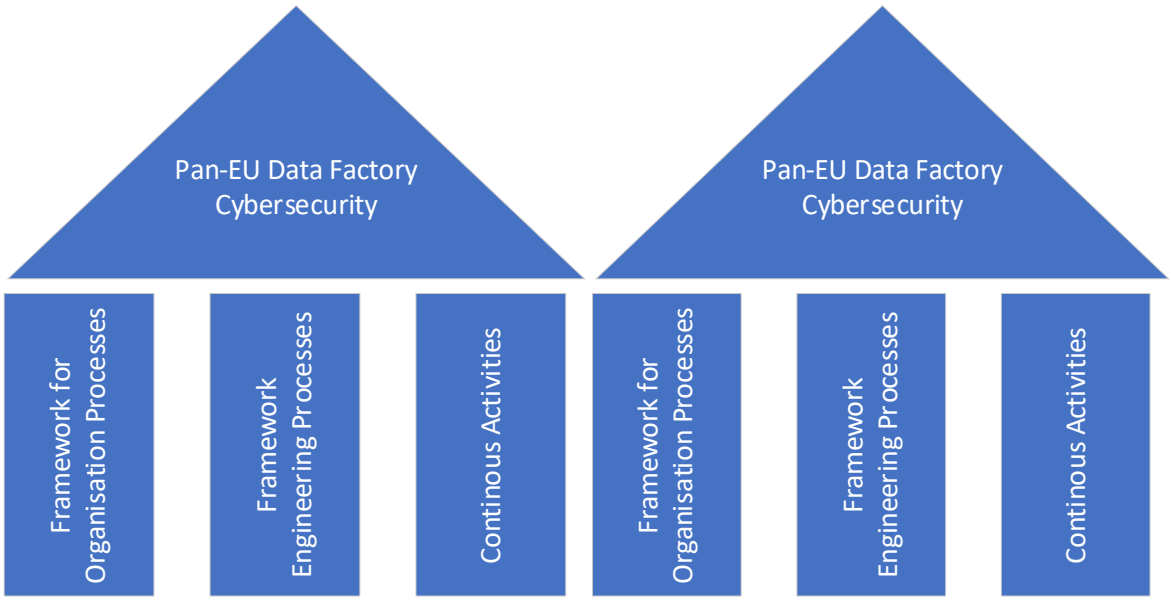

Figure 1: Overarching cybersecurity framework.

Operational Taxonomy for Data Security and Privacy

Taxonomy of operational Security and Privacy considerations for the Data Factory consists of five categories as described in Figure 2 and as referenced in NIST SP 1500-4r1:

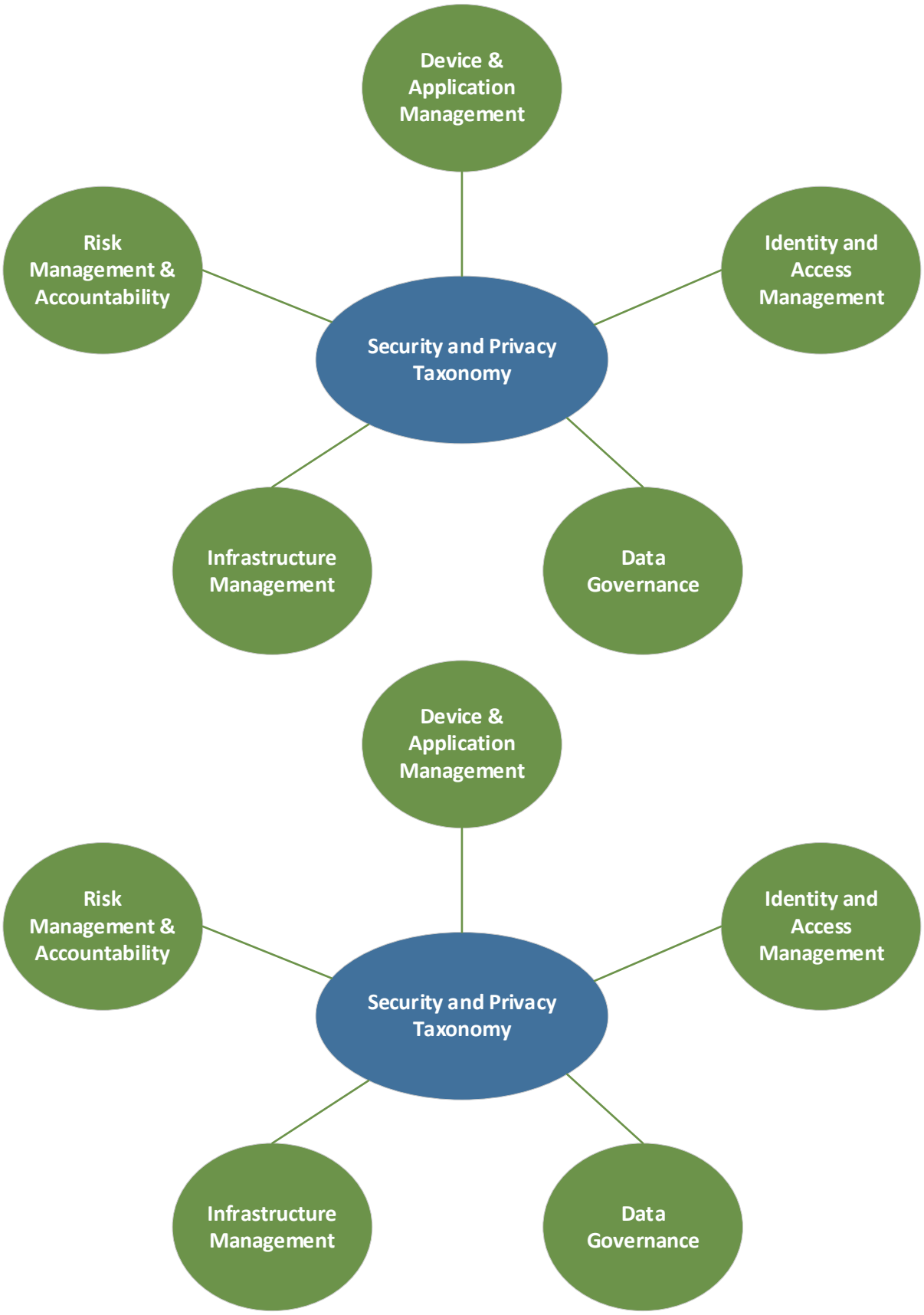

Figure 2: Data Factory security taxonomy.

Device & Application Management

Devices and applications (incl. tool chain) shall be registered, and their configuration shall be managed along their whole lifecycles.

- Policy Enforcement for:
  - Asset Management
  - Governance Model
  - Quality, health and configuration management

Identity & Access Management

Authentication and Access Control systems are among the most critical security components and is made-up of both process and technical activities:

- Identity Assurance Level which consists in identity proofing to ensure that an applicant is who they be up to a certain level of certitude;
- Authentication Assurance Level establishes that a subject attempting to access a digital service is in control of the technologies used to authenticate;
- Federation and Assertions which allows the conveyance of authentication attributes across networked systems;
- Digital Identity Risk Management to address false identity claim, authentication and federation errors (authenticator or identity assertions are compromised).

A description and specification of an Identity & Access Management mechanism for the pan-European Data Factory is provided in Chapter 3.

Data Governance

It refers to the overall management of the availability, usability, integrity, and security of the data employed in the pan-European Data Factory. Data Governance is addressed in Section 4.8.

Infrastructure Management

Infrastructure management involves security and privacy considerations related to hardware / software / network operation and maintenance. Some topics related to infrastructure management are listed below:

- Threat and vulnerability management;
- Monitoring and alerting;
- Configuration management;
- Malware resilience;
- System redundancy;
- System recovery.

Risk Management

A uniform methodology to evaluate the business risk level by conducting regular risk assessments shall be identified and applied. The purpose of the risk assessment is to discuss threats incl. attack vectors exploiting technical and/or procedural vulnerabilities leading to the compromise of the confidentiality, integrity and availability of infrastructure, applications and/or data. There are numerous standards, policies and frameworks describing Threat and Risk Analysis methodologies such as ISO 27005, NIST SP 800-37, BSI, MITRE ATT&CK, STRIDE etc. and every organisation must evaluate which methodologies suits its needs according to the organisational and operational context. It is not the purpose of this document to determine the adequate risk assessment methodology of the pan-European Data Factory. However, a high-level threat analysis is described in Section 4.9.

3 IAM CONCEPT

3.1 PURPOSE

Identity and Access Management (IAM) is a cornerstone to provide secure access to the data and services of the pan-European Data Factory. It shall provide the basis for a sovereign data exchange, ensure privacy consideration as well as access and usage rights. The purpose of the Identity and Access Management (IAM) is to verify and validate user identities and to grant or deny access to resources for a given context. IAM shall provide a consistent identification and authentication approach across all entities of the pan-European Data Factory. To ensure a secure interoperability and trust mechanism among all participants, a governance framework incl. authentication and access control policies shall be defined and agreed-upon with all entities and organisations.

Common vocabulary for Identity and Access Management that will be addressed in this document covers identification, authentication & authorisation, credentials management, federated identity and access management, as detailed in Table 1. The document also addresses the lifecycle process of the IAM product and of user accounts.

This part of this document focuses primarily on conceptual modelling, requirement specification and key considerations for a pan-European IAM and remains agnostic regarding technology and vendor.

3.2 DEFINITIONS

Table 1: IAM terminology.

Terms	Description
Authenticator Assurance Level (AAL)	A category describing the strength of the authentication process
Access Management System	Access management component to grant or deny access to resources (services, data, nodes)
Assertion	A statement from a “verifier” to a “relying party” that contains information about a user. Assertions may also contain verified attributes.
Authentication	Digital authentication is the process of determining the validity of one or more authenticators used to claim a Digital Identity
Authorization	Access privileges granted to a user, program, or the process or act of granting those privileges
Consumer	A role of a Participant with users & devices, searching / ordering services and maintaining a business relationship to Providers
Digital Identity	An attribute or set of attributes that uniquely describe a subject (individual or asset) in the pan-European Data Factory ecosystem

Federation Assurance Level (FAL)	A category describing the assertion protocol used by the federation to communicate authentication and attribute information (if applicable)
Federated Identity model	Conveyance of identity and authentication information across a set of networked systems.
Identity Assurance Level (IAL)	A category that conveys the degree of confidence that the applicant's claimed identity is their real identity.
Identity Proofing	Identity proofing establishes that a subject (a natural person for instance) is who he claims to be
Identity Provider (IdP)	A trusted component that issues and/or registers subscriber authenticators and issues electronic credentials to subscribers. This component is also sometimes called Credential System Provider (CSP). The Identity Provider issues assertions derived from those credentials.
Participant	Object in the pan-European Data Factory such as providers, consumers and devices
Relying Party	An entity that relies upon the subscriber's authenticator(s) and credentials or a verifier's assertion of a claimant's identity, typically to process a transaction or grant access to information or a system.
Provider	A role of a participant responsible for making data, services or asset available in the pan-European data factory ecosystem
Verifier	An entity that verifies the claimant's identity by verifying the claimant's possession and control of one or two authenticators using an authentication protocol.

3.3 SUMMARY OF USER TYPES: CONSUMER AND PROVIDER

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.

The means of contribution of a consortium or a contributor can be as follows (see also D 1 [4]):

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

It is possible to categorise the pan-European Data Factory Users in two major categories: Consumer and Provider. Table 2 describes the different users and their attributes. A user can be a provider as well as a consumer at the same time.

Table 2: Data Factory user types.

User Types	Description
User	A user is a natural or legal person having a technical account authorized or not to log in into a facility of the pan-European data factory. A user can also be a consumer and/or a provider incl. contributor. A user can also be named “a subscriber” as for example in NIST SP 800-63-3: Digital Identity Guideline.
Consumer	A user of the pan-European Data Factory who can browse / order services and usually maintains a business relationships with providers. A consumer consumes service instances and/or data and can also provide them to their own end-users.
Data-Owner	This role also has sovereignty over this data and can release it to other users (as consumers) for further processing. The Data Owner defines restrictions to the usage of his data in the form of policies.
Service Provider	A Service Provider defines and provides services to data owners (in the role of a consumer in that case) to expose the data. As an example: a Database service instance.
Instance Provider	An Instance Provider defines where and how a service runs, they take care of pipelines and orchestration of processes. Also instance providers can consume further Instance services
Node-Provider	A Node Provider supports the data factory with infrastructure and compute power. A Node Provider provides information and infrastructure on where to run services. In that context the Service Instance Provider becomes the consumer of nodes.

3.4 Definition of Roles

In Deliverable 1 [4] the role “user” was defined. Now, in the context of IAM, this now further differentiated into the following kinds of users as listed in Table 3.

Table 3: IAM user roles.

