CEF2 RailDataFactory

D1 – Data Factory Concept, Use Cases and Requirements

Version 1.1

Due date of deliverable: 31/03/2023

Actual submission date: 28/04/2023

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

Document status		
Revision	Date	Description
0.1	09/03/2023	Document template generated
0.2	24/03/2023	Major parts of content transferred from Confluence
0.3	29/03/2023	First complete draft
0.4	04/04/2023	Draft version submitted to advisory board
0.5	11/04/2023	Use case and requirements sections merged
0.6	19/04/2023	First review and commenting the advisory board comments
0.7	24/04/2023	Final version after addressing of all advisory board comments, sent for final consortium approval
1.0	28/04/2023	Version submitted to Sygma portal
1.1	03/05/2023	Correction on formatting and spelling errors

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	Sensitive – limited under the conditions of the Grant Agreement	

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

ACKNOWLEDGEMENTS

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

REPORT CONTRIBUTORS (IN ALPHABETICAL ORDER)

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

Note of Thanks

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

Disclaimer

Funded by the European Union. The information in this document is provided “as is”, and no guarantee or warranty is given that the information is fit for any particular purpose. The content of this document reflects only the authors’ views and does not necessarily reflect those of the European Union or the European Health and digital Executive Agency (HaDEA). Neither the European Union, nor the granting authority, nor the project consortium take any responsibility for any use of the information contained in this deliverable. The users use the information at their sole risk and liability.

Licensing

EXECUTIVE SUMMARY

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

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

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

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

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

ABBREVIATIONS AND ACRONYMS

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

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

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

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

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

1.1 Aim and Scope of the CEF2 RailDataFactory Study

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

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

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

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

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

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

1.2 Delineation from and Relation to other Works

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

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

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

The input from the TAURO project is, however, taken into consideration in particular in the derivation of use cases for the Data Factory, as covered in Chapter 5.

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

Within the sector initiative "Digitale Schiene Deutschland", Deutsche Bahn already started to set up some components of the Data Factory [11].

1.3 Aim and Structure of this Deliverable

This current document is the first deliverable D1 of the CEF 2 RailDataFactory project, covering the basic concept of the envisioned pan-European Data Factory, identified operational scenarios and use cases, their requirements specifically on the pan-European interconnectivity required for the Data Factory, and additional legal, regulatory or Cyber-security related aspects that have to be addressed.

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

The remainder of this document is structured as follows:

- In Chapter 2, the vision of the pan-European Data Factory is shortly introduced, together with key terminology used in the remainder of this document;
- In Chapter 3, a representative operations scenario is depicted, both from the perspective of rail operations, and from a technical perspective;
- In Chapter 4, the envisioned contributor concept of the Data Factory is introduced, and roles are defined;
- In Chapter 5, key Data Factory use cases are identified, with a distinction between general use cases and technical use cases especially relevant for the pan-European Data Factory backbone network in the center of this study, and requirements derived from the identified use cases are listed;
- In Chapter 6, legal, regulatory and Cyber-security related aspects are identified that have to be considered in the realization of the pan-European Data Factory.
- Finally, in Chapter 7, this document is concluded with a summary and the expected next steps in the study.

2 VISION AND DEFINITION OF THE DATA FACTORY

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

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

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

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

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

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

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

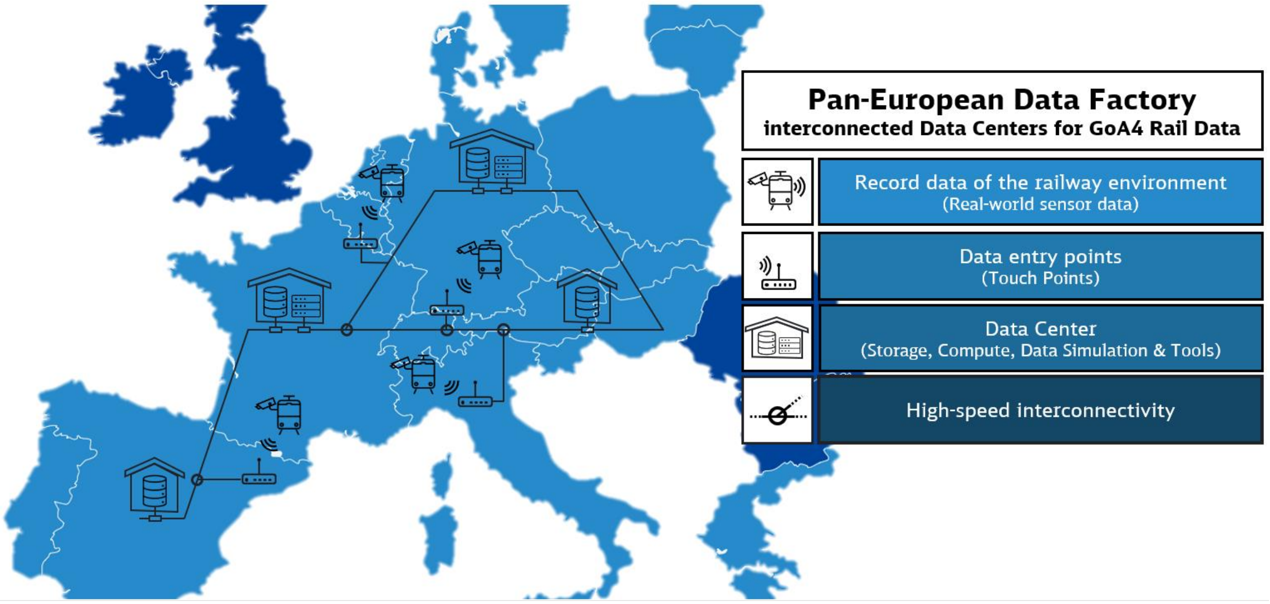

Pan-European Data Factory
interconnected Data Centers for GoA4 Rail Data
Record data of the railway environment
(Real-world sensor data)
Data entry points
(Touch Points)
Data Center
(Storage, Compute, Data Simulation & Tools)
High-speed interconnectivity

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

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

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

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

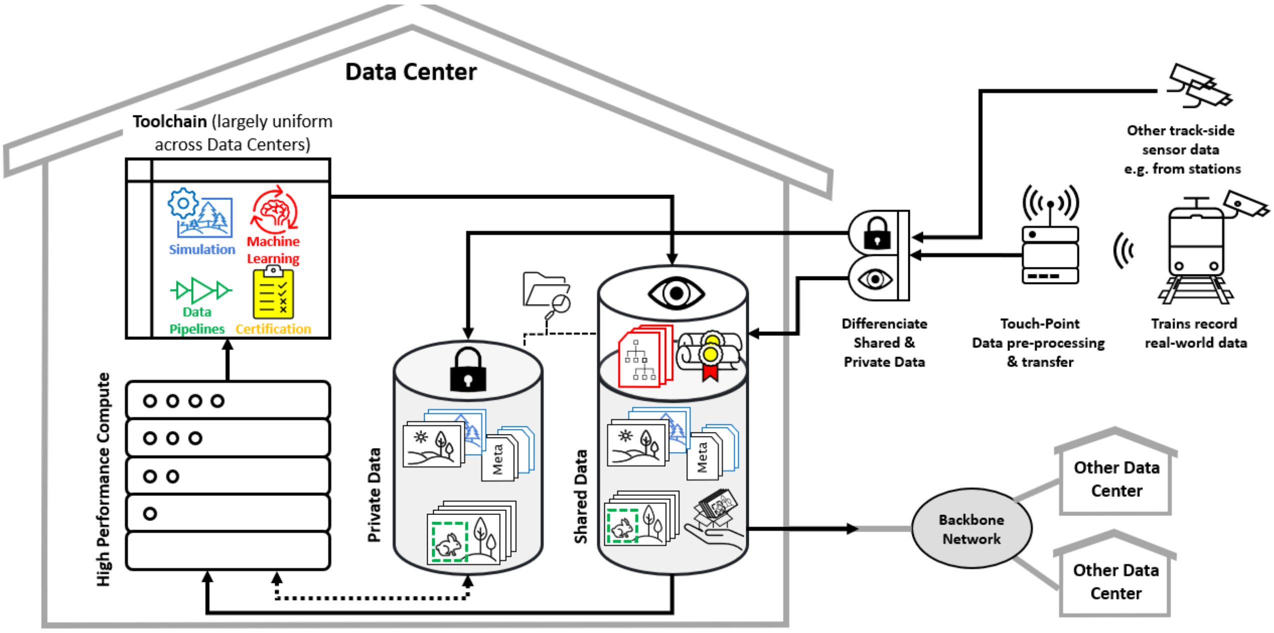

Data Center
Toolchain (largely uniform across Data Centers)
Simulation
Machine Learning
Data Pipelines
Certification
High Performance Compute
Private Data
Shared Data
Meta
Meta
Other track-side sensor data e.g. from stations
Differenciate Shared & Private Data
Touch-Point Data pre-processing & transfer
Trains record real-world data
Backbone Network
Other Data Center
Other Data Center

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

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

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

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

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

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

3 REPRESENTATIVE OPERATIONS SCENARIO

In this chapter, a representative operations scenario for the pan-European Rail Data Factory is described, both from the perspective of railway operations, and from a technical perspective.

3.1 OPERATIONS SCENARIO FROM RAILWAY OPERATIONS POINT OF VIEW

As an RU, flexibility is needed to send trains across Europe. A freight train can be scheduled to go from, e.g., the Port of Rotterdam in the Netherlands, head via Belgium to Metz in France, switch cargo there and return to the port of Rotterdam via Mannheim in Germany where cargo is switched once again.

As this train passes through the countries involved, responsibilities change. While the railway undertaking never changes along the journey, the train passes through multiple IMs' networks. These IMs require knowledge of RUs operation rules, and the RU needs knowledge about the IMs' track infrastructure. Additionally, the IMs supervise the train for as long as it is within their network. Envisioning an autonomously driving train, all of these functions will need to be automated. This requires standardized digital interfaces over which relevant data can be exchanged as well as relevant systems on-board and trackside to issue and interpret these data.

One of the big challenges in this regard is the development of AI models capable of recognizing the track, its surroundings and objects within the vicinity. This requires vast amounts of relevant data, e.g., from cameras, lidars and radars, so that a model can be trained to reliably detect tracks, catenaries, bridges and other objects. As these objects may differ from country to country, it is required that a train journeying through multiple countries is equipped with appropriate AI models to reliably recognize objects in these countries. Furthermore, in an operational system, cases can still arise where recognition is too inaccurate, resulting in a non-detected incident or a false positive. In these cases data needs to be captured and handed over to the responsible IM, so that this can take the appropriate steps, likely consisting of retraining and recertification of the corresponding AI model.

Additionally, further use cases can be envisioned like the recognition of a need for maintenance on or around the tracks. The communication of such information requires appropriate, standardized, communications channels, description and documentation methods as well as appropriate storage solutions.

In total, the Pan-European Data Factory hence needs to provide the means to

- transfer and store the collected sensor data from trains to the data entry points and to the Data Centers;
- generate artificial sensor data in simulations;
- use both types of data to train and certify AI models for fully-automated rail operation;
- exchange this data among data centers within the Data Factory in order to enable the training of AI models enabling cross-border rail traffic;
- enable RUs to download the AI models they need for operating in specific areas and enable them to document incidents.

3.2 Operations Scenario from Technical Point of View

A basic prerequisite for developing fully automated driving is the collection and then efficient distribution of all the required sensor data. ~~This means that the data collected on the train during the journeys is then transferred in full to the pan-European Data Factory via a secure path.~~ This means that the data collected by the train along the route is then transferred in full to the pan-European Data Factory via a secure connection. Once the data has been quality-assured and stored, it can also be made available to other participants and data centers in the network.

The models required for fully automated driving must also be generated. This requires special hardware and software, which must be explicitly available for this purpose. Since the training effort, as well as the re-training, of these AI models is extremely computationally intensive, and due to the computational properties of the AI training methods used typically requiring all necessary data to be available locally - i.e., at the same location - it must be possible to transfer the data to the data center intended for this purpose. Therefore, it is necessary that the railway companies can always transfer these data to the appropriate and designated data processing points and sinks.

Not only the transfer of the data and the training of an AI model is of great importance, but also the acceptance tests to achieve certification and thus approval of the AI model. This AI model is the core, which - after it is transferred to the train - will enable the fully automated train driving in the end.

In order to make the whole product complete, logging and monitoring mechanisms must of course be in place and made available everywhere, which track and observe the entire data life cycle. This also means that if incidents are recorded in the sensor, they can be investigated and evaluated by the railway and infrastructure companies and documented.

For the operational scenario, it is important to exchange data between data sources and data centers, to have all needed data available to train AI models, for instance used in neural networks. To train neural networks it is necessary to have the data locally because of different kind of reasons, e.g., performance, latency, cost to move data around, etc.

Figure 3 addresses this communication between data sources and data centers. They have to exchange different kinds of data, e.g., raw sensor data, metadata, annotations, trained neural networks, etc.