Categories	Role Name	Description
AI/Data	AI/ML Specialist	Designs, develops, and modifies AI applications, tools, and/or other solutions to enable successful accomplishment of mission objectives.
	Data Analyst	Analyses and interprets data from multiple disparate, sources and builds visualizations and dashboards to report insights.
	Data Architect	Designs a system’s data models, data flow, interfaces, and infrastructure to meet the

		information requirements of a business or mission.
	Data Officer	Holds responsibility for developing, promoting, and overseeing implementation of data as an asset and the establishment and enforcement of data-related strategies, policies, standards, processes, and governance.
	Data Scientist	Uncovers and explains actionable insights from data by combining scientific method, math and statistics, specialised programming, advanced analytics, AI, and storytelling.
IT	Database Administrator	Administers databases and/or data management systems that allow for the storage, query, and utilisation of data.
	Enterprise Architect	Develops and maintains business, systems, and information processes to support enterprise mission needs; develops information technology (IT) rules and requirements that describe baseline and target architectures.
	Network Operations Specialist	Plans, implements, and operates network services/systems, to include hardware and virtual environments.
	System Administrator	Installs, configures, troubleshoots, and maintains hardware, software, and administers system accounts.
	System developer	Designs, develops, tests, and evaluates information systems throughout the systems development lifecycle.
	System Testing & Evaluation Specialist	Plans, prepares, and executes tests of systems to evaluate results against specifications and requirements as well as analyse / report test results.
	Technical Support Specialist	Provides technical support to customers who need assistance utilising client level hardware and software in accordance with established or approved organizational process components. (i.e., Master Incident Management Plan, when applicable).
Software Engineering	DevSecOps Specialist	Selects/Deploys/Maintains the set of Continuous Integration/Continuous Deployment (CI/CD) tools and processes used by the development team and/or maintains the deployed software product

		and ensures observability and security across the lifecycle.
	Software Developer	Executes software planning, requirements, risk management, design, development, architecture, fulfils estimation, configuration management, quality, security, and tests using software development methodologies, architectural structures, viewpoints, styles, design decisions, and frameworks across all lifecycle phases.
	Software/Cloud Architect	Manages and identifies program high-level technical specifications, which may include application design, cloud computing strategy and adoption, and integration of software applications into a functioning system to meet requirements.
	Systems Security Analyst	Responsible for analysis and development of systems/software security through the product lifecycle to include integration, testing, operations and maintenance.

3.5 IDENTITY AND ACCESS MANAGEMENT CONTEXT AND LIFECYCLE ACTIVITIES

3.5.1 Context

The identity and access management must focus on ensuring the interoperability of identification, authentication and authorisation, based on conceptual design and architecture by adopting accepted architectures, protocols, international open standards, industry best practices, frameworks and policies.

3.5.2 Federated Identity Approach

As shown in D 2.1, the pan-European Data Factory is expected to be comprised of national data centers provided and operated by different entities. Consequentially, it is not realistic to assume a single centralised Access Management System, but rather each provider of a data center or other resource (e.g., toolchain) would provide and manage its own Access Management. Likewise, different entities would operate their own Identity System. To facilitate a joint usage of the pan-European Data Factory infrastructure, a Federated Identity Architecture could then be used that allows the conveyance of identity and authentication information across a set of networked systems or different domains. This is elaborated in more detail in this section.

A Federated Identity Architecture allows for the conveyance of identity and authentication information across a set of networked systems or different domains. It allows a given Identity Service Provider to provide authentication attributes and also user attributes to a number of separately-administered relying parties. A relying party is for instance a specific organization composing the pan-European Data Factory ecosystem (e.g., Railway undertaking of a specific EU country). In this scenario, every entity has its own Identity Service Provider that controls the identity data of the entity's user. With a federated IAM mode, a user of a pre-registered entity can log-in and authenticate against its home Identity System and access resources in external domains based on the specific access policies of

the target domain / organisation. Federated Identity Protocols include SAML, OpenID and OpenID Connect.

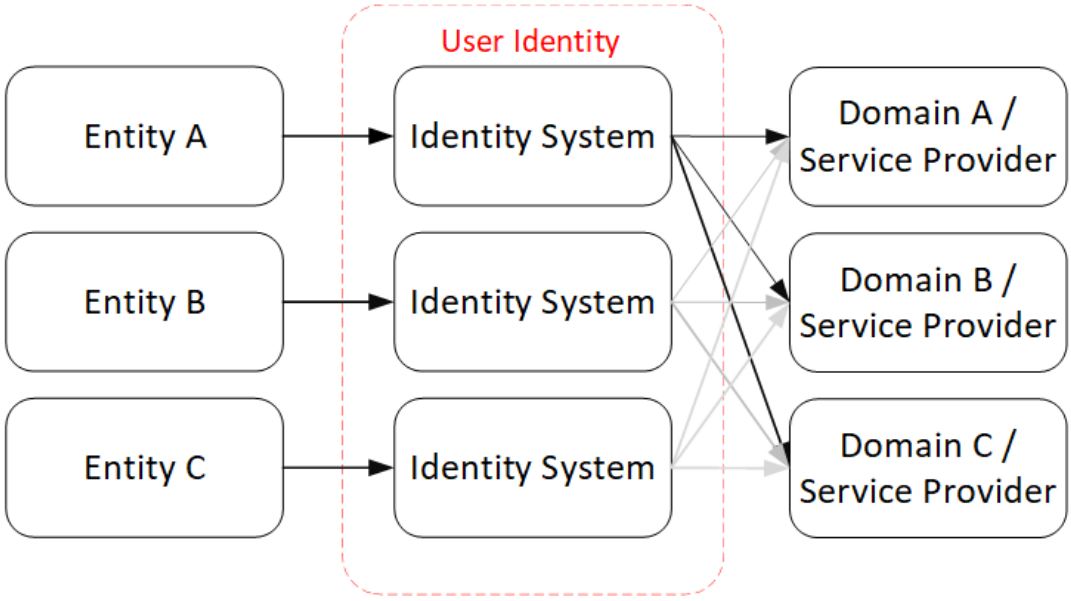

Figure 3: Federated identity approach.

3.5.3 Detailed Pan-European Data Factory Federated Identity Model

As part of a federated Identity and Access Management mechanism, several components must interact with each other in process steps. An overview of the different components and the way they interact with each other is described in Figure 4 and explained in Table 4.

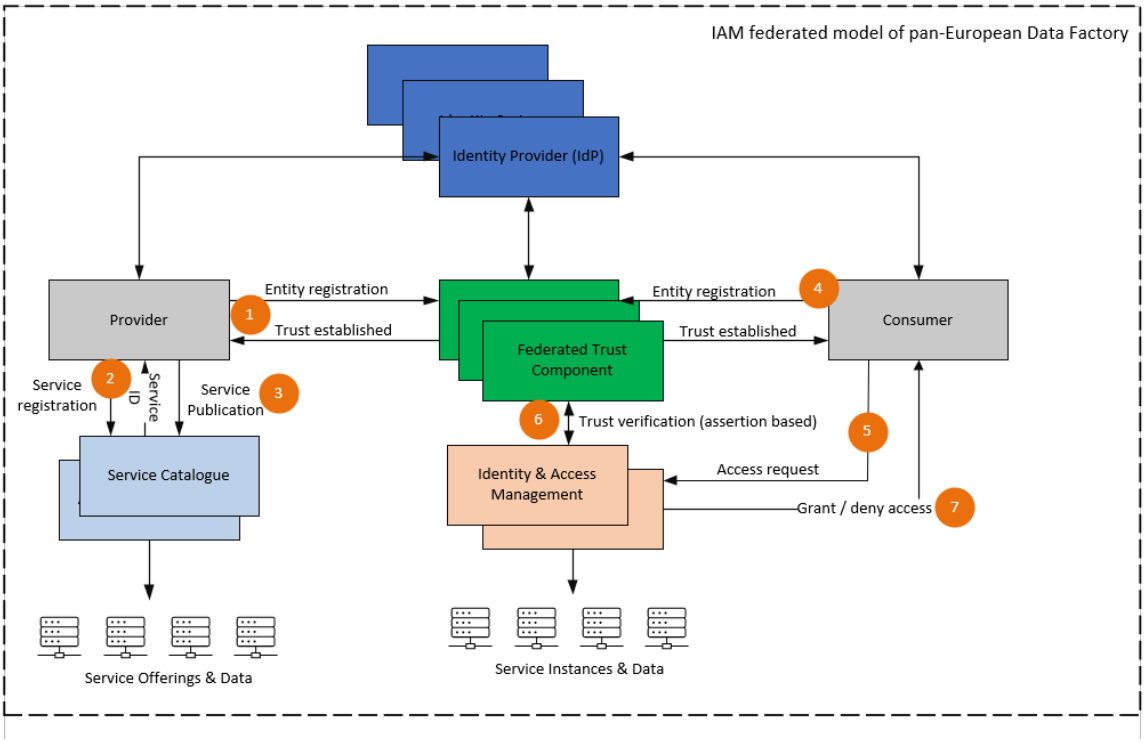

Figure 4: IAM service registration flow.

Table 4: IAM service registration flow.

Step	Purpose	Description
1	Provider Entity Registration	The provider entity / organization must first register in the pan-European Data Factory. In order to do that, the provider shall give details about its own Identity System / Identity Provider. The Identity System must fulfil IAL, AAL, FAL levels required by the pan-European Data Factory. Once the identity system of the provider has been verified, it becomes a federated identity system of the pan-European Data Factory
2	Service Offering Registration	The provider is able to register a Service into the pan-European Data Factory service catalogue.
3	Service Publication	Once released, the service is published in the Service Catalogue that can be browsed by the consumers seeking for services
4	Consumer Entity Registration	The consumer entity / organization must first register in the pan-European Data Factory. In order to do that, the consumer shall give details about its own and existing Identity System / Identity Provider. Once the identity system of the consumer has been verified, it becomes a federated identity system of the pan-European Data Factory
5	Service Access Request	The consumer reaches the Identity & Access Management component of the service provider the consumer wants to reach out
6	Trust verification	The consumer logs-in with its usual and pre-defined authentication mechanisms against his “home” Identity Provider. The Identity Provider forwards a set of attributes (Identity, Token, Timestamp...) via assertions to the Identity & Access Management component of the service provider. The Identity & Access Management component will verify the attributes and determine the type of access authorization the consumer has based on specific roles.
7	Grant / Deny Access	The Access Management grants or denies access to the requested services after evaluating the passed consumer’s attributes.

3.5.4 Authentication & Authorization Procedure

Figure 5 now describes the different steps as part of an authentication and authorisation procedure in a federated identity model.

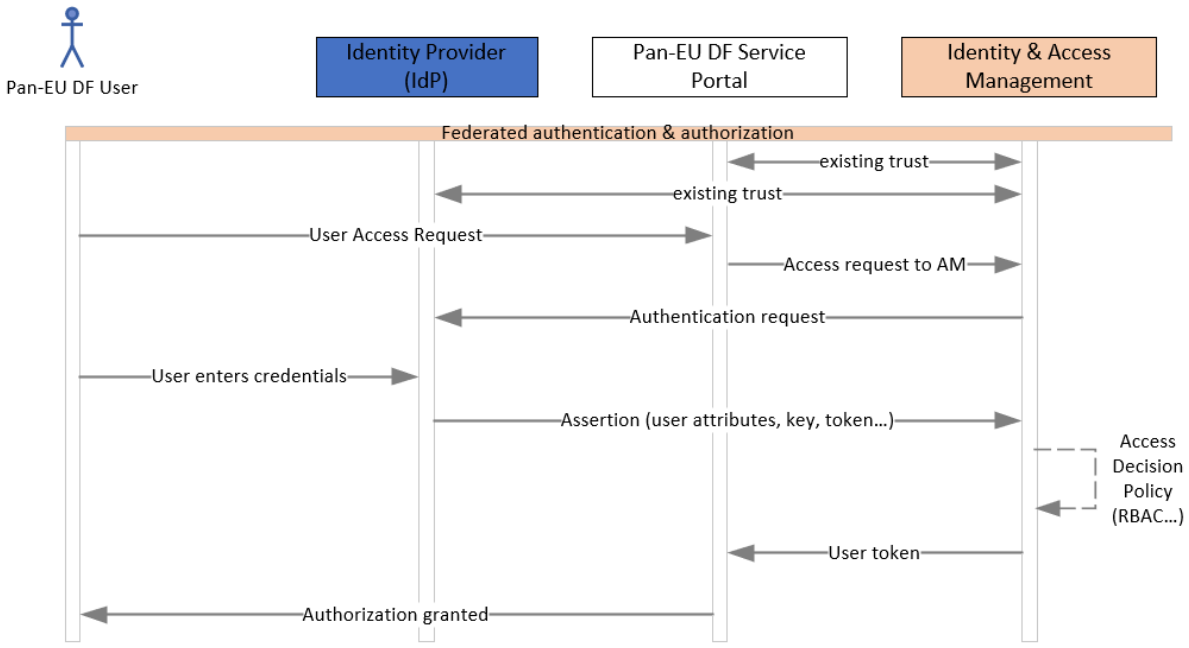

Figure 5: Authentication and authorization procedure.

A user accesses the pan-European Data Factory portal to reach a specific service. The Access Portal of the service provider forwards the login request to the trusted and known Access Management system. The Access Management system requests authentication from the home Identity Provider of the user. The user will provide its credentials (MFA, password...) against its home Identity Service Provider. The home Identity Service Provider validates the user inputs and provides attributes (also called assertion) about user's identity to the trusted Access Management system of the service provider. After verification of the user attributes (Name, role...), the access management service provider grants or denies access to the user based on pre-defined rules and policies: attributes, role-based access, etc...

3.5.5 IAM System Lifecycle

A system lifecycle consists of a series of identifiable stages or process steps through which an item goes, from its conception to disposal. Figure 6 describes the main phases of the IAM system lifecycle which complies with railway CENELEC standards and also with industrial security standards such as IEC 62443.

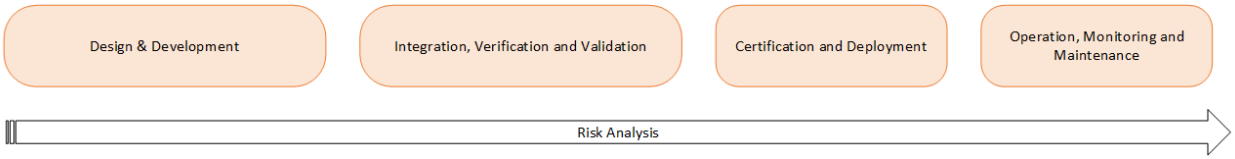

Figure 6: IAM system lifecycle phases.

Design and Developments Phase

The purpose of this phase is to specify the system requirements especially for IAM (see also Table 6), draft the system architecture and develop the Products accordingly.

Integration, Verification and Validation Phase

The purpose of this phase is to integrate and assemble the final product / system usually made up of different components. Within this phase validation tests are performed to demonstrate that the final product fulfils the requirements for IAM (see also Table 6).

Certification and Deployment Phase

The purpose of this phase is to deploy the final product and it usually terminates with a Site Acceptance Test. Certification and approval process from authorities also terminates during this phase.

Operation, Monitoring and Maintenance

The purpose of this phase is to operate, maintain and administer the IAM product. The product shall also be monitored continuously to detect intentional or unintentional incidents such as cybersecurity events, performance and health issues.

During the complete product / system lifecycle, a risk analysis must be continuously maintained to identify threats, vulnerabilities and to assess the potential impact and consequences of an incident.

3.5.6 Identity Lifecycle

A proper identity and authentication lifecycle management process is required to cover the areas of onboarding (e.g., issuing an authenticator), maintenance and administration (e.g., the authenticators) and to perform correct and secure the offboarding (e.g., withdrawing an authenticator from an identity). This is also described in Table 5.

Table 5: IAM identify lifecycle details.

IAM lifecycle activity	Description
Onboarding	Enrollment and identity proofing of applicants that wish to gain access to resources of the pan-European Data Factory. As a result of a successful identity proofing transaction, an authenticator is issued and bound to the verified identity of the applicant.
Operating / Maintaining	An updated authenticator shall be bound before existing authenticator's expiration. Compromised authenticators (Loss, theft, unauthorised duplication) are subject to suspension, revocation or destruction.
Offboarding	Revocation or termination of an authenticator refers to the removal of the binding of an authenticator and an identity. An online or digital identity ceases to exist when requested by the user or when the IdP determines that the user no longer meets its eligibility requirements.

3.6 IAM REQUIREMENTS SPECIFICATION

3.6.1 IAM Functional & Technical Requirements

Each organisation should pay attention to the following functional and technical requirements and implement them separately in order to make the system as secure as possible. These requirements enable the functioning of the pan-European Rail Data Factory as a loosely coupled system.

Table 6: IAM functional requirements.

#	Description
R1	Identity and access management shall be done in a distributed manner, thus not relying on a central and unique Identity Service Provider
R2	Identity and Access Management shall support federation (incl. SSO) where federation allows the conveyance of authentication attributes across domains of the pan-European Data Factory
R3	Identity and access management shall uniquely identify users and objects with an attribute or set of attributes in the pan-European Data Factory context
R4	Identity and Access Management shall support identification and authentication of devices and technical components such as Network devices, IoT Devices, Applications and Services
R5	Identity and Access Management shall support roles where a role defines what a user can do in the context of the pan-European Data Factory
R6	Identity and Access Management shall support attribute-based access control over defined subjects and objects of the pan-European Data Factory
R7	Identity and Access Management shall support SAML and OpenID Connect industry standard protocols for authentication and assertion
R8	Trust decision and trust enforcement shall be done by an Identity System Provider based on the security requirements and policies of the pan-European data factory
R9	Identity and Access Management shall support Identity proofing of users in accordance with the required Identity Assurance Level (IAL) usually pre-defined by an organisation policy
R10	Identity and Access Management shall support authentication process in accordance with the required Authenticator Assurance Level (AAL) usually pre-defined by an organisation policy
R11	Identity and Access Management shall support assertion mechanism in accordance with the required Federation Assurance Level (FAL) defined by an organisation policy
R12	Identity and Access Management shall support “Data Protection by Design and by Default” to ensure GDPR compliance
R13	Identity and Access Management shall support state of the art encryption mechanism

R14	All activities shall be logged and monitored

3.6.2 IAM Security Policy Requirements

Each organization should pay attention to the policy requirements listed in Table 7 and implement them separately in order to make the system as secure as possible. These requirements follow ISO 27001 and ISO 62443 as well.

Table 7: IAM security policy requirements.

#	Description
R1	Access Control Policy shall be developed, documented and disseminated among organisational users. The Access Control Policy shall address purpose, scope, roles, responsibilities and compliance.
R2	Role-Based Access Control (incl. privileged user accounts) Policy over defined subjects and objects shall be developed, documented and disseminated among an organisation
R3	Attribute-based Access Control Policy over defined subjects and objects shall be developed, documented and disseminated among an organisation
R4	An Identification and authentication policy shall be developed, documented and disseminated among organisational users. The Access Control Policy shall address purpose, scope, roles, responsibilities and compliance.
R5	The risk management strategy and risk assessment of the organisation shall always be taken into account when establishing IAM policies
R6	A specific identification and authentication policy for devices and /or type of devices shall be developed
R7	An incident management policy with associated procedures shall be established to manage IAM security incidents to cope with personal data and GDPR regulation

3.6.3 IAM On and Offboarding Requirements

Table 8: IAM on- and offboarding requirements.

#	Description
R1	An onboarding and offboarding process to register provider and consumer entities shall be defined
R2	An onboarding and offboarding process to register Services and Nodes shall be defined
R3	For both a consumer and a provider, there shall exist an Identity System proving the identity of the entity and its users
R4	The Identity System of a consumer or a provider must be identified and enrolled during the registration process of the pan-European Data Factory

R5	The Identity System of a consumer or a provider must meet IAL, AAL, and FAL levels of the pan-European Data Factory and those must be verified during the registration process
R6	Revocation of a trusted Identity System shall occur at the request of a registered entity
R7	Revocation of a trusted Identity System shall occur as soon as the Identity System no longer meets its eligibility requirements (e.g., IAL, AAL. FAL levels)
R8	Revocation of a trusted Identity System shall occur as promptly as practical following the detection of breach and/or unauthorized duplication (e.g., IdP impersonation)

3.6.4 IAM Organizational & Governance Requirements

Table 9: IAM governance requirements.

#	Description
1	An IAM framework (might be part of an overall cyber security framework) shall be developed with a set of rules and policies defining the minimum baseline to comply with to be part of the pan-European Data Factory
2	A central organization shall be in place to control and verify the correct implementation of the IAM framework within the pan-European data factory
3	IAM Administrators shall be appointed to operate IAM components/services and administer user / device accounts

3.7 IDENTIFICATION OF SOLUTIONS

3.7.1 IAM federation protocols

Federated IAM protocols enable the secure sharing of Identity and Access Management (IAM) information across multiple organisations or systems. These protocols establish a trust relationship between participating entities, allowing them to authenticate and authorise users from different domains or organizations without sharing sensitive credentials. The most commonly used protocols are SAML and OpenID Connect.

Specific Assertion Markup Language (SAML)

According to NIST SP 800-63, SAML is an XML-based framework for creating and exchanging authentication and attribute information between trusted entities over the Internet.

SAML Assertions are encoded in an XML scheme and can carry up to three types of statements:

- Authentication statements include information about the assertion issuer, the authenticated user, a validity period and other authentication information;
- Attribute statements contain specific characteristics related to the user. For example, subject “John” is associated with attribute “Role” with value “Administrator”;
- Authorisation statements identify the resources the user has permission to access. These resources may include specific devices, files and information on specific nodes.

OpenID Connect (OIDC)

OpenID Connect is a federated identity and authentication protocol built on top of Oauth 2.0 authorization framework. As part of an OpenID transaction, the Identity Provider (IdP) issues an ID Token , which is a signed assertion. The client (also called service provider or relying party) parses the ID Token to learn about the user and the primary authentication event at the IdP. This Token contains at minimum the following information about the user and authentication event:

- Identification of the IdP that issued the assertion;
- IdP-specific subject identifier representing the user;
- IdP-specific client identifier of the client at the IdP;
- Timestamp at which ID Token expires;
- Timestamp at which ID Token was issued.

3.7.2 Identification of federated IAM solutions

IAM solutions, or Identity and Access Management solutions, provide organizations with a comprehensive framework for managing user identities, authentication, and access control. These solutions streamline the process of granting and revoking access rights, enforcing security policies, and ensuring compliance with regulatory requirements. IAM solutions often include features such as centralised user provisioning, single sign-on (SSO), multi-factor authentication (MFA), and identity lifecycle management.

Microsoft Active Directory

Microsoft AD implements several scenarios and concepts to deal with “external identities” referring to the means to securely interact with users outside an organization:

- B2B collaboration: With this model, external users can authenticate with their preferred Identity Provider and then reach resources of the target organisation. External users are also managed in the same AD directory as internal employees but are annotated as guest users. Guest users can be managed the same way as employees with security groups;
- AD B2C (Business to Customer): With this model a specific and dedicated “AD B2C” directory is created and where the user objects are managed. External users can sign-in against external Identity Providers: SAML-based, Azure AD, social identities, etc.

Keycloak

This is an open-source Identity and Access Management solution which supports OpenID Connect and SAML 2.0 Identity standard protocols, as well Oauth 2.0. It can also connect to existing LDAP or Active Directory servers to retrieve existing user accounts. Keycloak also provides role-based and further fine-grained authorization services to manage access to resources.

4 DATA MANAGEMENT CONCEPT

4.1 INTRODUCTION

Data management refers to the comprehensive process of collecting, storing, protecting, and responsibly sharing data with others. The concept of connected data centers presents a strategic approach for managing and utilizing data that is stored in multiple data centers. This concept entails the establishment of consistent protocols and standards that ensure effective data management across various data centers.

Given that tasks cannot be performed independently by railway suppliers, IMs, and RUs, it is crucial for all stakeholders to collaborate and create a data ecosystem. This concept involves identifying the key components and elements required from a data perspective. The data flow may seem straightforward, but the sheer volume of information involved necessitates utmost caution in securely sharing this information.

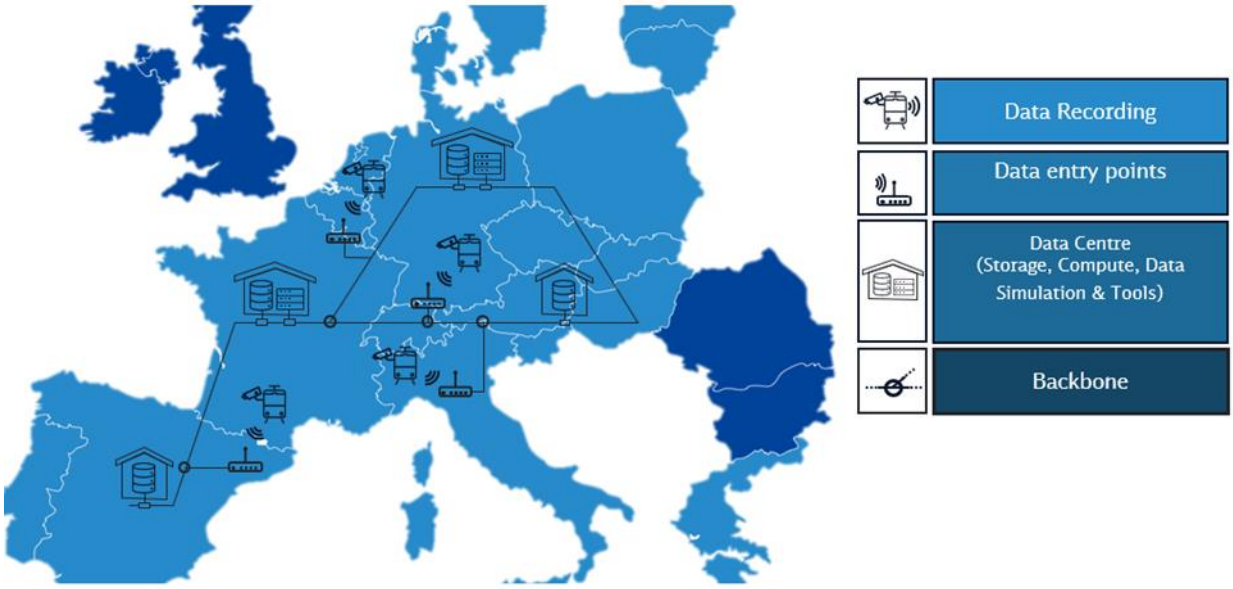

Figure 7: Pan-European data centres.

4.2 DATA FLOW

Data management involves the systematic organisation, storage, and sharing of datasets and data elements to facilitate easy access and retrieval when needed. A well-organised pool of datasets is critical and prerequisite for creating a sustainable rail data ecosystem.