How it will work

Trains travelling through Europe will record a very large amount of data due to the camera and sensor technologies installed in and on the trains. This collected data is then transferred via so-called Touchpoints or other technologies to the designated European facilities as the amount of data collected cannot feasibly be transferred using terrestrial mobile networks. This means that the data is first available to a facility in the respective country which reads out the sensor data. Once this data has been quality-checked and complemented with metadata, this metadata is shared as preliminary information with the other facilities via the **High-Speed Pan-European Railway Data Factory Backbone Network**.

Thus, it is possible to form a **Unified Data Management Catalog** allowing each facility to map all available data across Europe.

Users will be able to work with the data using the **uniform Tool Chain** with **harmonized protocols and data formats**. In this respect, identity and Access Management (IAM) will provide a centralized management of identities and access rights to the various services in a facility, as well as ensuring correct and secure authentication and authorization of users.

The IAM portal is the identity and access management system that connects each user to the correct access level in a secure manner. Individual IAM portals will grant access to the facilities, data and tools, see also Figure 3.

It should be noted that not all sensor data are immediately sent throughout Europe, but first all information (metadata) regarding these sensor data. The background to this is that this sensor data generates very heavy data volumes due to a large number of recording technologies, which have a considerable impact on the backbone network.

If now a user logs on to a European facility and detects via the Data Management Catalog that there is new or changed data, this user can decide himself to transfer this data in order to enrich his data master with further required data. Only with this approach is it possible to train AI-based models, which are then no longer bound to a specific location or country, but can act across Europe.

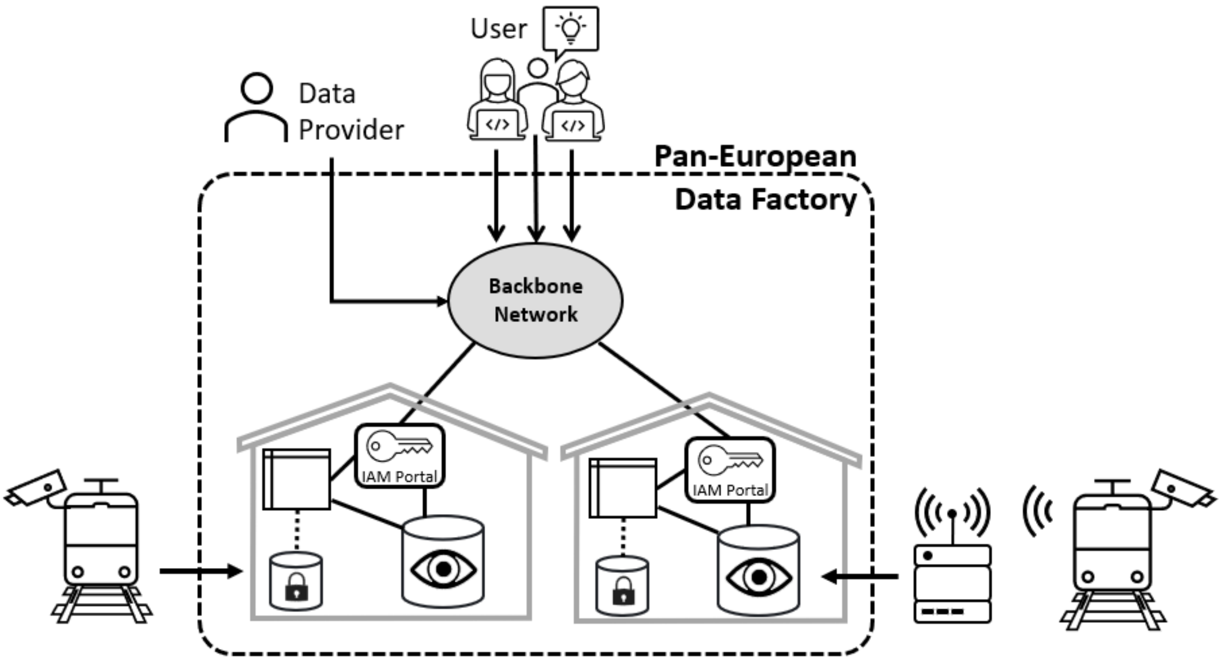


Figure 3. Representative operations scenario of the Data Factory (from technical perspective).

4 DATA FACTORY CONTRIBUTION CONCEPT AND ROLES

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

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

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

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

- Providing external computing power.

In the remainder of this document, roles as defined in Table 2 and illustrated in Figure 4 are used.

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

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

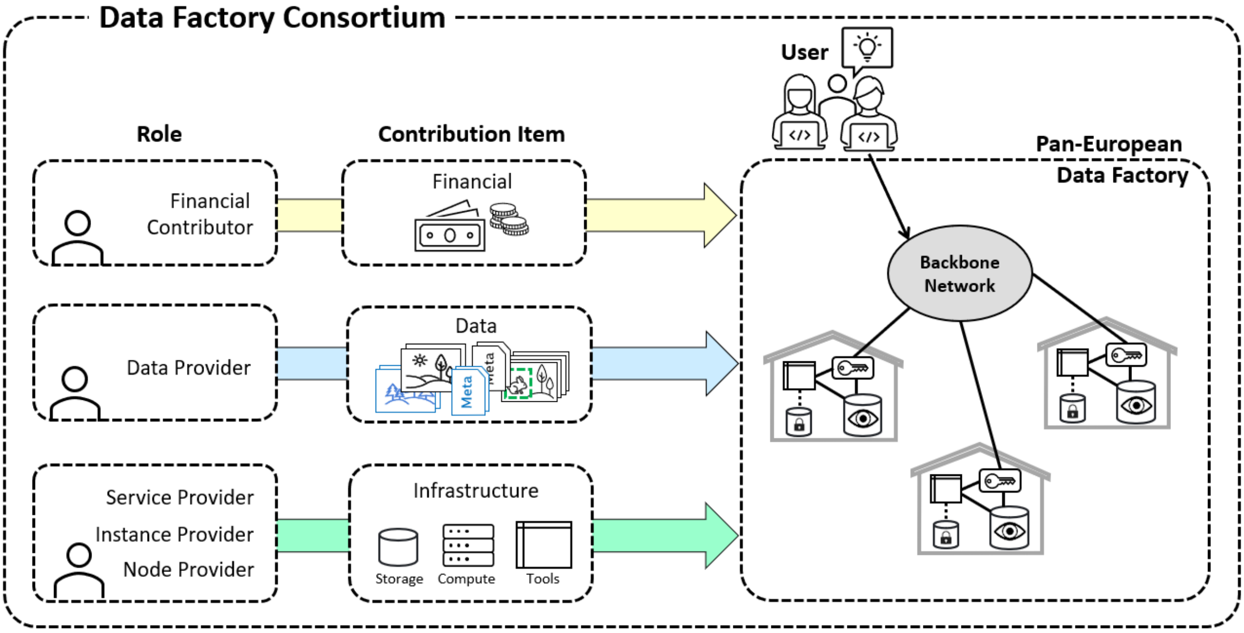


Figure 4. Illustration of the roles involved in the Pan-European Data Factory.

5 DATA FACTORY MAIN USE CASES AND REQUIREMENTS

This chapter addresses the main use cases identified for the Data Factory. These are based on earlier work by Shift2Rail TAURO project [4] and have been aligned with the views in the Europe's Rail FP2 R2DATO project [5].

We differentiate between **general** and **technical use cases.** The general use cases in Section 5.1 have been formulated from the perspective of the user or consumer, and the technical use cases in Section 5.2. address the the required pan-European backbone network and data platforms. Both sections also include the requirements derived from the use cases.

In this section, all requirements of the functional use cases are described and listed.

Note: These requirements are also known in R2DATO and are elaborated, specified and implemented there. In the CEF 2 project the necessary preparatory work is done and the necessary basic understanding is acquired.

Figure 5 shows all (general and technical) use cases covered in this deliverable and depicts their relation.

○ These technical use cases, highlighted in blue, are those relevant to CEF II. They describe the **Highspeed Railway Data Factory Backbone Network**.

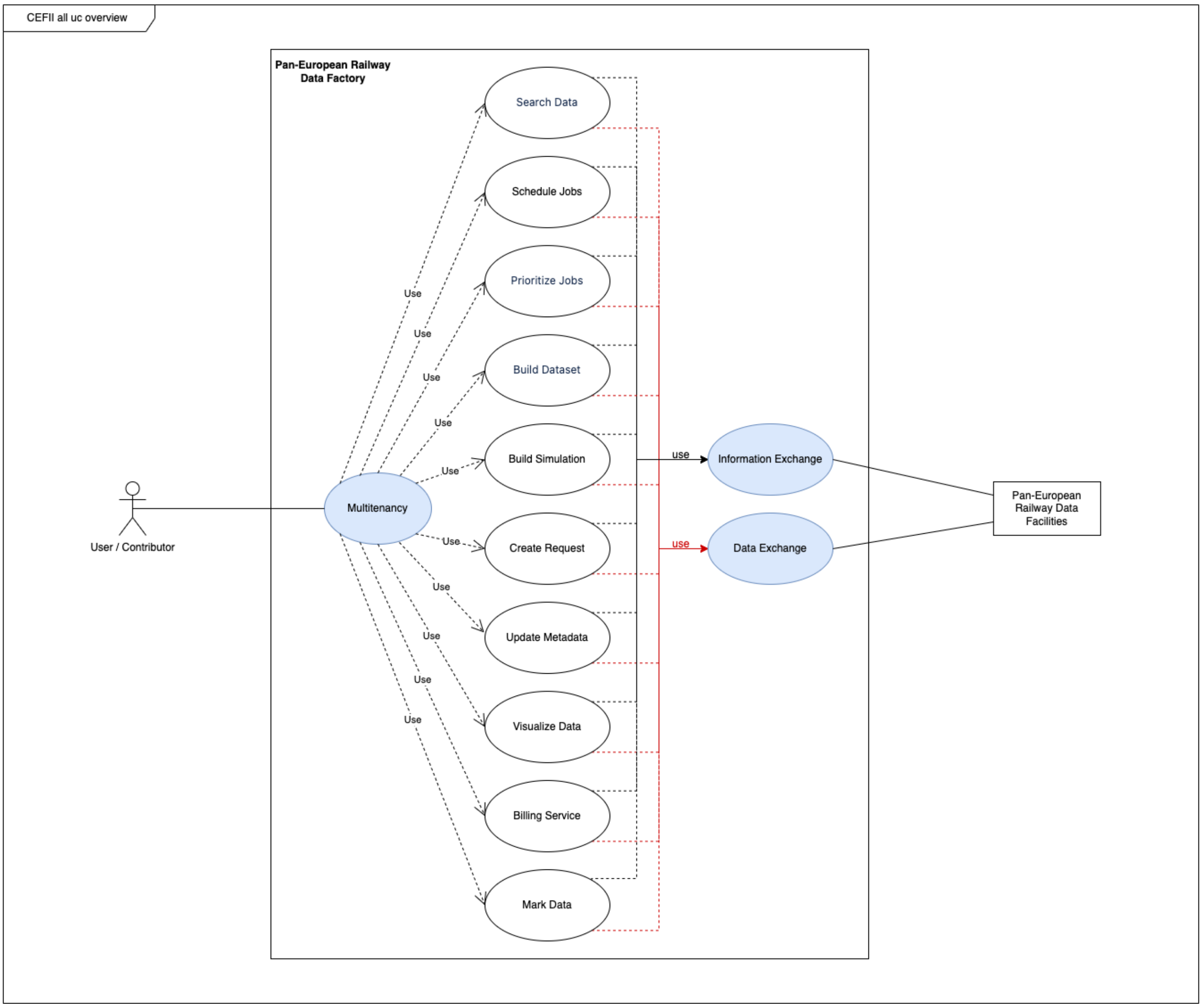


Figure 5. Overview of the general and technical use cases covered in this deliverable.

5.1 General Data Factory Use Cases and requirements

As stated before, in this section the general Data Factory use cases and requirements are listed from an user perspective.

5.1.1 Use Case 1: Search Data

In order to be able to search for specific resources within the Data Factory or a data center, there must be the possibility of a general search for all assets that are available to a user. Therefore, it must also be possible to send queries to all pan-European facilities connected to the Data Factory. This query, as well as the receiving of the queries, and the results must be manageable. This means that search queries are sorted and filtered, and that frequently used searches are saved and kept editable. The searches themselves, as well as their results can be filtered and sorted. Likewise, the resources to be searched for must be uniformly named and made available for a suitable search. Searches and filtering can be done for data, metadata, simulations, 3D assets, assets. All activities must be monitored and logged.