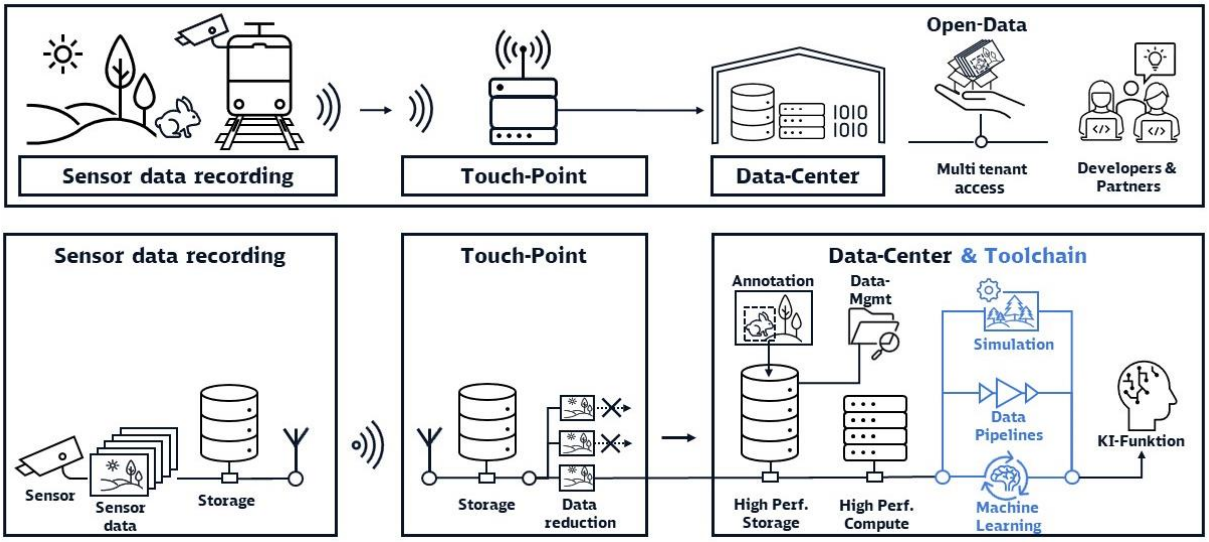

Figure 8: Data organisation and components (legend see Figure 11).

Creating a centralised data repository where all data is stored, all stakeholders can easily access and use the data for their respective needs. There is a segregation of private data (data that is not shared but available for processing / analysis / AI-training by its respective owner) and shareable datasets that can be accessed and used by a larger set of authorised users. When another organisation needs access to private data, permission to access the dataset can be defined and delegated. Proper Data management ensures that data is stored efficiently and made accessible to all stakeholders. This accessibility allows for faster modeling and training, which ultimately improves overall efficiency and productivity. Efficient data storage focuses on optimising resources and minimising overhead. Key considerations include organising data, choosing the right storage infrastructure, compressing and encoding data, utilising indexing and metadata, implementing data deduplication, and employing data archiving and tiered storage. These practices enhance data retrieval speed, reduce storage costs, and improve overall data management capabilities.

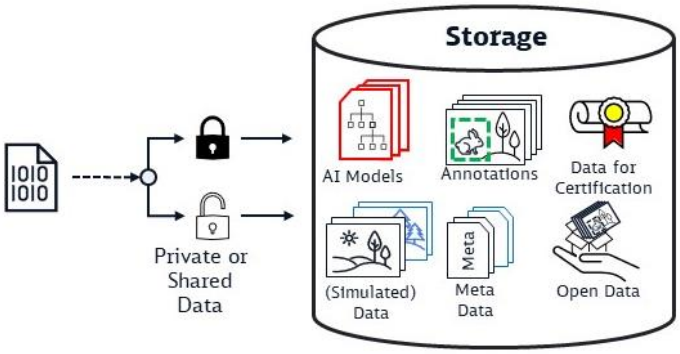

Figure 9: Notion of private and shared data (legend see Figure 11).

In the pan-European data management concept, stakeholders from different countries work collaboratively to collect, store, and manage data in a standardized manner. The data is systematically organised, and protocols are put in place to ensure everyone can easily locate and

utilize the data. This practice facilitates the seamless sharing of local datasets, allowing partners to derive the desired results much faster.

Sharing data is a critical aspect of data management as it promotes collaboration and helps stakeholders achieve more in less time. As more partners start to model and train with the data, there is a corresponding increase in productivity and efficiency, leading to end-to-end data supply chains with secure, sovereign and standardised data exchange. Data management also involves the development and implementation of data policies, data backup and recovery plans, and data archiving systems.

Added Values:

- Achieving sustainability through efficient utilisation of input data while maximising resource utilisation and implementing CO_2-saving standards and practices;
- Ensuring complete traceability of data chains from start to finish in the entire life cycle;
- Establishing consistency and compatibility among various sources and methodologies;
- Enhancing quality management by adopting consistent approaches instead of fragmented solutions;
- Facilitating modular production across multiple suppliers by implementing industry-standard data-driven practices;
- Serving as a foundation for the Digital Rail Twin, which aims to enhance maintenance, service life, and efficiency in rail transportation.

4.3 DATA ARCHITECTURE

Data architecture refers to the way data is structured, stored, and accessed within the referred infrastructure. In the context of the data factory, we have three main components – **Train / IoT devices, Touch-Point, and Data Centre,** and we need to define the data architecture for each of them. We will not delve deep into the physical and network layers but focus on the data layer of the architecture.

1. **Train / Other IoT devices – Data Logging Sub-System:** On the train, data is generated through various sensors on a persistent storage device. This process is facilitated by a **Data Logging Sub-System** that is connected to the local persistent storage device. To store such a massive amount of data, a distributed file system is used that captures data from the data logger and stores it in a machine with hard disks. The specifications of the hard disks depend on the data volume and high-end sensors, cameras, lidar, and other devices that can produce up to more than 1 GB of data per second. After the data has been stored on the hard drive, the required data storage, compression, and encryption methods need to be identified. The future scope is to get a data communication on the train (recording of sensor data) with a high bandwidth (e.g., on rail-certified switch) that is sufficient to offload the recorded data within a limited time frame. An important step on the train is that the data is distributed to data streams. In the past Sensors4Rail project, ROS (Robot Operating System) was used to capture these data streams. Various Steps are completed on the train:
    a. Data Generation: Data generated by Sensors is transferred through Data Logger to Storage.
        i. Time Synchronization of Data Streams: Data is distributed into streams for /further transmission.
    b. Data Distribution: Dividing the data into streams / topics / Blocks in ROS context.

c. Data Integrity Checksum
d. Auto Prioritization
e. Auto Reduction
f. Encryption
g. Compression

2. **Data Touchpoint:** The Touchpoint is an Edge device that has separate compute, storage and network components. It is located near the tracks at locations where the train stops for a period of time, such as e.g., at train stations or yards. The Touchpoint utilises high bandwidth wireless communication to offload the data from the train. Once the data is received a set of processes are started on the device:
    a. De-Compression / Compression: Data is Decompression in the touchpoint before the data can be evaluated. After the evaluation process the data is compressed again for next transmission. This is step is further research as to what the possible solution can be offered in future for this.
    b. Decryption / Encryption: Data security is ensured through encryption / decryption process on all the components.
    c. Auto Prioritization
    d. Auto Reduction
    e. Video Preview generation
    f. Remote Data Selection
    g. Auto (Pre-)Scenario Selection
    h. Offload of Function from the train
3. **Data Center:** The final component in our architecture has a central and important role in this flow. In the data center, the data is saved in a large data storage. In the following, tagging of the recorded sensor data and made manageable, visualisable and searchable by the data management. A system of high-performance computers is connected to the large data storage. This contains a software-based tool chain for the development of AI software. This includes tools for simulation, data pipelines for further processing and qualitative enhancement of the data, and tools for machine learning. This is used to train and test AI software for environment perception. Moreover, the data security, governance and compliance are ensured in all the components. All the measures have been placed to comply with internal and external standards.

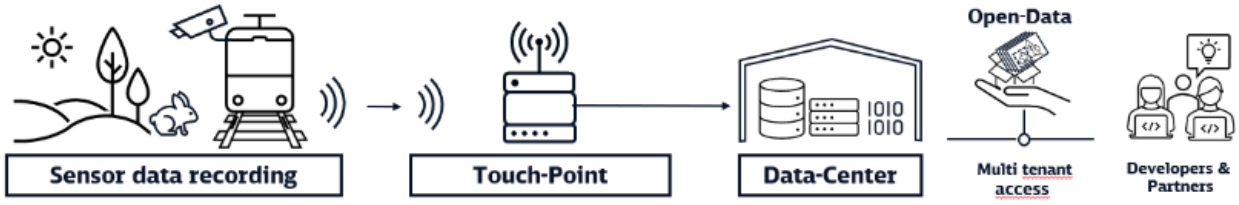

Figure 10: Data processing and transfer chain (legend see Figure 11).

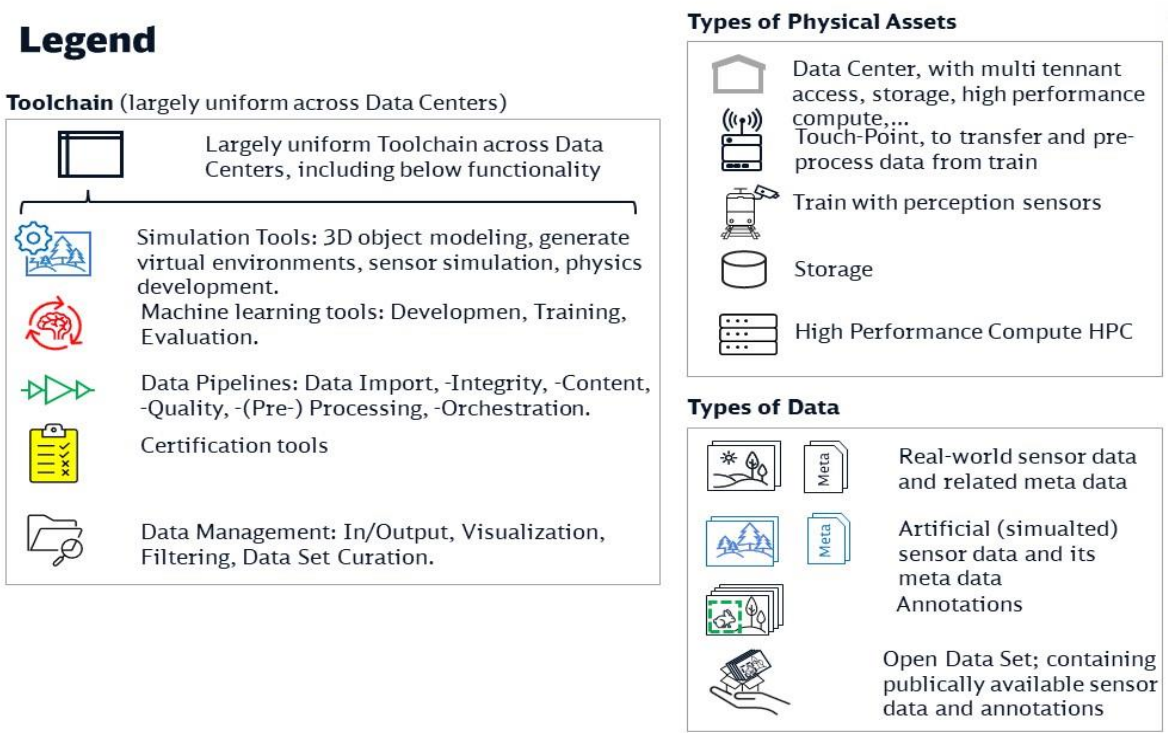

Figure 11: Legend for data classification.

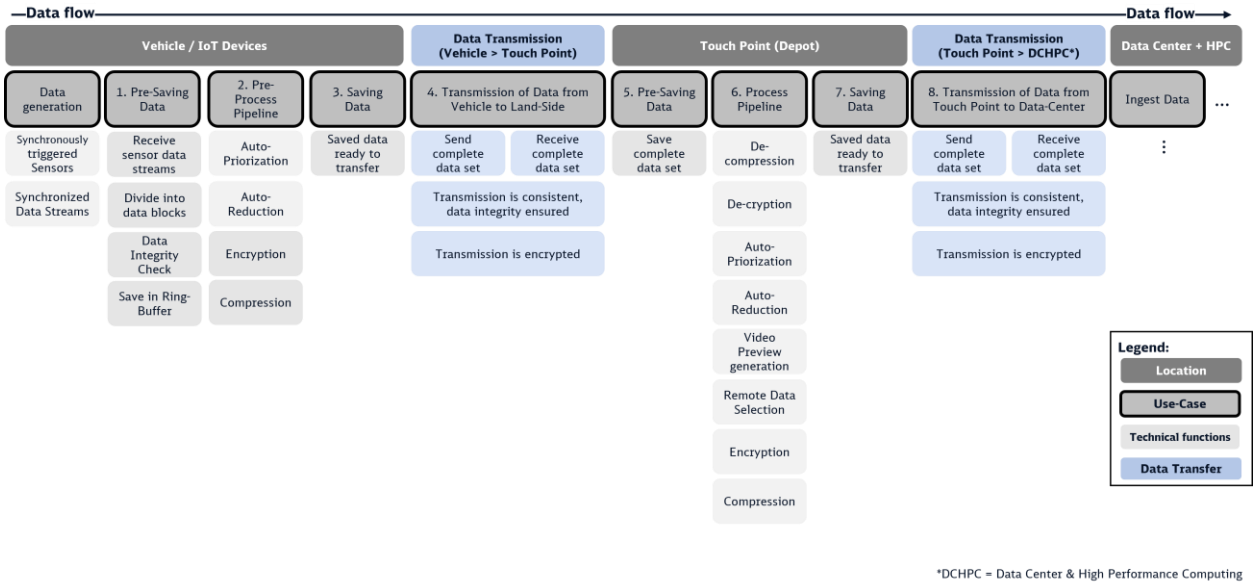

Figure 12: Functional allocation.

4.4 DATA CLASSIFICATION

To establish pan-European data centres, it is crucial to categorise the data. Categorisation helps us understand the responsibility and scope of the data, which is utilised in various use cases, such as annotations, simulation, training, and evaluation. Since the tasks cannot be done independently by railway suppliers, IMs, and RUs, it is essential for all participants in the Pan-European Data Centres to collaborate and create an ecosystem.

There are four primary categories of data that can be shared across the network: **Sensor data**, **Function data**, **Metadata**, and **AI/ML Data**.

1. **Sensor Data:** This data is the measurement data provided by the sensors and other IoT devices on the train and track. There is also synthetic sensor data that is generated by simulation;
2. **Function Data:** This data is generated by the subsystems and used to exchange information between different subsystems;
3. **Metadata:** Metadata describes additional information that is linked to Sensor or Function Data, such as time stamps, sensor parameters and software and hardware versions;
4. **AI/ML Data:** The data in this category describes network architectures, network models, and sensor datasets.

The following subcategories exist for the four primary categories above:

1. **Sensor Data**
    - **Perception Data**
        - Camera Images
        - Infrared Camera Images
        - Lidar Point Clouds
        - Radar Images and Point Clouds
        - Stereo Vision Camera Images
        - Depth Images / Depth Maps
        - Event Camera Images
    - **Localization Data**
        - GNSS
        - IMU/INS
        - Odometry
        - Ground Penetrating Radar
        - MAROS
    - **Map Data**
        - **Digital Register**
          The Digital Register is the central source of railway infrastructure-relevant objects. It provides topological track information, 3D landmark and various platform zones. Landmarks are derived from accurate kinematic measurements and are presented as linear or polygonal shapes. Only objects which exceed a minimum size, are above rails and within a certain distance to tracks are considered. All objects are defined in the Digital Register's object catalogue.
        - **Rail Horizon**
          The Rail Horizon can be regarded as a virtual sensor that is providing an arbitrary foresight based on digital map data and the current train position.
        - **OpenStreetMap (OSM)**
          OpenStreetMap is an open-source map data project that provides geographic information.
        - **Geographic Information System (GIS)**
          Geographic Information System is a system designed to create, manage and visualise geographic data.
2. **Function Data**

- Bus data
- Subsystem outputs (e.g., detected objects, assessed scenes, …)

3. **Metadata**
    - **Descriptive Metadata**
        - Title or name of the data
        - Owner
        - Date of creation
        - Keywords or tags
        - Summary or abstract
        - Language
    - **Structural Metadata**
        - File format or data format
        - File size
        - Data organization or structure
        - Relationship to other data or files
        - Data version or revision information
    - **Administrative Metadata**
        - Data ownership or rights information
        - Access permissions and restrictions
        - Data provenance or source
        - Data creation or modification history
        - Data retention or expiration policies
    - **Technical Metadata**
        - File type or format specifications
        - Encoding or compression methods
        - Data resolution or quality information
        - Software or tools used for data creation or processing
        - Hardware specifications or requirements
    - **Contextual Metadata**
        - Data subject or topic
        - Geographic location or coordinates
        - Temporal information (time and date)
        - Related events or context
        - Cultural or historical relevance
    - **Preservation Metadata**
        - Data preservation methods or strategies
        - File integrity checks or checksums
        - Migration or conversion information
        - Backup or recovery procedures
        - Long-term storage requirements
    - **Rights Metadata**
        - Copyright or intellectual property information
        - License information (e.g., Creative Commons)
        - Terms of use or distribution
        - Attribution requirements
        - Privacy or confidentiality considerations
        - Approvals
4. **AI/ML Data**

- **Neural Network Architectures**
  Neural network architectures describe the design (i.e., layers and functions) of a neural network.
- **Neural Network Models**
  A network model is the result when training a neural network architecture on a (training) dataset.
- **Datasets**
  Datasets are collections of selected data samples for a specific use case. An of open multi-modal sensor dataset in the railway context is OSDaR23. It contains 45 annotated sequences with synchronised visual camera, infrared camera, lidar, and radar data in combination with precise GNSS and IMU information.

4.5 DATA FORMATS

Pan-European data centers adhere to industry-standard data formats that facilitate efficient storage, retrieval, and processing of data. Here are some common data formats that can be:

1. **Sensor Data**
   - **Perception Data**
     - **Images**
       Cameras capture the environment as images. But there are also radar sensors that output their data as images. Depending on camera and radar type these images differ in the representation and have to be interpreted respectively. Images can differ in the number of channels, the resolution as well as the dynamic range. The most common data formats to store images are JPEG (Joint Photographic Experts Group) and PNG (Portable Network Graphics). Common formats might often not be sufficient to store all the available information captured by the sensor. Therefore, it might be necessary to store these images in raw image formats. There are open formats existing, such as DNG, that can be used. For some sensors it makes sense to use customised formats which can be created with technologies such as Protocol Buffers.
     - **Point Clouds**
       Sensors such as LiDAR (Light Detection and Ranging) or Radar generate point cloud data, which represents the 3D coordinates of objects in the physical environment. There are multiple open formats available to store point cloud data such as PCD (Point Cloud Data) or LAS (LiDAR Data Exchange Format).
   - **Localization Data**
     - **GNSS and IMU/INSS**
       Contains data about the position and orientation in world coordinates and the acceleration and velocity of the train. These floating-point numbers with associated keys are commonly stored in files such as JSON or CSV as well as databases.
     - **Ground Penetrating Radar and MAROS**
       The idea behind these technologies for localization is to create a unique fingerprint of the current area and store it in a map. This allows to later localise the train based on matching fingerprints. The data format depends

on the used sensor and there are also novel technologies on the market, such as the MAROS (Magnetic Railway Onboard Sensor).

- **Map Data**
  - **Digital Register**
    The Digital Register is currently under development and uses Protocol Buffers for storing the map data. In addition to this, GeoJSON and JSON are currently used formats.
  - **Rail Horizon**
    The Rail Horizon is an onboard system and utilizes Protocol Buffers to generate the virtual horizon from the existing map data.
  - **OpenStreetMap (OSM)**
    Maps can be stored in OSM XML or PBF (Protocol Buffer Binary Format) formats.
  - **Geographic Information System (GIS)**
    Various GIS formats, such as Shapefile (SHP) or GeoJSON, can be used to store map data including road networks, landmarks, and points of interest.

2. **Function Data**
   - Depends on the specific applications. This can contain classical onboard BUS data as well as function output such as object lists. These can be stored in JSON files as well as customized Protocol Buffer formats.
3. **Metadata**
   - **General**
     Metadata mainly consists of hierarchical key-value pairs. This data can be stored in files as well as in databases. JSON is commonly used when storing this data as files. There are many SQL (Structured Query Language) and NoSQL databases available that are suitable for storing metadata.
   - **Annotations**
     Annotations are a special type of metadata that is used for machine learning and to validate the output of algorithms. Annotation or Labelling describes the process of marking objects in camera, lidar or radar data by either marking the points or pixels that belong to a specific object or by enclosing the objects with bounding boxes or polygons in the respective sensor data. In addition, attributes can be assigned for each label that further describe the labelled object. Annotations are commonly stored in JSON files. ASAM is working on a standardisation of the JSON structure for annotations in the automotive area with the ASAM OpenLabel standard. DB Netz AG is supporting this effort and has adapted this standard for rail [8].
4. **AI/ML Data**
   The data formats for Machine Learning applications depend on the specific use case and the used framework (e.g., TensorFlow, PyTorch). With the Open Neural Network Exchange (ONNX) there is an open format that allows sharing network models between different frameworks.

This topic is open to discussion as a unified interchangeable format can be defined once the concept is implemented and agreed upon.

An essential aspect to consider for this metadata directory is the comprehensive functional depiction of the operational context surrounding the data. In simpler terms, it is crucial to include a readme or a wiki that precisely outlines the procedures involved in utilising the dataset. For instance, this documentation should encompass a thorough explanation of the dataset's functional components

and the requisite AI program, which are indispensable for successfully conducting time-sensitive tests.

4.6 DATA QUALITY

The data quality rules have been defined based on the data category and each data category has a particular quality to uphold. We are only drafting the high-level data quality rules and this document does not fulfil the scope of doing into the details of all the data quality rules. As far as the collaborative datasets is considered, we must maintain a level of data quality on our dataset otherwise no further development could benefit.

Sensor Data Quality Rules

Here are some general and key aspects of sensor data quality rules:

1. Accuracy: Sensor data should accurately reflect the measured physical phenomena or variables. The sensors should be calibrated and validated regularly to ensure accurate measurements within specified tolerances.
2. Precision: Sensor data should provide precise measurements with minimal variability and uncertainty. This requires using sensors with appropriate resolution and sensitivity, as well as employing techniques to reduce noise and interference in the data.
3. Consistency: Sensor data should exhibit consistency over time and across different sensors of the same type. Consistency can be ensured by standardizing sensor calibration procedures, data acquisition techniques, and environmental conditions during data collection.
4. Completeness: Sensor data should be complete, capturing all relevant measurements and associated metadata. Missing or incomplete data points should be minimized, and techniques such as interpolation or data imputation may be employed when necessary.
5. Timeliness: Sensor data should be collected and made available in a timely manner to enable real-time or near real-time applications. Delays in data acquisition and transmission should be minimized to ensure timely decision-making.
6. Validity: Sensor data should be validated to ensure that it represents the intended physical phenomena or variables. This may involve cross-referencing sensor measurements with ground truth data or comparing readings from multiple sensors to detect anomalies or inconsistencies.
7. Reliability: Sensor data should be reliable and trustworthy. Sensors should undergo regular maintenance and quality assurance procedures to prevent malfunctions or drifts in measurements. Redundancy measures, such as using multiple sensors for verification, can enhance reliability.
8. Metadata: Sensor data should be accompanied by comprehensive metadata, including information about sensor specifications, calibration history, sampling rates, and any relevant contextual information. This metadata helps in interpreting and analysing the sensor data accurately.
9. Data Quality Assurance: Regular data quality checks should be performed, including outlier detection, data cleansing, and statistical analysis. Anomalies, errors, or inconsistencies should be identified and addressed promptly to maintain data integrity.
10. Documentation: Detailed documentation of sensor deployment, calibration processes, data collection protocols, and any modifications or maintenance performed on the sensors should

be maintained. This documentation aids in traceability, replication, and analysis of the sensor data.

Key Aspects of certain sensor types

- Camera: Image resolution, focal distance, focal point.
- Lidar: point density, range, radial resolution, horizontal resolution, measurement accuracy.
- Radar: objects, their distance, speed, and direction.
- Multi-modal data: calibration precision, synchronicity (max time offset of the acquisition time stamps), accuracy of time synchronisation within the system (e.g., min accuracy, max accuracy, average error, ...)
- time series data (so video data / data streams): frequency, max level of frame drops (frames getting lost), variance in frequency (e.g., 10Hz means one data point every 100ms - in reality this can be maybe between 95ms and 105ms)
- and combination of multiple, e.g., multi-modal time series data

Annotation quality

- Unique Identifier for each Annotation
- max allowed error level
- pixel accuracy for boxes
- lidar accuracy
- Attribute Quality
- Annotation Type Quality
- Annotation Rule Quality
- Data Quality based in Classes.

Here are some general and key aspects of annotation data quality rules:

1. Accuracy: Annotations should accurately represent the intended labels or characteristics of the data. Annotators should possess sufficient expertise and follow rigorous guidelines to minimize errors and inaccuracies.
2. Consistency: Annotations should be consistent across different annotators and annotation tasks. Inter-annotator agreement measures, such as kappa scores, can be used to assess consistency and guide improvements in annotation guidelines and training.
3. Completeness: Annotations should cover all relevant aspects specified in the annotation guidelines. No crucial information or features should be omitted or overlooked during the annotation process.
4. Granularity: Annotations should be appropriately granular, depending on the specific requirements of the task. Fine-grained annotations provide more detailed and specific information, while coarse-grained annotations offer broader categorizations. The granularity should align with the objectives of the AI application.
5. Objectivity: Annotations should be objective and unbiased, avoiding subjective interpretations or personal biases. Annotators should follow standardized guidelines and avoid introducing their own subjective judgments.
6. Contextual Understanding: Annotators should possess a good understanding of the context and domain-specific knowledge relevant to the data being annotated. This enables accurate labelling and avoids misinterpretation.

7. Regular Quality Control: Regular quality control checks should be conducted to identify and rectify any annotation errors or inconsistencies. These checks can involve expert review, validation by multiple annotators, or using pre-annotated gold-standard data for comparison.
8. Documentation: Detailed documentation of the annotation process, guidelines, and any specific conventions used should be maintained. This helps maintain transparency and allows others to replicate or review the annotation work.
9. Feedback and Iterative Improvement: Annotators should receive feedback on their work, including discussions on ambiguous cases and areas for improvement. Iterative cycles of annotation and review can help refine the annotation guidelines and enhance data quality.

Metadata quality

Rules are designed to ensure that the metadata associated with a dataset is of high quality, accurate, and comprehensive.

Metadata quality rules:

1. Accuracy: The metadata should provide accurate and reliable information about the AI model, including its purpose, functionality, and limitations. It should reflect the actual behaviour and performance of the model.
2. Completeness: The metadata should be comprehensive and include all relevant information about the AI model. This includes details about the model's architecture, training data, hyperparameters, and any other essential elements that contribute to its functioning.
3. Consistency: The metadata should be consistent in its terminology, formatting, and structure. It should adhere to predefined standards and conventions to ensure ease of understanding and comparison with other models.
4. Clarity: The metadata should be clear and easily understandable to different stakeholders, including researchers, developers, and end-users. It should avoid jargon and technical language whenever possible and provide clear explanations of key concepts and terms.
5. Relevance: The metadata should focus on providing information that is directly relevant to the AI model and its application. It should prioritise important details while avoiding unnecessary or redundant information.
6. Up to date: The metadata should be regularly updated to reflect any changes or updates to the AI model. This includes modifications to the model's architecture, training data, or any other relevant aspects that may impact its performance or usage.
7. Accessibility: The metadata should be easily accessible to users and should be available in a format that facilitates easy retrieval and understanding. It should be well-organized and accompanied by appropriate documentation or references for further clarification.
8. Metadata should be available.
9. Metadata should be readable and known formats.
10. All the metadata formats should be based on FAIR principles [9]

4.7 METADATA MANAGEMENT

Metadata management involves the organisation, documentation, and control of metadata associated with the train's operations, components, and systems. It plays a crucial role in ensuring the reliability, traceability, and interoperability of data used by autonomous train systems. Here are some key aspects of metadata management for autonomous trains:

Metadata Identification and Definition:

Identify the types of metadata relevant to autonomous train operations, such as train configuration, sensor data, maintenance records, operational parameters, and safety guidelines.