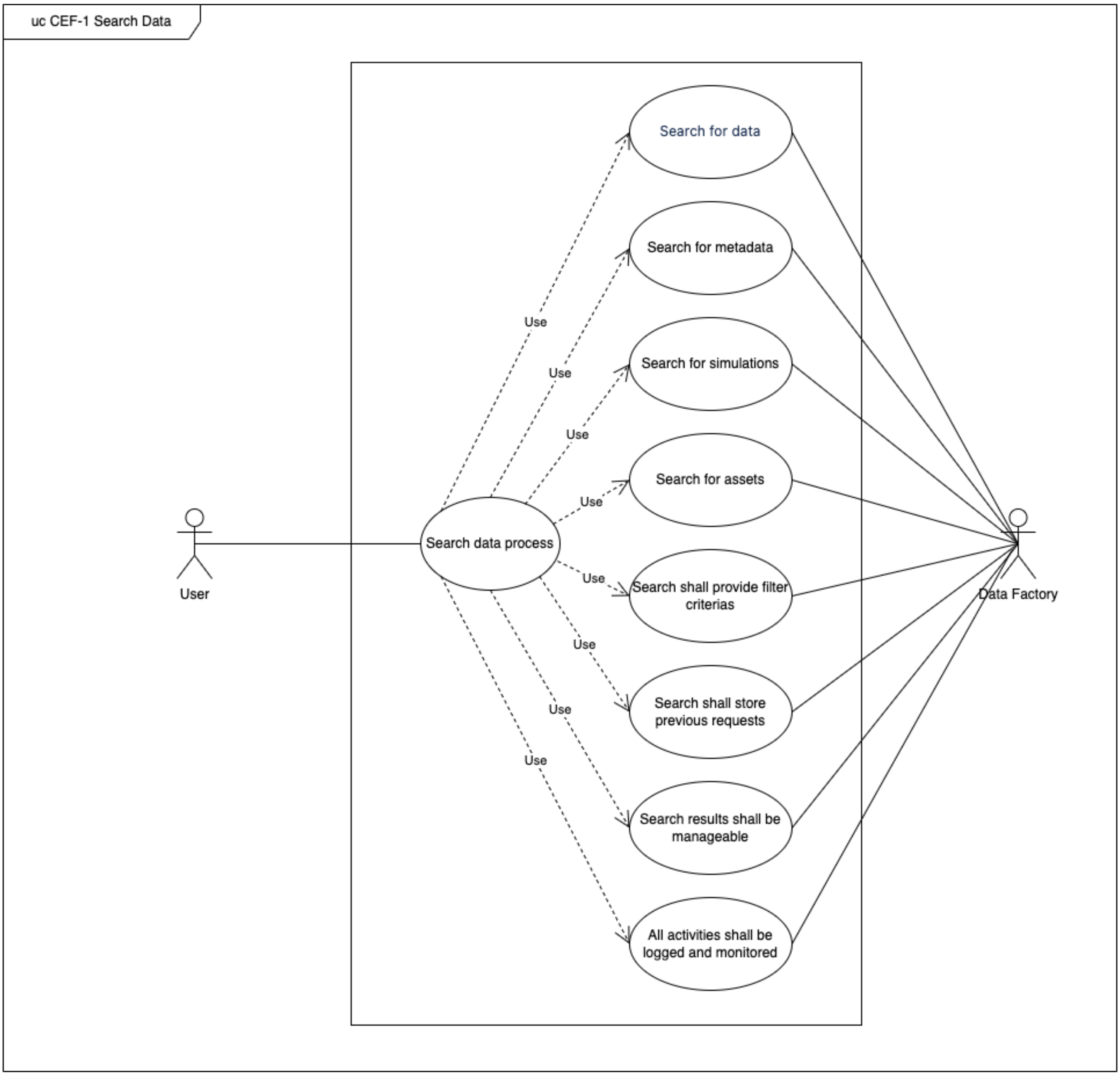

uc CEF-1 Search Data
User
Search data process
Use
Search for data
Search for metadata
Search for simulations
Search for assets
Search shall provide filter criterias
Search shall store previous requests
Search results shall be manageable
All activities shall be logged and monitored
Data Factory

Figure 6. Use Case 1: Search data within the Pan-European Data Factory.

Requirements for Use Case 1: Search Data

Requirements	Description
Search for data	The search for data includes the possibility for a user to search for data within the entire Data Factory. All data available to the user should be able to be found and searched.

Search for metadata	The search for metadata includes the possibility for a user to search for all metadata within the entire Data Factory. All metadata available to the user should be able to be found and searched. Metadata are structured information and contains information about characteristics of other data.
Search for simulations	If a user needs to generate artificial data or needs to render virtual environments, this is done by simulation. In order to find various types of simulations, those needs to be searchable.
Search for assets	Within the Data Factory, it should be possible to search for any type of required asset. This search can range from a report, raw sensor data and software to hardware.
Search shall provide filter criteria	In order to make the searches within the Data Factory more pleasant and more fluid or efficient for the user, the search must be simplified by means of predefined filters. In this way, it is possible to search for desired attributes more quickly.
Search shall store previous requests	In order to be able to search even more efficiently and pleasantly, and to avoid the user from having to create his eventual longer and more complex searches from scratch, there is the possibility to save all the previous searches. Thus, it is possible to keep the searches and make adjustments to get more efficiently to the information searched for. Note: The storage of search queries is a user-specific and user-dependent function and can be activated by the user himself, per search, if required. The user himself can manage this service.
Search results shall be manageable	When a search returns the results, these may be several answers depending on the search. Therefore, it should be possible to manage and sort the results to provide an overview appropriate to the user.
All activities shall be logged and monitored	Every activity or event, trigger, exchange and access must be logged and monitored

5.1.2 Use Case 2: Schedule Jobs

A number of different processes will run within the Pan-European Data Factory. It must therefore be possible to set the jobs required for training AI models, running simulations or rendering image material in such a way that they are processed correctly and in the best possible way. This requires a scheduler that not only enables this, but which is also able to query connected facilities about their current workload. On the one hand, this feedback must be displayed transparently to the user so that he knows where his job is running and, on the other hand, how long the expected processing time is. Further the user must have the possibility to be informed about the progress any time. It must also be ensured that all data required for processing is transported securely to the correct location. If the job has been processed and completed according to the specifications, the user must be informed of this.

Here, same as for use case 1, all activities must be recorded and monitored.

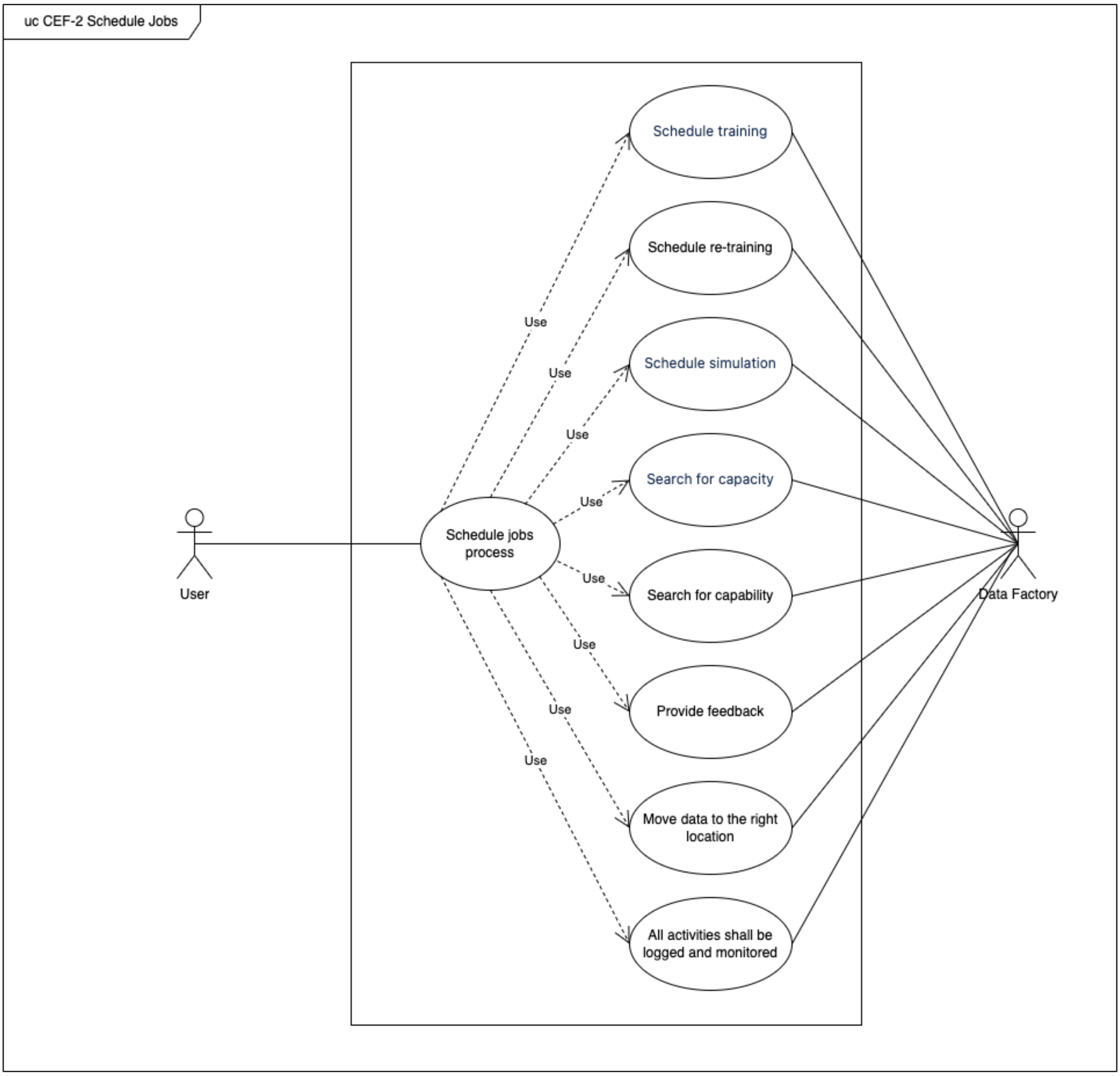


Figure 7. Use Case 2: Schedule jobs within the Pan-European Data Factory.

Requirements for Use Case 2: Schedule Jobs

Requirements	Description
Schedule training	In order to obtain an AI model, an AI algorithm must first be appropriately learned, i.e. trained. And this training is pushed into a queue by an automated job. If the required capacities are not immediately available, the job remains in this queue until the job can start running.
Schedule re-training	As with the training of an AI model, so also with the re-training of an AI model, this job is automatically written into a queue to wait for the processing, immediately the required capacities are not free. The job

	is executed immediately as soon as sufficient free capacities are available again.
Schedule simulation	Also, when running the generation of artificial data by means of a rendering of a simulation, this job is automatically written to a queue and executed as soon as sufficient free capacity is available for it.
Search for capacity	In order not to run into the risk of jobs getting stuck or not being able to be executed because capacities within the data center are currently scarce and in use, it must be possible to perform a search across the entire data factory and start jobs immediately.
Search for capability	Assuming that in production not all data centers connected within the Data Factory will be equipped in the same way, it should be possible to search the available services of the individual data centers.
Provide feedback	The user gets transparency about running jobs and where they run and about the job history..
Move data to the right location	For ML training, it is important that the visual material is located at the same place where the training takes place. Therefore, it must be ensured that all components and image materials important for the training are available at the required location in time for the start of the training.
All activities shall be logged and monitored	Every activity or event, trigger, exchange and access must be logged and monitored

5.1.3 Use Case 3: Prioritize Jobs

If a job needs to be adjusted, rescheduled or cancelled within the pan-European Data Factory before processing, then a search must be available to find the jobs that have not yet finished. Likewise, such a search must be performed if a job needs to be reprioritized. Also, the search must provide a search result display that can be sorted according to user preferences. For the reprioritization different levels are available, which can assign a new priority after a selection and confirmation. Again, all activities must be recorded and monitored.

In general, prioritizing a job has two aspects. On the one hand, an automatism must ensure that the resources are perfectly utilized in order to be able to process as many jobs as possible at the same time. And on the other hand, there must be a guidance to be able to classify the urgency of high-priority jobs. Note: It is assumed that users of the Pan-European Data Factory can naturally determine the order of priority of their own jobs. For the prioritization among the jobs of different users, likely some governance has to be setup, which is beyond the scope of this study.