Define the specific metadata elements, their formats, and their relationships to ensure consistent and standardised metadata across the train system.

Metadata Collection and Storage:

Establish mechanisms to collect and store metadata from various sources within the autonomous train system, including sensors, control systems, maintenance logs, and external interfaces.

Implement appropriate data storage and management systems, such as databases or data lakes, to securely store and retrieve metadata.

Metadata Documentation and Cataloging:

Create a metadata catalog or repository that documents the available metadata, including its meaning, source, format, and any associated business rules or constraints.

Maintain a comprehensive metadata dictionary or data catalog to provide a centralized reference for metadata definitions, attributes, and relationships.

Metadata Governance and Quality Control:

Implement metadata governance processes to ensure metadata integrity, consistency, and compliance with regulations, standards, and internal policies.

Establish quality control mechanisms to validate and verify the accuracy, completeness, and reliability of metadata, especially for critical functions like safety-critical systems.

Metadata Interoperability and Integration:

Define data exchange standards and protocols to facilitate interoperability between different autonomous train systems, subsystems, or external interfaces.

Establish metadata mapping and transformation mechanisms to enable seamless integration and data sharing between different metadata sources and formats.

Metadata Lifecycle Management:

Define metadata lifecycle processes, including metadata creation, modification, archiving, and deletion.

Implement version control and change management practices to track and manage changes to metadata over time, ensuring proper documentation and audit trails.

Metadata Security and Access Control:

Implement appropriate access controls and security measures to protect sensitive metadata from unauthorized access, modification, or disclosure.

Define access privileges and roles to restrict metadata access based on user roles, responsibilities, and the principle of least privilege.

Metadata Discovery and Search:

Provide tools or interfaces for users to discover and search for relevant metadata based on specific criteria or attributes.

Implement metadata search capabilities, including metadata indexing, tagging, or search algorithms, to facilitate efficient and effective discovery of metadata.

Figure 13: Metadata management.

4.8 DATA GOVERNANCE

4.8.1 Introduction

The purpose is to establish a comprehensive data governance framework for the pan-European Data Factory. The framework aims to ensure effective data management, quality, and security while promoting interoperability and standardisation. It outlines the key principles, policies, and guidelines for governing data in the automated train ecosystem.

4.8.2 Objectives

The primary objectives of the data governance framework are as follows:

a) Facilitate seamless data exchange and interoperability among Pan-European Data Centres.
b) Ensure data privacy, security, and compliance with relevant regulations and standards.
c) Enhance data quality, accuracy, and reliability for optimal train operations and passenger safety.
d) Establish roles, responsibilities, and accountability throughout the data lifecycle.
e) Enable data-driven decision-making, optimisation, and innovation in the Pan-European Data Centres.

4.8.3 Approach

After careful examination of various widely accepted and frequently referenced data governance frameworks the following approaches have been created. For reference, please refer to McKinsey [10], Eckerson [11], PwC [12], SAS [13], DGI [14], Gaia-X [15] and DAMA DMBOK [16].

When comparing different data governance operating models, there are several factors to consider. Here are some key points of comparison:

a) Centralised vs. Decentralised: Data governance operating models can be centralised or decentralised. In a centralised model, there is a single governing body responsible for data governance across the organisation. This promotes consistency and alignment but may face challenges in accommodating diverse needs. In a decentralised model, data governance is distributed across various units or departments, allowing for more tailored governance practices but potentially leading to inconsistencies and duplication of efforts.
b) Top-Down vs. Bottom-Up: The approach to data governance can be either top-down or bottom-up. In a top-down model, governance is driven by senior leadership, who define policies, standards, and guidelines that are then implemented throughout the organisation. This ensures a strong focus on governance principles but may result in limited flexibility and engagement at lower levels. In a bottom-up model, governance initiatives start at the operational level, with units taking ownership of their data governance practices. This allows for more agility and local context but may lead to fragmentation and lack of overall coordination.
c) Process-Oriented vs. Data-Oriented: Data governance operating models can also vary in their focus. Process-oriented models prioritise establishing clear processes, workflows, and accountability mechanisms for data governance. This ensures consistency and efficiency but may overlook the specific characteristics and quality of the data itself. Data-oriented models, on the other hand, emphasise understanding and managing the data itself, including data quality, metadata, and data lineage. This provides a more granular and holistic approach to governance but may require additional effort to define and maintain processes.
d) Maturity and Complexity: Data governance operating models can range in maturity and complexity. Some organisations may have relatively simple models with basic governance structures and processes, while others may have sophisticated models with multiple tiers of governance, specialised committees, and advanced data management capabilities. The choice of operating model should align with the organization's current maturity level and its ability to handle the complexity of data governance requirements.
e) Ultimately, the choice of a data governance operating model depends on the organisation's specific needs, culture, resources, and strategic objectives. It is important to carefully assess these factors to select an operating model that best fits the organisation's unique circumstances.

Two possible approaches to governing the data ecosystem: the Unified Central Data Space(UCDS) and the Decentralized Data Space (DDS). This approach has been taken from the Gaia X data space framework [15].

1. **Unified Central Data Space - UCDS**
2. **Decentralised Data Space - DDS**

In the **Unified Central Data Space (UCDS)**, data producers, data owners, and data consumers share a centralised data space for data sharing. Data producers or owners publish the data within a centralised network designated for data sharing. It is the responsibility of the data producer or owner to maintain the data and metadata quality and standards. Once the dataset and metadata are published, they should be regularly maintained and revised according to predefined norms. Two types of data exist: private and shareable. Private data is governed by specific rules, including time-based access and non-shareable agreements, while shareable data can only be shared within the consortium.

In the **Decentralized Data Space (DDS)**, data producers or owners host the data in their private data spaces, while metadata is shared across the data ecosystem. When data consumers find the appropriate dataset, they submit a request to access the data, and the data producer or owner responds with an approval or denial. In this model, governance is distributed with 80% responsibility lying with the data producer or owner and 20% with the data consumer. This distribution implies that the primary responsibility for governance, including decision-making, oversight, and control, lies with the data producer or owner. The data producer or owner is the entity or individual who generates or possesses the data in question. They hold the majority share of responsibility, accounting for 80% of the governance process.

The data producer or owner is expected to take the lead in establishing and enforcing governance policies and practices related to the data. They have the authority to determine how the data is collected, stored, processed, and shared. They are responsible for ensuring compliance with legal and ethical guidelines, as well as protecting the data from unauthorised access or misuse.

On the other hand, the data consumer is assigned 20% of the governance responsibility. The data consumer refers to the party or parties who utilise or access the data produced by the data owner. They have a smaller role in the governance process compared to the data producer/owner.

As a data consumer, their responsibility may include adhering to the governance policies set by the data producer/owner, respecting data privacy and security measures, and using the data in accordance with any agreed-upon terms or restrictions. They may also provide feedback or suggestions to the data producer / owner regarding the data's quality, usability, or specific requirements.

By assigning a majority share of governance responsibility to the data producer/owner, this model emphasizes their role as the primary custodian of the data. This arrangement acknowledges that the data producer / owner has the most intimate knowledge of the data, its context, and its intended purpose, and thus, should have a higher level of control and decision-making authority. Meanwhile, the data consumer's role is recognised as important but secondary to the data producer / owner in the overall governance framework.

4.8.4 Data Governance Framework

Governance Structure The data governance structure for Pan-European Data Centres might consist of the following entities and their roles:

a. **European Data Governance Council (EDGC)**: The EDGC will serve as the central governing body responsible for overseeing data governance policies, standards, and initiatives across Europe. It will establish guidelines, promote collaboration, and facilitate coordination among stakeholders.
b. **National Data Governance Authorities (NDGA):** Each participating country will have a designated National Data Governance Authority responsible for implementing and enforcing the data governance framework within their jurisdiction. They will ensure compliance, address local requirements, and act as points of contact for data-related matters.
c. **Train Operators**: Train operators will be responsible for data collection, management, and sharing within their respective operations. They will adhere to the data governance policies and standards set forth by the EDGC and the National Data Governance Authorities.
d. **Data Owners and Data Producers:** Entities responsible for generating and owning data, such as infrastructure managers, and service providers, will comply with the data governance framework. They will ensure data quality, accuracy, and timeliness while adhering to data sharing protocols.
e. **Data Consumers**: Authorised entities, including research institutions, government agencies, and analytics providers, may access and analyse the data for research, optimization, and decision-making purposes. Data consumers will adhere to data access policies and protocols established by the framework.

Data Governance Policies and Standards The data governance framework will include the following policies and standards:

a. Data Classification and Metadata: Guidelines for classifying data based on sensitivity, criticality, and usage. Metadata standards will be established to ensure consistency and facilitate data discovery.
b. Data Access and Security: Protocols for granting authorised access to data, including authentication, authorisation, and encryption mechanisms. Security measures will be implemented to safeguard data from unauthorised access, breaches, or misuse.
c. Data Privacy and Compliance: Policies to protect passenger privacy and ensure compliance with relevant data protection regulations, such as the General Data Protection Regulation (GDPR) in Europe.
d. Data Quality and Integrity: Standards and processes to maintain data quality, accuracy, completeness, and integrity throughout its lifecycle. Data validation, cleansing, and verification mechanisms will be implemented.
e. Data Retention and Archiving: Guidelines for data retention periods, archival practices, and data disposal in compliance with legal and regulatory requirements.
f. Data Sharing and Interoperability: Protocols and formats for seamless data sharing and interoperability among automated train systems across Europe. Standards for data exchange interfaces and communication protocols will be established.

Roles and Responsibilities The data governance framework will define the roles and responsibilities of stakeholders involved in the automated train ecosystem:

a. European Data Governance Council (EDGC): Oversee and set policies, standards, and guidelines at the European level. Promote collaboration and harmonization among stakeholders.
b. National Data Governance Authorities: Implement and enforce the data governance framework within their respective jurisdictions. Ensure compliance, provide guidance, and resolve data-related issues.
c. Train Operators: Collect, manage, and share data in compliance with the framework. Ensure data quality, security, and adherence to data governance policies
d. Data Owners and Data Producers: Generate and own data, ensuring compliance with data governance policies. Maintain data quality, accuracy, and timelines
e. Data Consumers: Access and analyse data for authorised purposes. Adhere to data access policies and protocols defined by the framework.

4.8.4.1 Conclusion

The data governance framework for pan-European Data Factory establishes a robust structure, policies, and standards to ensure effective data management, quality, and security. By adhering to this framework, stakeholders can promote interoperability, innovation, and safety in the automated train industry. Continuous collaboration, monitoring, and evolution of the framework will be essential to adapt to technological advancements and changing requirements in the future.

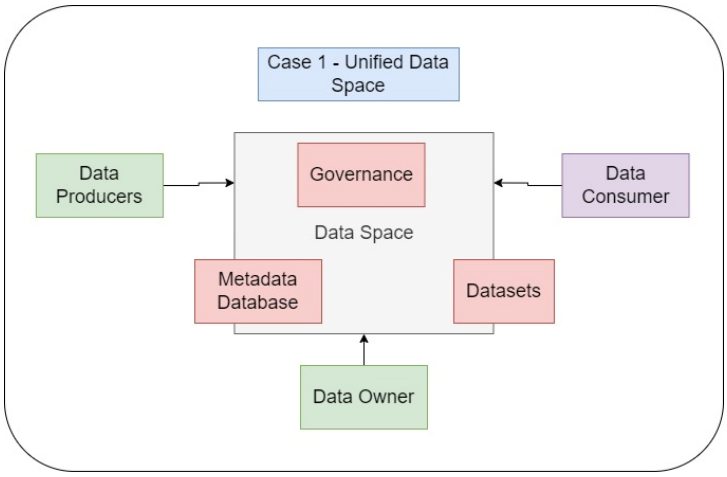

Figure 14: Unified data space.

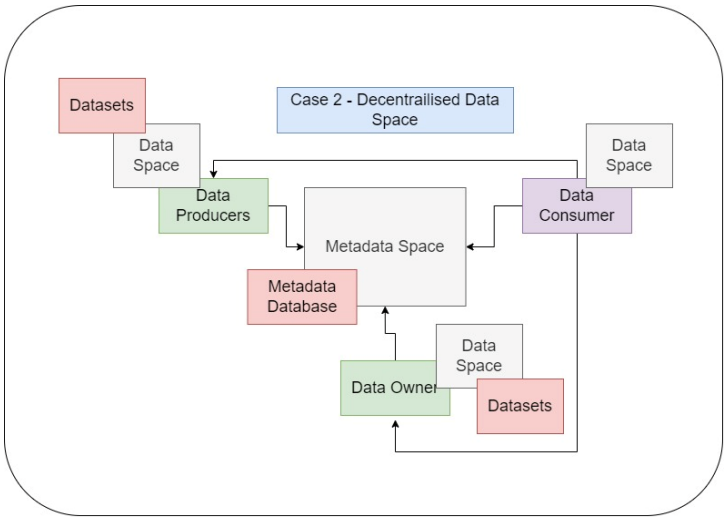

Figure 15: Decentralised data space.