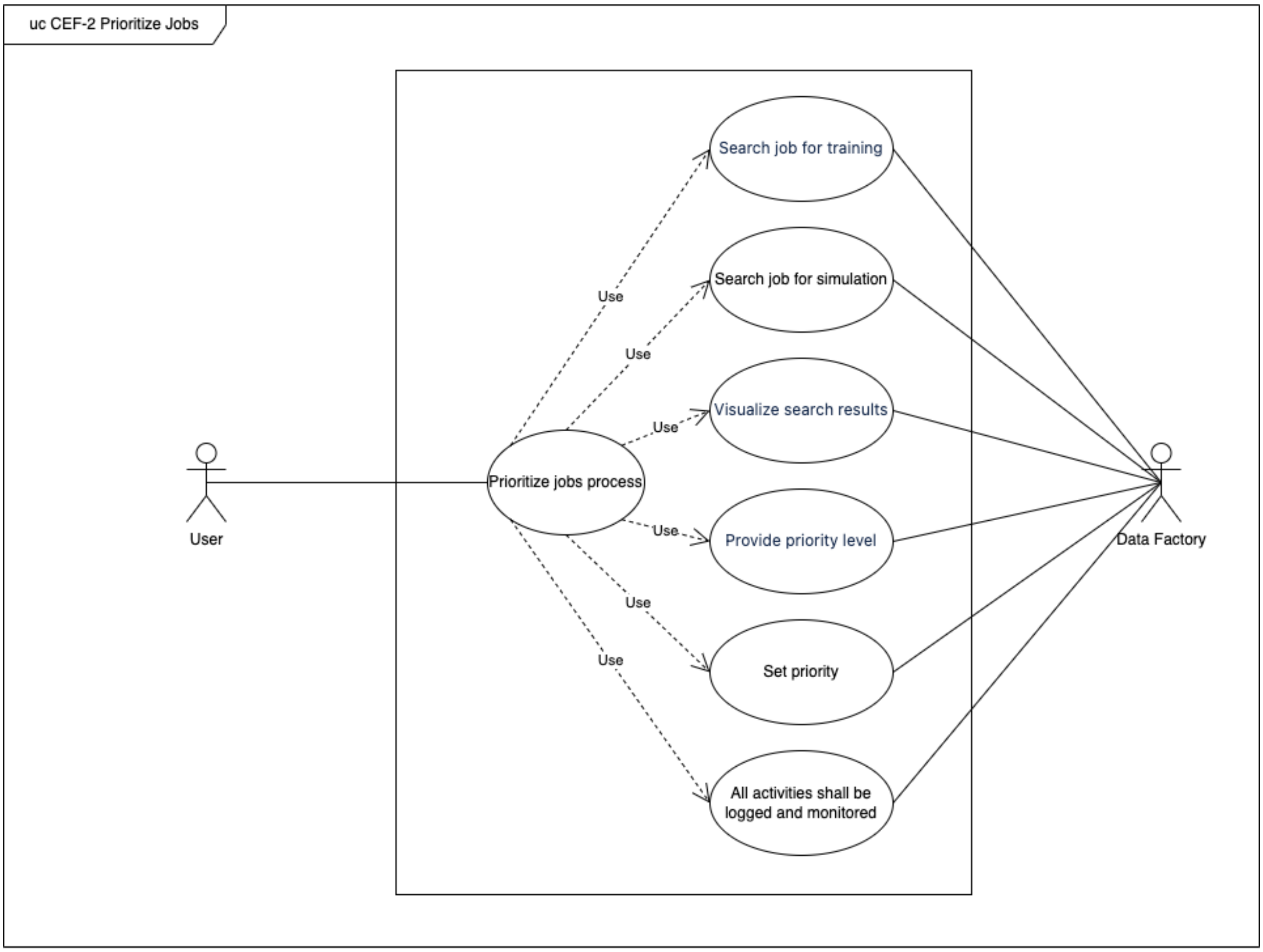


Figure 8. Use Case 3: Prioritize jobs within the pan-European Data Factory.

Requirements for Use Case 3: Prioritize Jobs

Requirements	Description
Search jobs for training	In order for the user to get an overview of his ML jobs, to be able to reprioritize if necessary, he must be able to search specifically for one job or for several jobs.
Search jobs for simulation	In order for the user to get an overview of his render jobs, to be able to reprioritize if necessary, the user shall be able to search specifically for one simulation job or for several simulation jobs.
Visualize search results	The results of the user's search for jobs, they are displayed to him, with all available work statuses and also where in a queue the job currently is.
Provide priority level	Within the whole system there must be a selection of different priority levels from which users can choose for their jobs. By default, all jobs are given the same priority level.

Set priority	Via the search, a user can have his current and pending jobs displayed. The user can then also assign new priority levels to these jobs and change them.
All activities shall be logged and monitored	Every activity or event, trigger, exchange and access must be logged and monitored

5.1.4 Use Case 4: Build Dataset

Probably the most important cause for creating data sets within the pan-European Data Factory is for training the AI models. In order to be able to create as many and as complete datasets as possible for this purpose, there must be a search function that recognizes the data available for this purpose in all connected facilities and makes it available to the user as a result. These results must be visualized and made available to the user, and they must also be markable in order to be able to retrieve and transfer them if the machine learning (ML) training does not take place locally. It is important to note that if the data originates from other connected data sources, these must be marked accordingly in the data set if they are to be deleted. If a data set is no longer needed, it must be deleted. However, before the data can be deleted, it must be checked which dependencies exist for this data and whether these can also be deleted, for example a search query or the search history. Care must also be taken to ensure that no data is removed that may also be required for the billing process. Again, all activities are recorded and monitored.

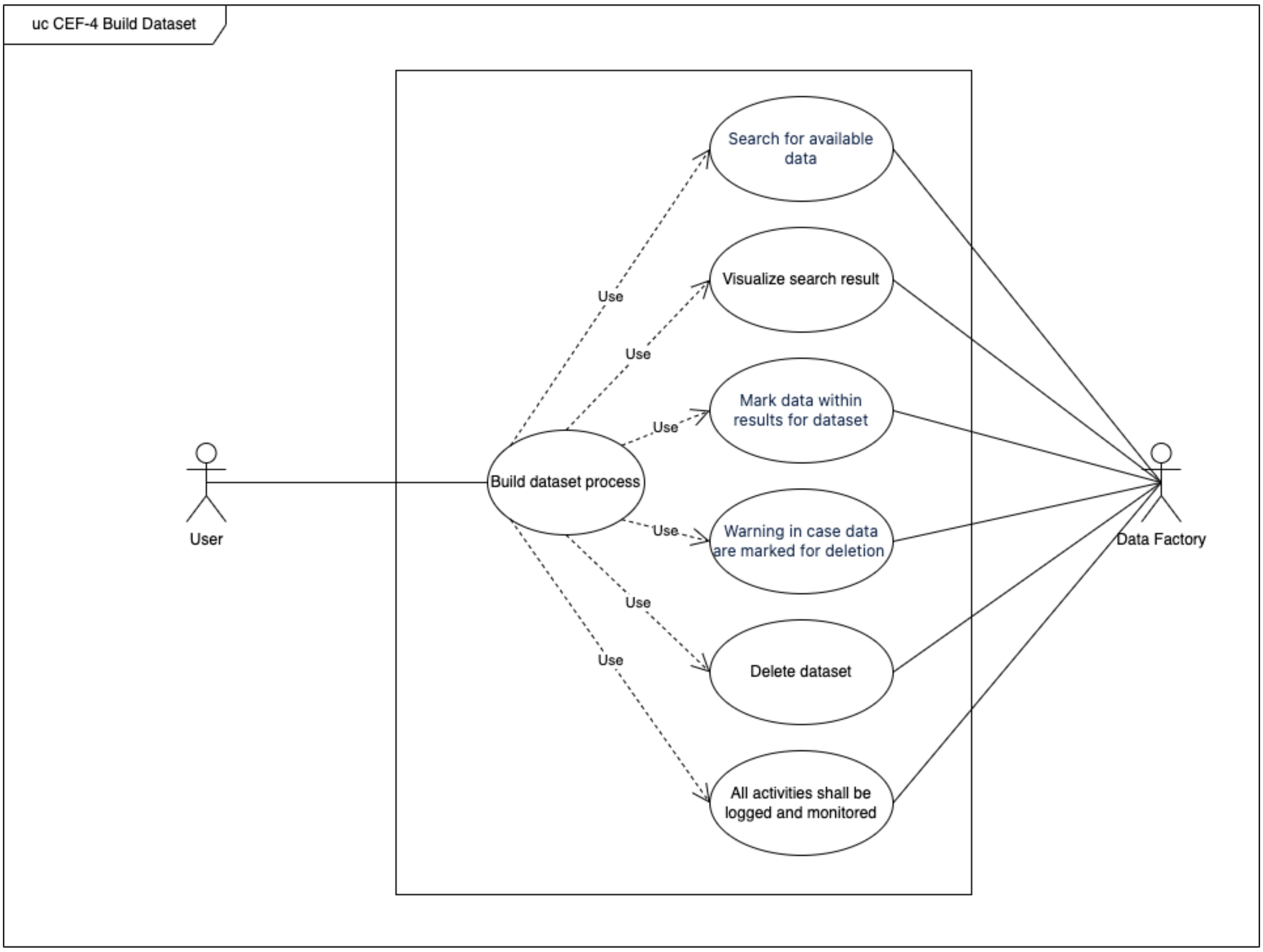

uc CEF-4 Build Dataset
User
Build dataset process
Use
Use
Use
Use
Use
Use
Search for available data
Visualize search result
Mark data within results for dataset
Warning in case data are marked for deletion
Delete dataset
All activities shall be logged and monitored
Data Factory

Figure 9. Use Case 4: Build dataset within the Pan-European Data Factory.

Requirements for Use Case 4: Build Dataset

Requirements	Description
Search for available data	In order to have sufficient data available for ML training, it may be necessary to search for or request additional data in order to generate a complete data set. Therefore, it is necessary to be able to query for available data within the Data Factory.
Visualize search result	Again, it is advantageous to look at the searched and returned results before a data transfer is to take place. Therefore, the results must be visually viewable.
Mark data within results for dataset	Once a user has viewed and filtered the results after a search and determined which data should be transferred and added to a dataset, it should be possible to select all the required data of the result and add it to the dataset.
Warning in case data marked for deletion	Of course, it can also happen that datasets required over time are formed and are in use, just as it can happen that older data is to be deleted in a data center. If this is the case, then this data to be

	deleted should be displayed in all datasets of the Data Factory with a warning.
Delete dataset	Likewise, there must be a possibility, due to various reasons, to delete data sets that are no longer needed and that were previously created.
All activities shall be logged and monitored	Every activity or event, trigger, exchange and access must be logged and monitored

5.1.5 Use Case 5: Build Simulation

In order to be able to train situations in the environment of automated train driving that one would like to avoid or that are very difficult to generate, there is the service of generating them by means of simulations. So that not every contributor has to create its own simulations, the pan-European Data Factory also offers the option of searching for existing simulations. It is also possible to search for 3D assets if you need to create your own simulation. It is also possible to combine 3D assets with an existing scene. Furthermore, an import of own created 3D assets is advantageous and provided as a service, as well as to view the entire composition in advance as a preview. All activities must be monitored and recorded.

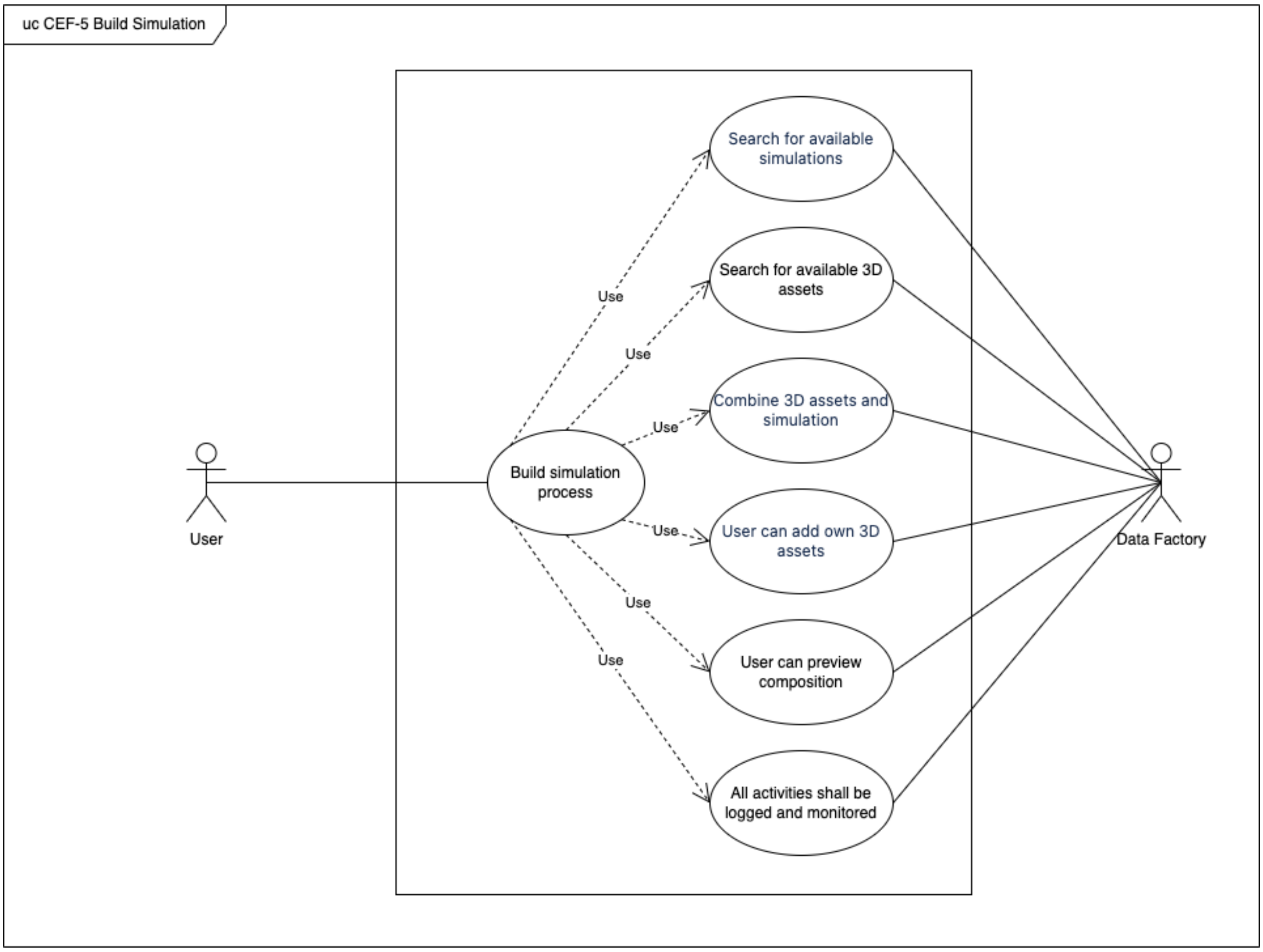

uc CEF-5 Build Simulation
User
Build simulation process
Use
Use
Use
Use
Use
Use
Search for available simulations
Search for available 3D assets
Combine 3D assets and simulation
User can add own 3D assets
User can preview composition
All activities shall be logged and monitored
Data Factory

Figure 10. Use Case 5: Build simulation within the Pan-European Data Factory.