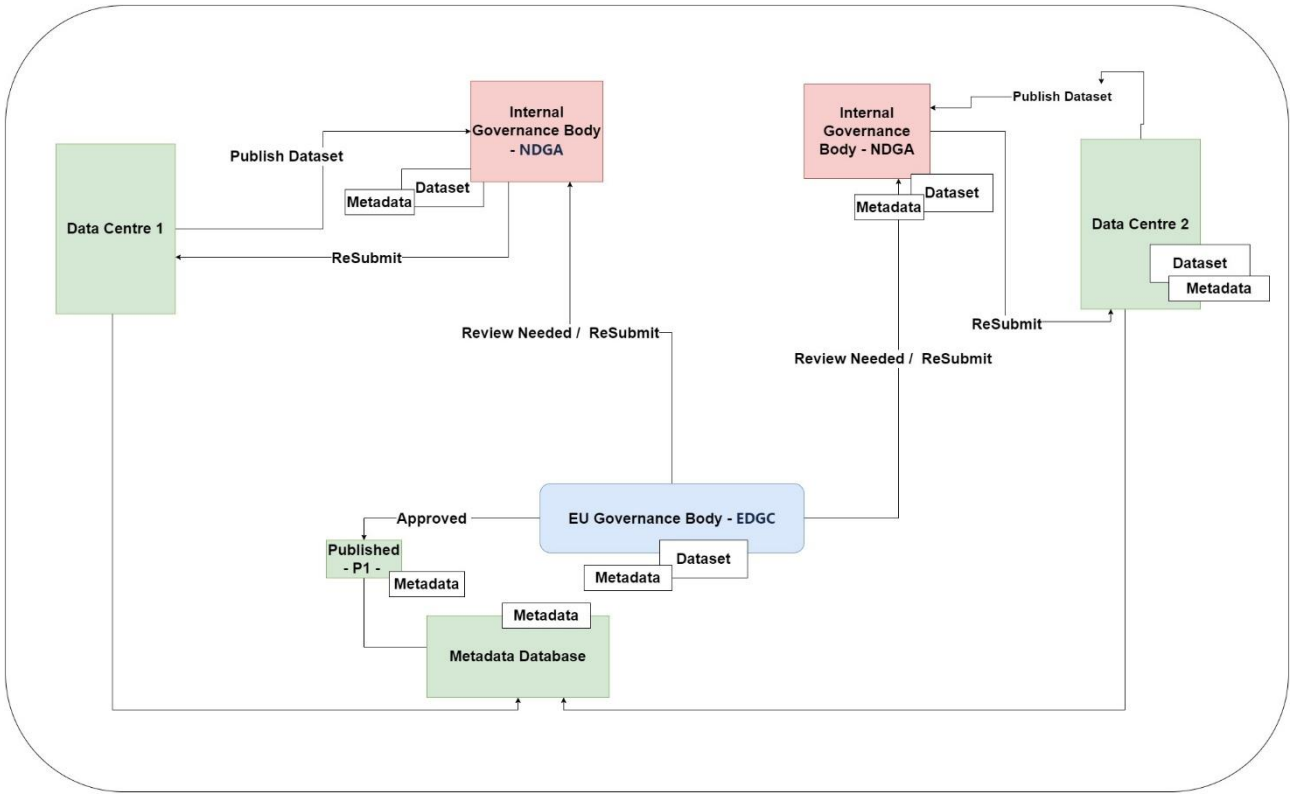

Figure 16: Data governance framework & process.

4.9 DATA SECURITY

The primary purpose of this chapter is to identify European data governance and security frameworks incl. security requirements ruling data security aspects. The second purpose of this chapter is to describe a security scheme to fulfil the identified security requirements. The third purpose of this chapter is to provide a high-level threat analysis incl. attack method, security vulnerabilities and mitigation measures for the Pan-European Data Centres.

European Regulations related to data security:

- European Data Governance ACT (DGA) (EU) 2022/868
- NIS 2 Directive (EU) 2022/2555
- General Data Protection Regulation (GDPR) (EU) 2016/679 (only applicable if the processing of personal data occurs)

Extract of Security Requirements

Table 10: Data security requirements.

Source	Article	Requirement description
DGA	Article 5, Conditions for re-use	Public sector bodies shall, in accordance with Union and national law, ensure that the protected nature of data is preserved. They may provide for the following requirements: a) to grant access for the re-use of data only where the public sector body or the competent body, following the request for re-use, has ensured that data has been: • anonymised, in the case of personal data; and • modified, aggregated or treated by any other method of disclosure control, in the case of commercially confidential information, including trade secrets or content protected by intellectual property rights b) to access and re-use the data remotely within a secure processing environment that is provided or controlled by the public sector body c) to access and re-use the data within the physical premises in which the secure processing environment is located in accordance with high security standards, provided that remote access cannot be allowed without jeopardising the rights and interests of third parties
		In the case of re-use allowed in accordance with paragraph 3, points (b) and (c), the public sector bodies shall impose conditions that preserve the integrity of the functioning of the technical systems of the secure processing environment used.
		Unless national law provides for specific safeguards on applicable confidentiality obligations relating to the re-use of data referred to in Article 3(1), the public sector body shall make the re-use of data provided in accordance with paragraph 3 of this

		Article conditional on the adherence by the re-user to a confidentiality obligation that prohibits the disclosure of any information that jeopardises the rights and interests of third parties that the re-user may have acquired despite the safeguards put in place. Re-users shall be prohibited from re-identifying any data subject to whom the data relates and shall take technical and operational measures to prevent re-identification and to notify any data breach resulting in the re-identification of the data subjects concerned to the public sector body. In the event of the unauthorised re-use of non-personal data, the re-user shall, without delay, where appropriate with the assistance of the public sector body, inform the legal persons whose rights and interests may be affected.
	Article 12, Conditions for providing data intermediation services	The provision of data intermediation services referred in Article 10 shall be subject to the following conditions: (j) the data intermediation services provider shall put in place adequate technical, legal and organisational measures in order to prevent the transfer of or access to non-personal data that is unlawful under Union law or the national law of the relevant Member State; (k) the data intermediation services provider shall without delay inform data holders in the event of an unauthorised transfer, access or use of the non-personal data that it has shared (l) the data intermediation services provider shall take necessary measures to ensure an appropriate level of security for the storage, processing and transmission of non-personal data, and the data intermediation services provider shall further ensure the highest level of security for the storage and transmission of competitively sensitive information; (o) the data intermediation services provider shall maintain a log record of the data intermediation activity.
GDPR	Article 25 - Data Protection by design and by default	1. Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as

		the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects. 2. The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed.That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility.In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons. 3. An approved certification mechanism pursuant to Article 42 may be used as an element to demonstrate compliance with the requirements set out in paragraphs 1 and 2 of this Article.
	Article 30 - Records of processing activities	Each controller and, where applicable, the controller's representative, shall maintain a record of processing activities under its responsibility.
	Article 32 – Security of Processing	In assessing the appropriate level of security account shall be taken in particular of the risks that are presented by processing, in particular from accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to

		personal data transmitted, stored or otherwise processed.
	Article 33 - Notification of a personal data breach to the supervisory authority	In the case of a personal data breach, the controller shall without undue delay and, where feasible, not later than 72 hours after having become aware of it, notify the personal data breach to the supervisory authority competent in accordance with Article 55, unless the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons
NIS 2	Article 21 - Governance	Member States shall ensure that the management bodies of essential and important entities approve the cybersecurity risk-management measures taken by those entities in order to comply with Article 21, oversee its implementation and can be held liable for infringements by the entities of that Article.
	Article 21 – Cybersecurity risk-management measures	Member States shall ensure that essential and important entities take appropriate and proportionate technical, operational and organisational measures to manage the risks posed to the security of network and information systems which those entities use for their operations or for the provision of their services, and to prevent or minimise the impact of incidents on recipients of their services and on other services. Taking into account the state-of-the-art and, where applicable, relevant European and international standards, as well as the cost of implementation, the measures referred to in the first subparagraph shall ensure a level of security of network and information systems appropriate to the risks posed. When assessing the proportionality of those measures, due account shall be taken of the degree of the entity's exposure to risks, the entity's size and the likelihood of occurrence of incidents and their

		severity, including their societal and economic impact.

Summary of identified requirements from the EU Data security frameworks

- Implementation and enforcement of appropriate and proportionate state of the art security controls (technical, organisational and legal) through “security-by-design”
- A risk management approach shall be defined and enforced to determine appropriate security controls
- In the event of a data breach related to personal data (PII), the competent authority shall be informed in a timely manner by the data controller (within 72 hours)
- A documentation and record of personal data processing activities shall be maintained

Data Security Scheme Data must be protected against data breach and the confidentiality, integrity and availability of the data shall be maintained. Data Security can be structured around 4 pillars:

- **Technical Measures**: It is related to the equipment, components, devices, and associated documentation or other media which pertain to cryptography, or to the security of telecommunications and information systems;
- **Organisational Measures**: It is related to a program designed by an organization to maintain the cyber security of the entire organisation’s assets to an established level of confidentiality, integrity and availability;
- **Operational Measures**: it is related to the security controls (i.e., safeguards or countermeasures) for an information system that are primarily implemented and executed by people (as opposed to systems);
- **Legal Measures**: It is related to contractual clauses between parties to ensure appropriate data protection safeguards.

Figure 17 describes the structure for the overall management of data security:

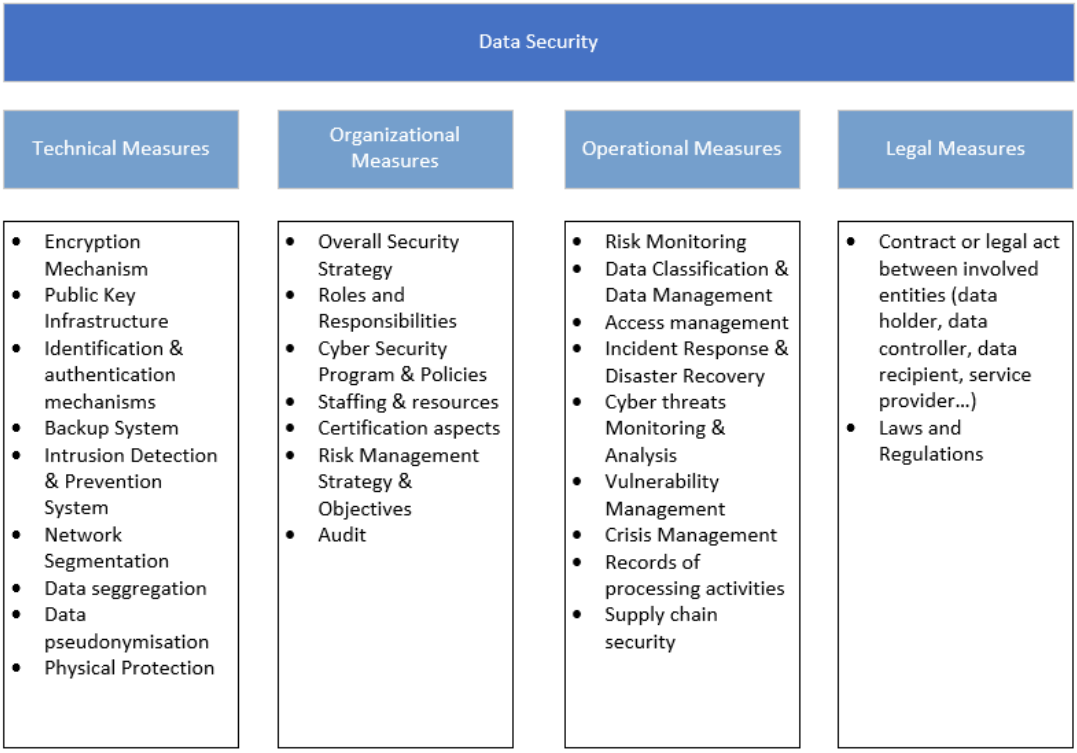

Figure 17: Data security scheme.

Threat Analysis

This part describes the baseline for threats, vulnerabilities and attack methods on the Pan-European Data Centres and can be considered for detailed Threat and risk analysis. It also describes mitigations to the identified threats.

Possible attack impacts on data may include:

- Data integrity breach;
- Data confidentiality breach;
- Loss of data availability.

Table 11: Data threat analysis.