Requirements for Use Case 5: Build Simulation

Requirements	Description
Search for available simulations	If a user wants to be able to produce artificial data by simulation as quickly and easily as possible, an intelligent search must be available to look for simulations that already exist. This can also result in a great time saving.
Search for available 3D assets	It can also happen that a user of a simulation finds out that further 3D objects are still needed to complete the scene and these are not available in the current data center. For this purpose, there is another search function to query all data centers connected to the Data Factory.
Combine 3D assets and simulation	In case a user wants to create a new artificial scene or edit or extend an existing one, there is the possibility to add more 3D assets to this scene.
User can add own 3D assets	If a user wants to bring own or new and currently not existing 3D assets into the Data Factory, there will be a special import process for this.
User can preview composition	Before the user wants to render his simulation, which can be quite lengthy, he is offered the possibility to view this simulation beforehand, in order to be able to make adjustments and changes to it ahead of time, if necessary, before the job is placed in the queue accordingly.
All activities shall be logged and monitored	Every activity or event, trigger, exchange and access must be logged and monitored

5.1.6 Use Case 6: Create Request

In each connected data center or facility it should be possible for the user to create and send his own queries ad hoc. These requests can have the following content:

- Requests to get more data;
- Requests to get resource needs (it is likely that not every connected data center is equally equipped in terms of software, tools, service and hardware);
- Request for more simulations or simulation data or their results;
- Request for new or currently also not yet existing 3D assets;
- Request for metadata or other information.

All requests and activities shall be recorded and monitored.

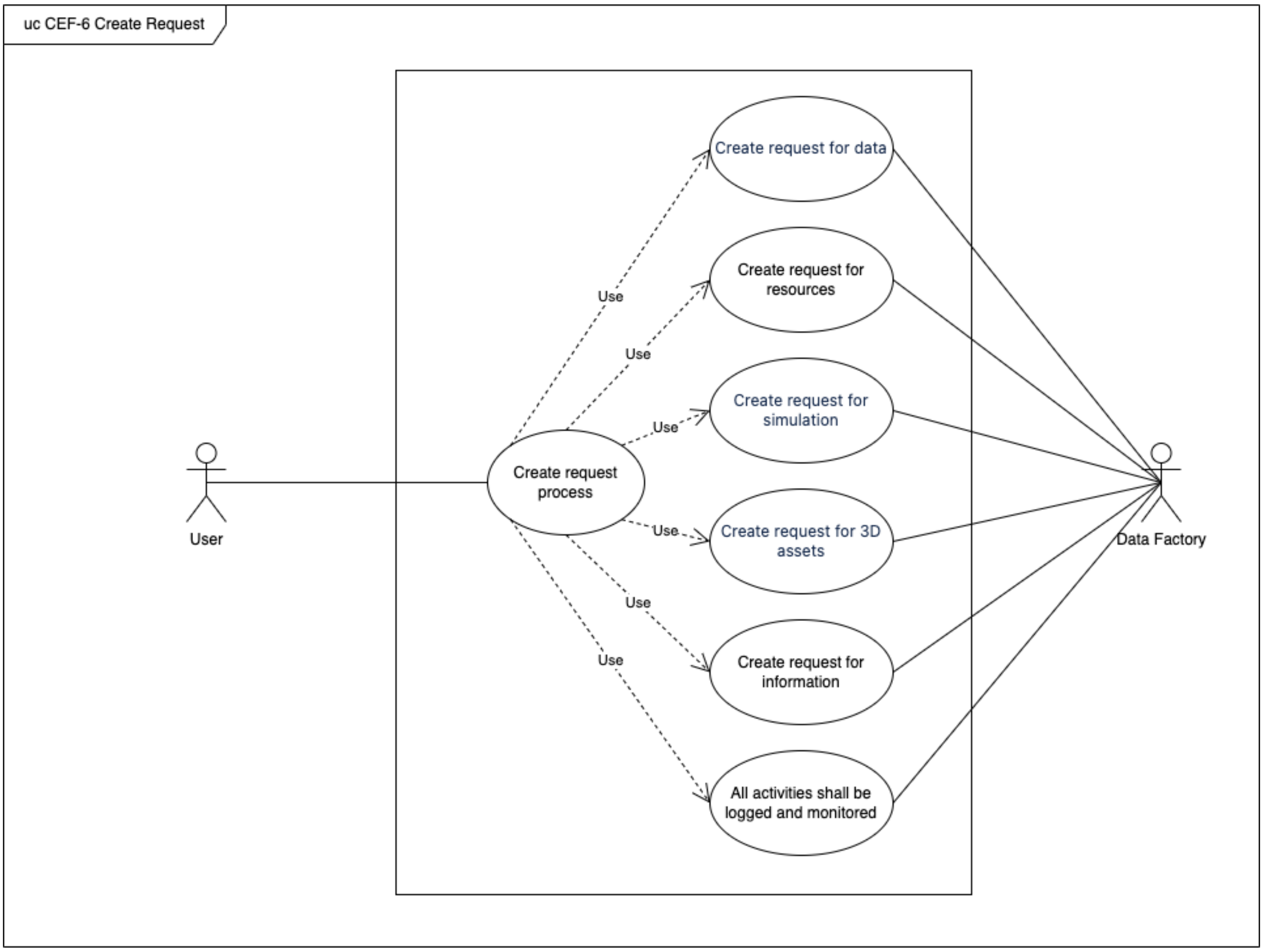

uc CEF-6 Create Request
User
Create request process
Use
Use
Use
Use
Use
Use
Create request for data
Create request for resources
Create request for simulation
Create request for 3D assets
Create request for information
All activities shall be logged and monitored
Data Factory

Figure 11. Use Case 6: Create request within the Pan-European Data Factory.

Requirements for Use Case 6: Create Request

Requirements	Description
Create request for data	Each search must, of course, generate a corresponding request. This request generates a request for data and sends it automatically via the backbone network to all connected data centers of the Data Factory.
Create request for resources	Each search must, of course, generate a corresponding request. This request generates a request for resources and sends it automatically via the backbone network to all connected data centers of the Data Factory.
Create request for simulation	Each search must, of course, generate a corresponding request. This request generates a request for simulations and sends it automatically via the backbone network to all connected data centers of the Data Factory.
Create request for 3D assets	Each search must, of course, generate a corresponding request. This request generates a request for 3D assets and sends it

	automatically via the backbone network to all connected data centers of the Data Factory.
Create request for information	Each search must, of course, generate a corresponding request. This request generates a request for information and sends it automatically via the backbone network to all connected data centers of the Data Factory.
All activities shall be logged and monitored	Every activity or event, trigger, exchange and access must be logged and monitored

5.1.7 Use Case 7: Update Metadata

In the context of qualified data and data management, the management of metadata is indispensable and necessary to enable an exact search for data. Based on this, metadata must be extendable by further fields and be able to be filled in by the user. Likewise, editing metadata is important, as well as not deleting obsolete or no longer needed metadata and metadata fields. A search over metadata, their contents and the fields themselves is to be made possible, as well as an arranging after a hierarchical order. The structure of the metadata is to be kept constant in all connected facilities.

Again, all activities are to be stored and monitored.

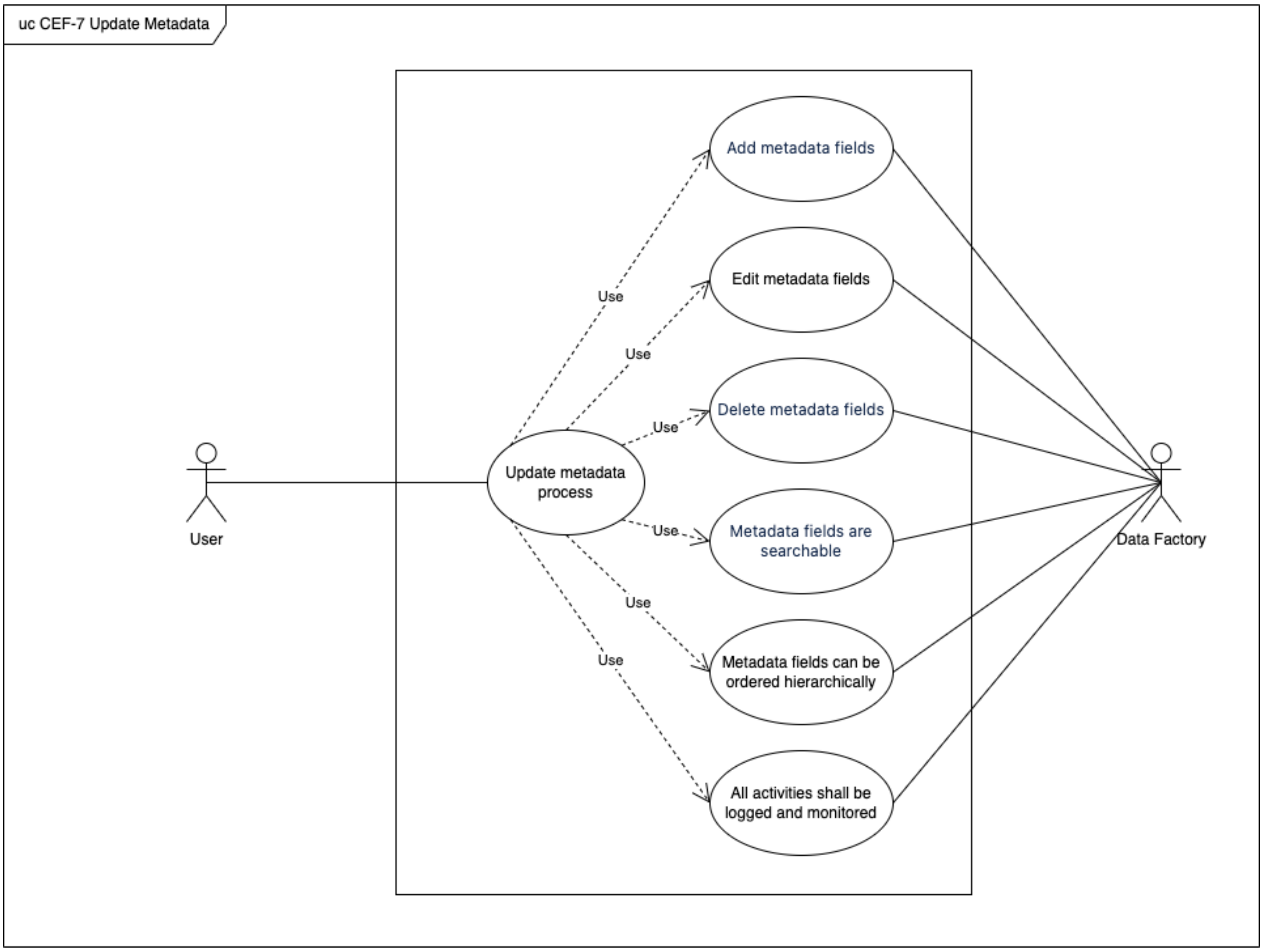


Figure 12. Use Case 7: Update metadata within the Pan-European Data Factory.