High level and sub-level descriptions of threats		**Attack method / exploited vulnerabilities**	
Threats regarding back-end servers: infrastructure and applications of	Back-end servers used as a means to extract or compromised data (“data breach”)	1.1	Abuse of privileges by staff **(insider attack)**
		1.2	**Unauthorized remote access** to the server (enabled for example by backdoors, unpatched Remote Access Server, user identity theft…)

the pan-EU Data Factory		1.3	**Unauthorized physical access** to the server (conducted by connecting a media or device to the server)
		1.4	**Information breach** by unintended sharing of data (e.g. admin errors)
	Services from back-end server such as tool chain being disrupted, affecting global operation	2.1	**Attack on back-end server stops its functioning (e.g. DDOS attack)** for example it prevents it from interacting with data touchpoint / train and providing services the rely on.
Threats to the Pan-EU Data Factory communication channels e.g. inter Data Factory communication	Spoofing of messages / data being transferred inside the pan-European data factory network	3.1	**Spoofing of messages** by impersonation
	Communication channels used to conduct unnoticed changes (manipulation) or deletion of data	4.1	Communication channels permit **manipulation** of data
		4.2	Communication channels permit **erasure** of data
		4.3	Communication channels permit **introduction of code / code injection**
	Communication channels permit untrusted / unreliable messages to be accepted	4.4	Accepting information from an **unreliable or untrusted source** (Man in the middle attack / session hijacking or impersonation)
	Denial of service attacks on communication channels to prevent data transfer and data availability	4.5	**Sending a large number of data** to network components, so that it is unable to provide the expected services
	An unprivileged user is able to gain privileged access to communication channels components	4.6	An attacker is able **to elevate the privileges** of a user that he initially compromised (for example root access to network components)
	Viruses and Malware embedded in communication protocols are able to infect the data factory back end servers	4.7	Attacker can **execute arbitrary code** to launch malware attack by profiting from unpatched or 0-day vulnerabilities affecting the communication protocols
Threats to the data factory back-end	Misuse or compromise of update procedures	5.1	Compromise the data factory software monitoring and management server to **deploy malicious code** through patch

servers regarding its update procedures			/ software update distribution towards the back-end server
		5.2	**Compromise** of cryptographic keys preventing software update distribution leaving back-end servers vulnerable to cyber attacks
Threats to the Data Factory back-end servers related to unintended human actions	Legitimate users / actors take actions facilitating unintentionally the course of a cyber attack	6.1	A trusted user (developer, administrator, IT security engineer...) is tricked (**social engineering attack**) to take an action to unintentionally load a malware
		6.2	**Defined security procedures** not being properly followed and/or applied
Threats to the Data and Code of the pan-EU Data Factory hosted on computing systems: Back-end servers, data touchpoint etc.	Data extraction	7.1	**Extraction of** copyright or confidential / company secret
		7.2	**Unauthorized access to Personal Identifiable Information** (PII) such as user identity, payment account information etc.
		7.3	**Extraction** of cryptographic keys allowing to decrypt data at rest or in-motion

Mitigations to the identified threats related to the infrastructure of the “back-end servers”

Table 12: Threat mitigations for back-end servers.

Threats to “Back-end servers”	Mitigation / Countermeasure
Ref 1.1	Security Controls and policies such as “principle of least privilege”, “role-based access control” and “need-to-know principle” are defined and enforced
Ref 1.2	Security Controls and policies are applied to secure the remote access mechanism and to minimise unauthorised access. Both process and technical security controls shall be defined and a list of controls can be found in the following standards: NIST CSF, IEC 62443, ISO 27001 etc.
Ref 1.3	Through system design and access control mechanisms (technical and process controls) it should not be possible for unauthorised personnel to access personal or system critical data
Ref 1.4	Awareness training, data monitoring and incident response mechanisms shall be applied to minimise the risk related to unintended data breach

Ref 2.1	DOS / DDOS attacks protection shall be defined during system design. Procedures for Incident Response and Disaster Recovery must also be available to ensure the business continuity

Mitigations to the identified threats related to the “communication channels”

Table 13: Threat mitigations for the communication channels.

Threats to “communication channels”	Mitigation / Countermeasure
Ref 3.1	The authenticity and integrity of messages shall always be verified
Ref 4.1	Access control techniques and encryption mechanisms shall be applied to protect system data/code
Ref 4.2	
Ref 4.3	
Ref 4.4	The authenticity and integrity of messages shall always be verified
Ref 4.5	DOS / DDOS attacks protection shall be defined during system design. Procedures for Incident Response and Disaster Recovery must also be available to ensure the business continuity
Ref 4.6	Measures to prevent, detect and respond to unauthorized access shall be employed. Security Controls and policies such as “principle of least privilege”, “role-based access control” and “need-to-know principle” are defined and applied.
Ref 4.7	Measures to protect systems against embedded viruses/malware should be identified and applied. Measures to identify and respond to malware incident shall be in place

Mitigations to the identified threats related to the “update process” and central management system

Table 14: Threat mitigations for the update process.

Threats to “Update Process”	Mitigation / Countermeasure
Ref 5.1	Secure software update procedures shall be employed. Authenticity and integrity of the software update binaries shall always be verified. The central software management system shall be secured with state-of-the-art security measures
Ref 5.2	Security controls incl. technical and process measures shall be implemented for storing and managing the cryptographic keys during their entire lifecycles.

Mitigations to the identified threats related to “unintended human actions facilitating a cyber attack”

Table 15: Threat mitigations for unintended human actions.

Threats to "Unintended human actions"	Mitigation / Countermeasure
Ref 6.1	Role-based awareness campaign shall be defined and rolled-out to minimise human errors. Strong mechanisms for access privileges shall be defined (need-to-know, role-based-access, etc.).
Ref 6.2	Security procedures shall be defined, and the organization shall seek for global acceptance and commitment. Security procedures shall be defined appropriately and consider usability and simplicity.

Mitigations to the identified threats related to "Data / Code"

Table 16: Threat mitigations for data and code.

Threats to "Data / Code"	Mitigation / Countermeasure
Ref 7.1	State-of-the-art authentication and access control techniques and encryption mechanisms shall be enforced to protect system data and source code from being stolen or manipulated. Continuous data monitoring (incl. system logs collection) and Incident Response procedures shall be defined and enforced to minimise the consequence of data breach.
Ref 7.2	
Ref 7.3	Security controls shall be implemented for storing cryptographic keys e.g. use of Hardware Security Module (HSM)

5 DATA TRANSFER CONCEPT

In order to secure data transfer, it is imperative to secure data in transit but also to ensure integrity of data. The latter is important to ensure data corruption, or even tampering can be detected.

5.1 DATA TRANSPORT SECURITY

Securing the transfer of data is of utmost importance in today's digital landscape. With the increasing threats of unauthorized access and data breaches, it is crucial for organizations to implement robust security measures during the data transfer process. Two key aspects of data transfer security are encryption and securing the transport channels.

Encryption plays a pivotal role in protecting data during transmission. Implementing at least two-factor encryption ensures an added layer of security. This involves utilizing encryption algorithms that require multiple factors, such as passwords and security tokens, to access and decrypt the data. By employing this approach, organizations can significantly reduce the risk of data compromise.

Securing the transport channels is equally vital in safeguarding the data during its journey. One effective method for achieving this is through the use of a Virtual Private Network (VPN). A VPN establishes an encrypted connection between the sender and receiver, regardless of their physical location. By routing the data traffic through a secure tunnel within the VPN, organizations can prevent

unauthorized access and protect the confidentiality of the transmitted information. Within the public networks, where the risk of interception is high, it is essential to take additional precautions. The use of encrypted data tunnels within the VPN provides an extra layer of protection against potential attacks and eavesdropping attempts. These tunnels ensure that the data remains encrypted throughout the entire transfer process, safeguarding it from unauthorized access.

To facilitate secure communication, the key exchange process is crucial. When establishing a VPN connection, the necessary key exchange occurs automatically during the connection setup. This ensures that the communication between the parties is encrypted and protected from interception. Symmetric encryption, such as Private Key Encryption, is commonly employed in this process. With symmetric encryption, both communication partners share the same key, enabling efficient encryption and decryption of the data.

Considering the potential threats posed by quantum computing, it is important to future-proof data transfer security. Asymmetric encryption algorithms, which rely on mathematical problems that are challenging for conventional computers to solve, may be compromised by quantum computers in the future. Therefore, organizations should proactively adopt post-quantum cryptography techniques to ensure the long-term security of their data transmissions.

In addition to VPNs, organizations can also leverage Multiprotocol Label Switching (MPLS) to direct IP data traffic. MPLS enables efficient and secure communication by providing different levels of service for various data streams. It establishes separate virtual channels for routing the data, allowing organizations to prioritize and control the traffic flow effectively.

To enhance the security of the overall system, a modular approach can be implemented, which includes the utilization of Black Channel Communication. This approach enables secure data exchange between closed networks without revealing the content of the communication. The modular design ensures flexibility and adaptability to evolving security requirements, providing a robust and confidential data transfer solution.

In conclusion, ensuring the security of data transfers necessitates a combination of measures such as at least two-factor encryption, encryption of transmission paths via a VPN, symmetric encryption for key exchange, and the modular design of the data encryption system. By implementing these security measures, organizations can establish a secure and confidential data transfer process, mitigating the risks of unauthorized access and data breaches.

5.2 DATA INTEGRITY

In order to ensure data integrity, it must be made sure that data is not tampered with at any point in the processing chain, but also that data originates from a trusted source.

The integrity of data is crucial to ensure that information remains accurate, complete, and unaltered. Within the pan-European Data Factory, proactive measures must be taken to detect potential data anomalies and ensure data integrity throughout the entire lifecycle of the data. The following are important aspects of ensuring data integrity in a Data Factory.

One important aspect of ensuring data integrity is the use of ALCOA principles. ALCOA stands for Attributable, Legible, Contemporaneous, Original, and Accurate. These principles serve as guidelines for proper documentation and recording of data to ensure its integrity. Furthermore, the

ALCOA concept has been expanded with the ALCOA+ principle, which emphasizes the importance of data completeness, consistency, permanence, and availability. Another important aspect of ensuring data integrity is the accurate documentation of data migration. Within the network of partners within the Rail Data Factories, these data movements must be transparently and comprehensively documented, including how the data was imported and tested. Extensive preparatory work such as data cleansing, mapping, and conversions must be performed before importing data into target systems. Accurate documentation enables the tracing of the migration process, identification of errors, and ensures data integrity. In order to ensure data integrity, checksums can be derived from the data. Ideally this is done as early in the processing chain as possible, e.g. at a data touch point. These checksums must be exchanged with any consumer of the data, e.g. the data center. This way data integrity can be verified at any time. These checksums must also be exchanged with other data factories consuming the data so that they are also capable of ensuring data integrity.

To detect potential data anomalies early on and take appropriate measures, a proactive monitoring system must be implemented within the network and among the partners. This system can include data analysis and automated monitoring mechanisms that detect deviations or irregularities in the data. Early detection of anomalies allows for prompt action and restoration of data integrity. Ensuring data integrity requires an ongoing process that includes proactive measures, clear guidelines, and comprehensive documentation. By adhering to the ALCOA principles and the ALCOA+ principle, as well as accurately documenting data migration, it can be ensured that data remains accurate and complete throughout its lifecycle.

Guaranteeing the integrity of data is a fundamental requirement for any data factory. It starts with the creation of data and extends to its transmission, where it is essential to ensure that the data comes from a trusted source. To achieve this, each data source within the factory should be uniquely identified using a robust Public Key Infrastructure (PKI). This identification process increases the overall trustworthiness of the data and ensures that it can be reliably traced back to its source.

In addition, robust measures must be taken to secure any data transmission within the factory. This includes the use of strong encryption protocols to protect data as it is transferred from one location to another. Encryption ensures that data remains confidential and is protected from unauthorised access during transmission.

In addition to encryption, other security measures such as secure protocols and authentication mechanisms should be used to verify the identity of the parties involved in the data transfer. This can prevent data from being manipulated or intercepted in transit.

By focusing data quality management with data integrity and implementing strong security measures, data factories can build trust in their data sources and maintain the confidence of their stakeholders. Robust data security practices are essential to mitigate risk and ensure the integrity of the entire data lifecycle. Additional measures should be taken such as ensuring every data transfer is secured, and that data is encrypted at rest.

These functions must be implemented both in the vehicles and in the Data Touch Points and central locations to enable transparent documentation, such as for approval, and to prevent manipulation.

This is crucial to build trust in the data and minimise potential risks such as data anomalies and manipulations.

6 CONCLUSIONS AND OUTLOOK

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.

Building upon this, Deliverable D 2.1 provides an architectural analysis, functional zones, building blocks and considerations for implementation, as well as orchestration and operation.

This Deliverable, D 2.2 takes this input to provide concepts for IAM, Data Management and Data Transfer.

For the IAM concept, roles, requirements and a security and governance concept for the pan-European Rail Data Factory are evaluated. Due to federated IAM being well understood in the IT industry, in general, this relies entirely on industry standard and readily available frameworks.

For the data management concept, governance, security, data types and models are evaluated further. In future work, this must be mapped onto a concrete implementation.

For the data transfer concept, data security and data integrity measures are described on a high level. In future work requirements need to be derived.

This work will serve as an input to the further work in this study, in particular the development of a network concept in D 2.3, as well as a commercial and operational assessment in D 3 and deployment strategies in D 4.

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] 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

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

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

[7] 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

[8] Rail Label, see https://github.com/DSD-DBS/raillabel

[9] FAIR Principle, see https://www.go-fair.org/fair-principles/

[10] Bryan Petzold, Matthias Roggendorf, Kayvaun Rowshankish, and Christoph Sporleder, "Designing data governance that delivers value", 2020. [Online]. Available: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/designing-data-governance-that-delivers-value

[11] Dave Wells, "The Path to Modern Data Governance", 2019. [Online]. Available: https://www.eckerson.com/articles/modern-data-governance-problems

[12] PwC, "Global and industry frameworks for data governance", 2019. [Online]. Available: https://www.pwc.in/consulting/technology/data-and-analytics/govern-your-data/insights/global-and-industry-frameworks-for-data-governance.html

[13] SAS, "The SAS® Data Governance Framework: A Blueprint for Success", 2018. [Online]. Available: https://www.sas.com/content/dam/SAS/documents/marketing-whitepapers-ebooks/sas-whitepapers/en/sas-data-governance-framework-107325.pdf

[14] Gwen Thomas, "The DGI Data Governance Framework". [Online]. Available: https://datagovernance.com/wp-content/uploads/2020/07/dgi_data_governance_framework.pdf

[15] Gaia-X, „Gaia-X Framework". [Online]. Available: https://docs.gaia-x.eu/framework/

[16] The Global Data Management Community. Publications >> Books Referenced in DMBoK V2. Available: https://www.dama.org/cpages/books-referenced-in-dama-dmbok