Requirements for Use Case 7: Update Metadata

Requirements	Description
Add metadata fields	To be able to describe data more precisely and provide further added value, additional characteristics can be added to data in the form of metadata.
Edit metadata fields	These characteristics in the form of metadata must also remain editable.
Delete metadata fields	It must also be possible to delete metadata that is no longer required
Metadata fields are searchable	Furthermore, it must be ensured that this additional metadata can also be found and recorded within a search.
Metadata fields can be ordered hierarchically	If a larger amount of metadata exists and in order not to lose track of it, the metadata must also be able to be ordered and edited hierarchically.
All activities shall be logged and monitored	Every activity or event, trigger, exchange and access must be logged and monitored

5.1.8 Use Case 8: Visualize Data

Displaying data in a data center is a key function for visually accessing required data. Among other things, dashboards are used for visualization, which can be created and also shared by the user. Within these dashboards, data including their annotations are displayed. Also possible is the display of metadata of each visualized sensor type, as well as the visualization of extended data and various reports. All activities can be recorded and monitored.

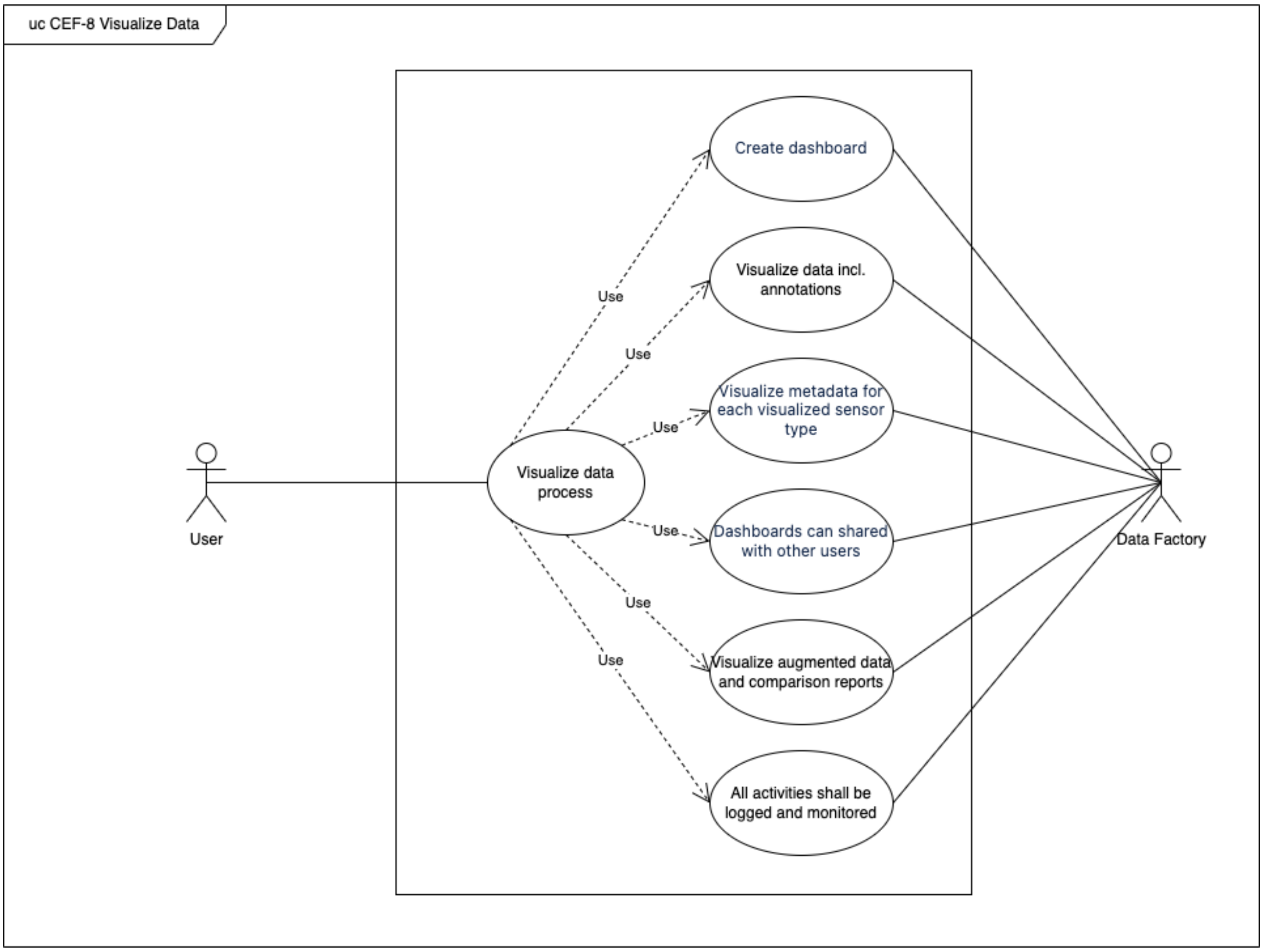

uc CEF-8 Visualize Data
User
Visualize data process
Use
Create dashboard
Visualize data incl. annotations
Visualize metadata for each visualized sensor type
Dashboards can shared with other users
Visualize augmented data and comparison reports
All activities shall be logged and monitored
Data Factory

Figure 13. Use Case 8: Visualize data within the Pan-European Data Factory.

Requirements for Use Case 8: Visualize Data

Requirements	Description
Create dashboard	To be able to display data and also larger amounts of visualizations, some kind of dashboards is needed. With this method, the data and results can be displayed.
Visualize data incl. annotations	In a dashboard it must be possible to display data including annotations associated with the data.
Visualize metadata for each visualized sensor type	All available metadata for a digital twin must also be viewable via a dashboard, as well as all metadata for a sensor image.
Dashboards can be shared with other users	Once created by a user, dashboards can be made available to other users as needed, ensuring efficient work and visualization across the whole Data Factory via the backbone network.
Visualize augmented data and comparison reports	Artificially generated data with its existing metadata, as well as various reports, can also be displayed within a dashboard.

All activities are logged and monitored	Every activity or event, trigger, exchange and access must be logged and monitored

5.1.9 Use Case 9: Billing Service

For correct and secure billing, all information on resource usage required by a user must be recorded and stored. It is also necessary to record and assign the user to tasks and jobs. Grouping users by tenants simplifies the display and provides the accounting department with a better overview. In order to be transparent to the users, all billing information is displayed, as well as a preliminary cost preview. At the end of the billing period, the billing statement is automatically generated and made available for download. All activities are recorded and monitored here as well.

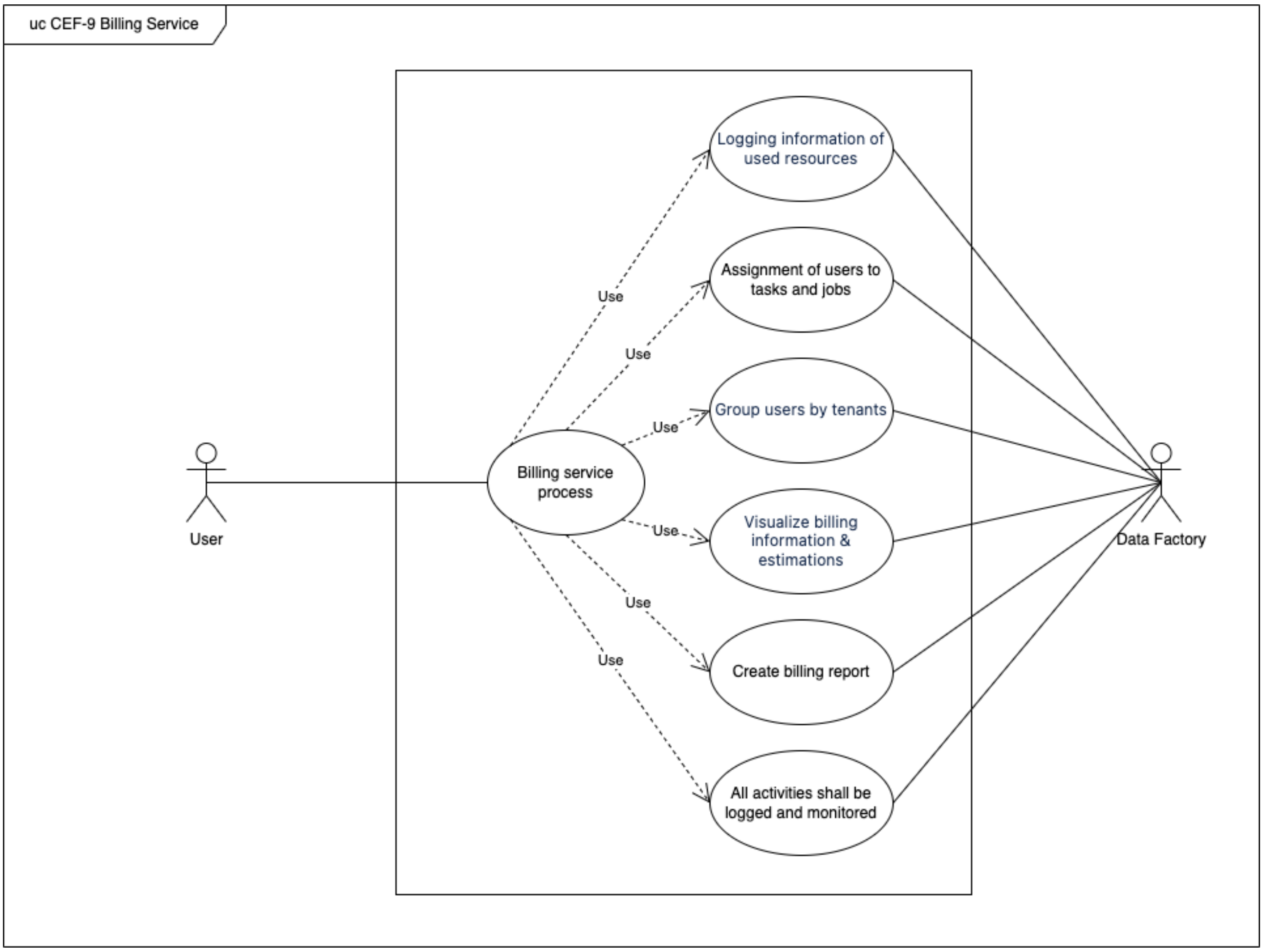

uc CEF-9 Billing Service
User
Billing service process
Use
Use
Use
Use
Use
Use
Logging information of used resources
Assignment of users to tasks and jobs
Group users by tenants
Visualize billing information & estimations
Create billing report
All activities shall be logged and monitored
Data Factory

Figure 14. Use Case 9: Billing service within the Pan-European Data Factory.

Requirements for Use Case 9: Billing Service

Requirements	Description

Logging information of used resources	For the billing service, it is extremely important that the information of all resources used is accurately recorded.
Assignment of users to tasks and jobs	As well it is also important to record all users and their exact assignments to tasks and jobs, so that every activity of a user is traceable and provable at any time.
Group users by tenants	In order to be able to bill and display not only user-dependent, but also across tenants, users must be able to be grouped by their tenant
Visualize billing information & estimations	It must be possible to see online at any time what has been consumed, what the breakdown of consumption is and what the forecast is up to the end of the billing date. The system must be able to provide information at any time about which user has used which resources and when.
Create billing report	If needed a billing report can be created and downloaded.
All activities shall be logged and monitored	Every activity or event, trigger, exchange and access must be logged and monitored

5.1.10 Use Case 10: Mark Data

Because of different reasons and for security it is necessary to mark data within the data centers as private or shareable. In order to do this it is important the user can set and configure permissions by his own, as well authorized users can manage the access within their tenants too, as well to create access to private data. In order to be able to restrict access to data for sensitive actions or certain services, corresponding functionalities are required, which are made available via the data center. Here, too, all activities are recorded and monitored.

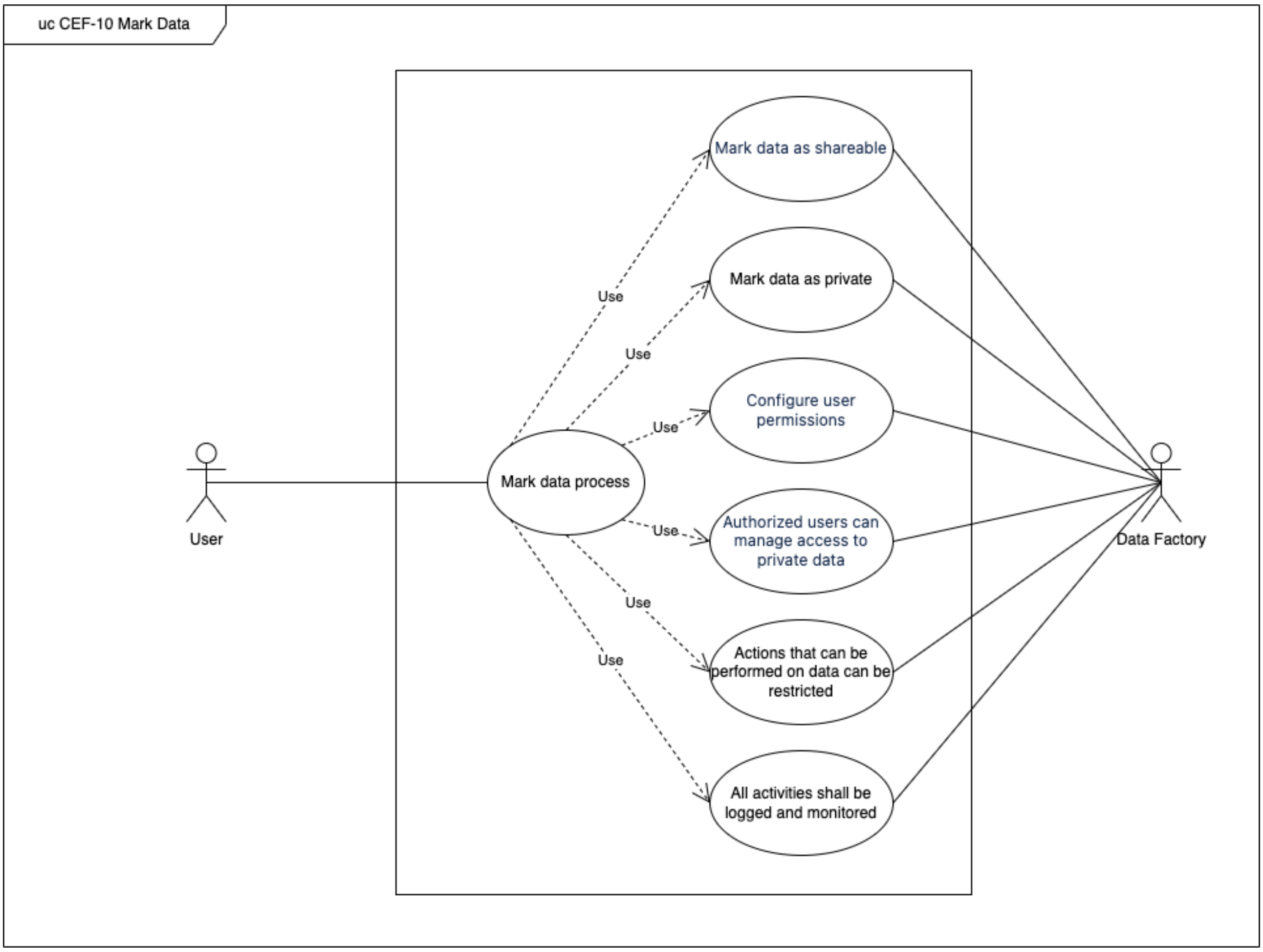

uc CEF-10 Mark Data
User
Mark data process
Use
Use
Use
Use
Use
Use
Mark data as shareable
Mark data as private
Configure user permissions
Authorized users can manage access to private data
Actions that can be performed on data can be restricted
All activities shall be logged and monitored
Data Factory

Figure 15. Use Case 10: Mark data within the Pan-European Data Factory.

Requirements for Use Case 10: Mark Data

Requirements	Description
Mark data as shareable	In a data center, it should be possible to explicitly mark data for distribution
Mark data as private	In a data center it should be possible to explicitly mark data as private
Configure user permissions	As the owner of data, I would like to be able to assign my own permissions to the data for which I am authorized. This ensures more accurate processing of this data.
Authorized users can manage access to private data	Authorized users have the possibility to manage the access to private data.
Actions that can be performed on data can be restricted	In some cases it will be necessary not to allow certain actions on specific data, these must then be able to be restricted

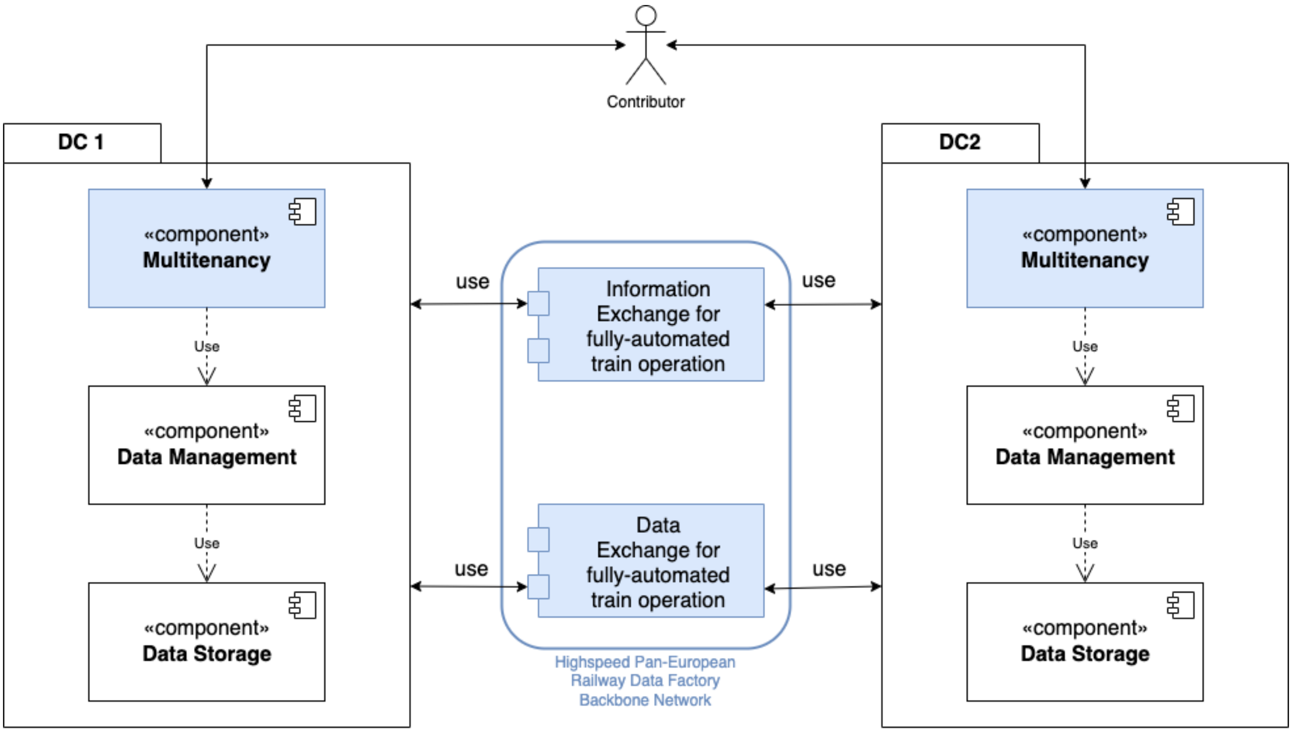
All activities shall be logged and monitored	Every activity or event, trigger, exchange and access must be logged and monitored

5.2 Technical Data Factory Use Cases and requirements

In this section, as mentioned earlier, the focus is on the technical use cases and their requirements. The technical use cases, which are particularly related to the pan-European Data Factory connectivity backbone and the data platform, are the focus of this study. To create a common, secure and efficiently functioning Pan-European Railway Data Factory, all participating facilities must be sufficiently connected, in terms of stability and high performance, so that they can exchange information and data with each other at any time.

Here, there is a separation between **information** (e.g., metadata, requests, reports, flaggings, etc.) and **data** (e.g., heavyweight sensor data), as the structures can differ significantly from each other, as well as the amounts of information or data. This leads to a different control and methodology in understanding and implementation.

In Figure 16, an overview on the technical use cases is given, before these are detailed in the subsequent sections.


Contributor
DC 1
DC2
«component» Multitenancy
Use
«component» Data Management
Use
«component» Data Storage
«component» Multitenancy
Use
«component» Data Management
Use
«component» Data Storage
use
use
use
use
Information Exchange for fully-automated train operation
Data Exchange for fully-automated train operation
Highspeed Pan-European Railway Data Factory Backbone Network

Note:
DC x can be a Data Center as well a Data Source

Figure 16. Overview on the identified technical use cases.

5.2.1 Use Case T1: Information exchange for fully-automated train operation

This use case describes the efficient transfer of information and notification from a distributed data source to another data source or data center. A distributed data source can be just a data source or as well another data center. In this context, distributed data sources mean an environment for storing and exchanging data on demand. The difference between a data source and a data center is that a data center will not only store and share data, but also provide tools and services which a data source will not support.

Efficient transfer is achieved if this is fully-automated as far as possible and the systems are coordinated with each other as far as possible, this applies to both hardware and software. The interoperability of all parties involved must be ensured. This is the only way to ensure secure and reliable information transmission. Formats, taxonomy and ontology must also be standardized and coordinated and must be checked before/during import and saving (Note: Which detailed implications this has on the information exchange between data centers is to be analysed).

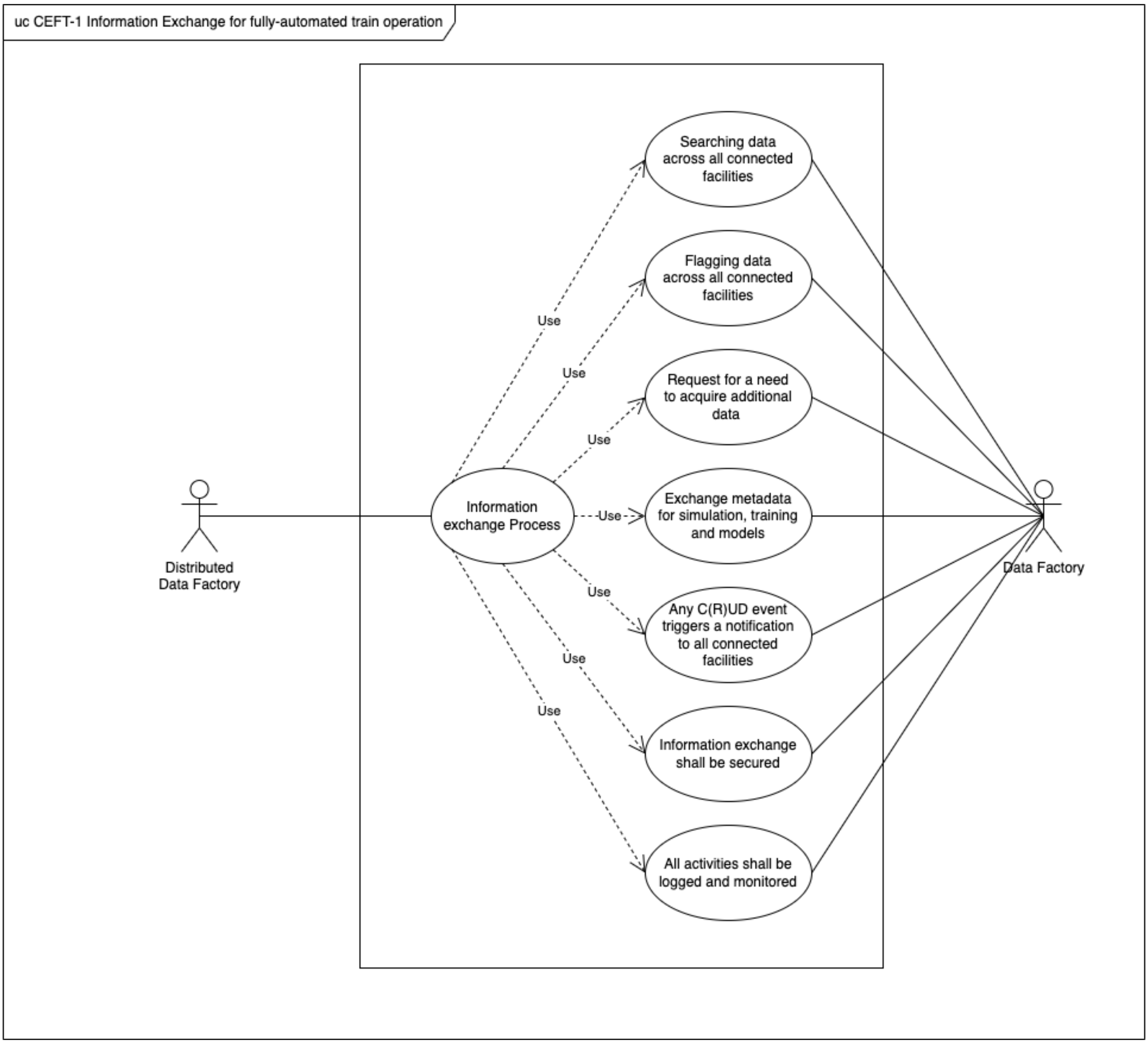

uc CEFT-1 Information Exchange for fully-automated train operation
Distributed Data Factory
Information exchange Process
Use
Searching data across all connected facilities
Flagging data across all connected facilities
Request for a need to acquire additional data
Exchange metadata for simulation, training and models
Any C(R)UD event triggers a notification to all connected facilities
Information exchange shall be secured
All activities shall be logged and monitored
Data Factory

Figure 17. Use Case T1: Technical information exchange within the Pan-European Data Factory.

Requirements for Use Case T1: Information Exchange for fully-automated train operation

Requirement	Description
Searching data across all connected facilities with a response within a few seconds.	When a user initiates a search query for information in one of the connected data centers, this query is automatically sent via the backbone network to all data centers connected to the Data Factory. The user shall obtain a response within a few seconds. Motivation: The capability to do a quick search for available data across facilities is considered essential for an efficient usage of the Pan-European Data Factory.
Flagging data across all connected facilities	When a dataset is defined in a data center and uses data originating from other data centers, these datas shall be flagged to indicate their usage.
Request for a need to acquire additional data	If the desired data is not listed in the catalogue the user may request the creation of these data.
Exchange metadata for simulation, training and models	In order to be able to build up a data catalogue, so that it is possible to identify which data is available and retrievable in which connected facility, the metadata must be exchanged for this data
Any C(R)UD event triggers a notification to all connected facilities	Meaning of (R) The (R) means readable and will not change any data, but it could be possible a READ is logged or triggers an event. CRUD means that data can be created, read, updated and/or deleted. As soon as this happens, the other data centers in the Data Factory must be informed so that it can be ensured that up-to-date processing is always guaranteed. This should also be automated within the backbone network.
Information exchange shall be secure	The exchange of information through the backbone network should always be secure, automated, high-performance and efficient.
All activities shall be logged and monitored	Every activity or event, trigger, exchange and access must be logged and monitored

5.2.2 Use Case T2: Data Exchange for fully-automated train Operation

This use case describes the efficient transfer of data from a distributed data source to a data center with shared access. Efficient transfer is achieved if this is automated as far as possible and the systems are coordinated with each other as far as possible, this applies to both hardware and software. The interoperability of all parties involved must be ensured. This is the only way to ensure secure and reliable data transmission. Formats, ontology and taxonomy must also be standardized and coordinated, and data consistency and integrity must be checked before/during import and saving. Also a takeover of data must ensure versioning is kept intact. Additionally common data quality standards must be fulfilled.

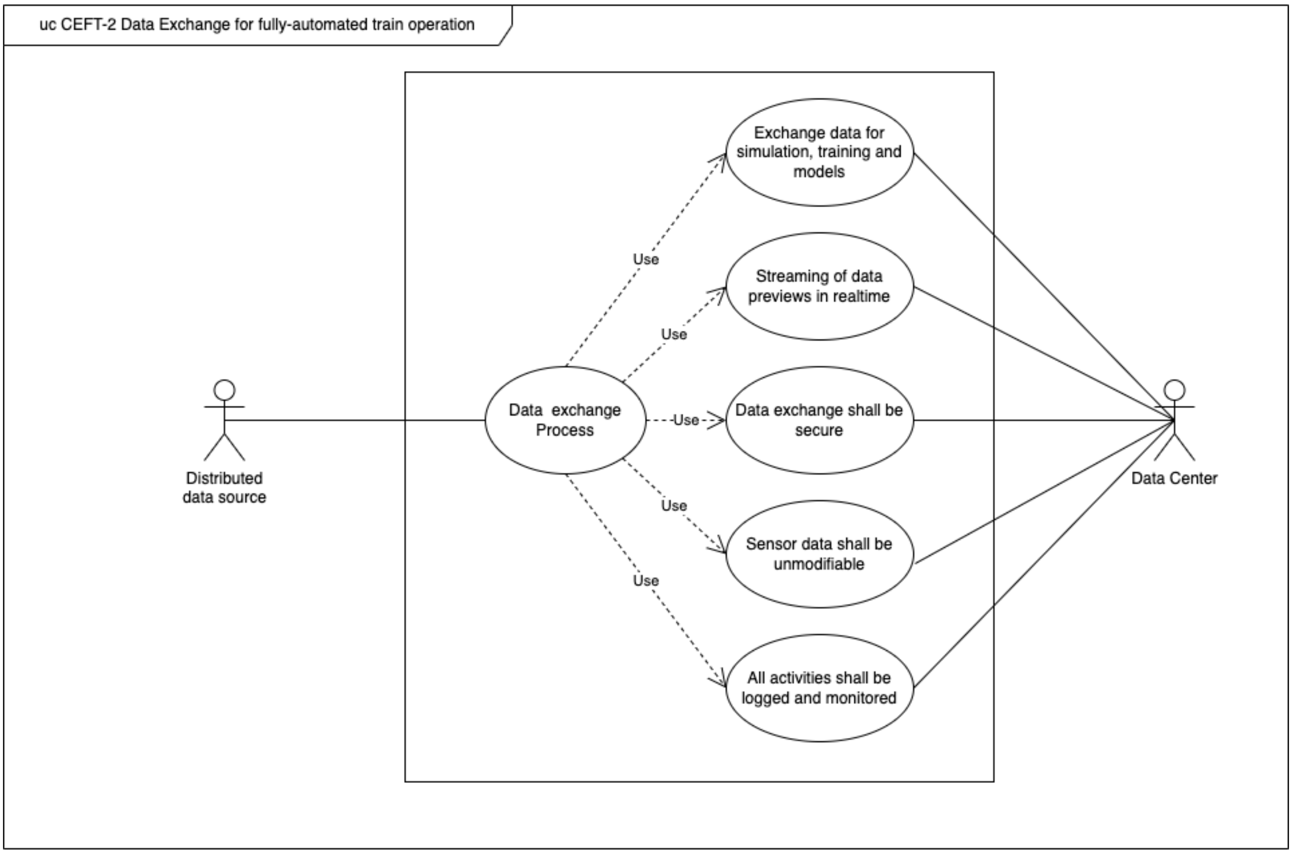

uc CEFT-2 Data Exchange for fully-automated train operation
Distributed data source
Data exchange Process
Use
Use
Use
Use
Use
Exchange data for simulation, training and models
Streaming of data previews in realtime
Data exchange shall be secure
Sensor data shall be unmodifiable
All activities shall be logged and monitored
Data Center

Figure 18. Use case T2: Technical data exchange within the Pan-European Data Factory.

Requirements for Use Case T2: Data Exchange for fully-automated train operation

Requirements	Description
Exchange data for simulation, training and models	It must also be possible to exchange large amounts of data (possibly on the order of 10s of Terabytes for one data set) between the connected facilities, for example to transfer sensor data streams or bigger simulation scenarios.
Streaming of data previews in real time	It must be possible to look at a preview of the data including annotations from the connected facilities in real time (i.e., with the preview starting maximum few seconds after the user's request, similar to the streaming experience on popular services such as YouTube) via the data catalogue
Data exchange shall be secure	Every data exchange must be secure, highly automated and performant

Sensor data shall be unmodifiable	Raw sensor data shall never be changed, unless they are deleted. Especially for ML Training it is important to train AI Models on unadulterated and unmodified data.
All activities shall be logged and monitored	Every activity or event, trigger, exchange and access must be logged and monitored

5.2.3 Use Case T3: Multitenancy

This use case describes the capability of a data factory to host multiple tenants. This allows for the sharing of services and resources in the data factory between multiple contributors. It is envisioned that this way a contributor who only provides partial services by themselves can book additionally required services.

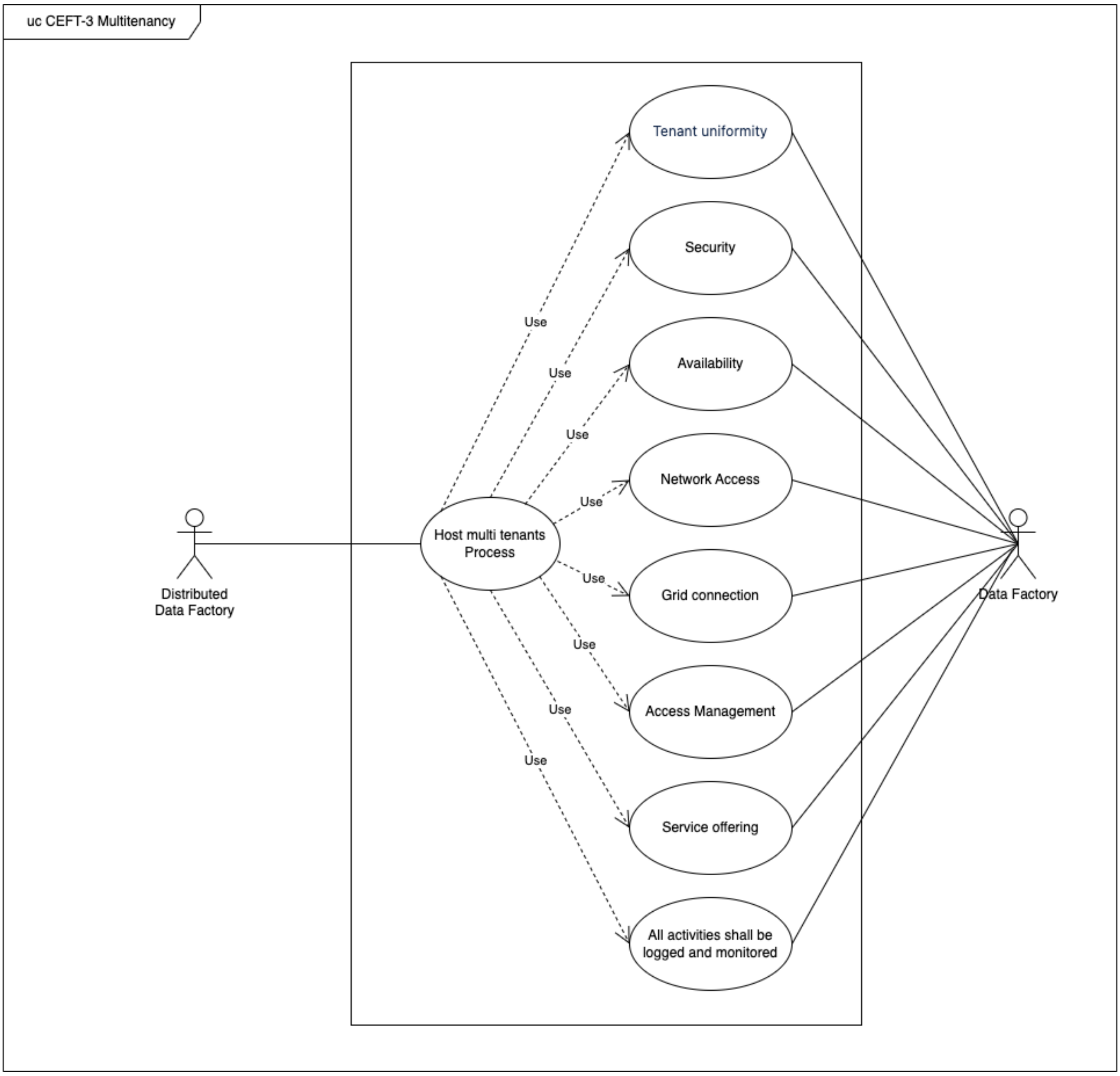


Figure 19. Use case T3: Multitenancy.

Requirements for Use Case T3: Multitenancy

Requirements	Description
Tenant uniformity	The systems and toolchain present in a data factory can be used by different tenants in a uniform way so that it is open to contributors whether they want to provide all services necessary for AI training and homologation or just a subset and then rent the rest of the services from another contributor.
Security	Tenants are isolated environments. This means that data and workloads inside of a tenant are concealed from other tenants data and workloads by default.

Availability	A tenant is available for its requester for the entire time it is required by the requester (or at least the time it is paid for).
Network Access	A tenant is accessible via the network for its respective users.
Grid connection	Each tenant in a data-factory is connected to the pan-European data-factory system as if they were an individual data-factory. *Note: Bookable resources and services must only be announced by the provider of the data-factory and not individual tenants within.*
Access Management	Each tenant owner can manage access to and access levels within their isolated environment.
Service offering	A data-factory can offer services via its multitenancy functionality that can securely be consumed by other contributors. This allows them to discover and book services from other contributors. This enables a contributor to scale when needed, make use of data-locality when they do a training mostly based on another contributors data or rely on another contributors services permanently.
All activities shall be logged and monitored	All activities are monitored and logged. Note: This is required to trace resource usage which is important for billing purposes, as well as a security requirement.

6 LEGAL, REGULATORY AND CYBER-SECURITY ASPECTS

In this section, legal, regulatory and Cyber-security related aspects are listed which have to be considered in the context of the pan-European Data Factory, and which will consequently be subject to analysis in the remainder of the CEF2 RailDataFactory study. A particular challenge is obviously that the Data Factory as an international infrastructure and ecosystem has to fulfil the superset of different laws and regulations as required in the involved countries. Due to the composition of the CEF2 RailDataFactory consortium with partners from Germany, France and the Netherlands, the subsequent points have initially been gathered from the perspective of the three countries, assuming these should be largely representative also for the remainder of Europe.

6.1 DEFINITIONS

The CEF2 RailDataFactory study in general uses the following definitions from the NIST Big Data Interoperability Framework [6].

Data Governance: refers to the overall management of the availability, usability, integrity and security of the data employed in an enterprise

Data Provider: Introduces new data or information feeds into the Big Data system. Makes data available to itself or to others. This maps to the definition of Data Provider in Section 4.

Data Application Provider: Executes a life cycle to meet security and privacy requirements as well as System Orchestrator-defined requirements

Data Framework Provider: Establishes a computing framework in which to execute certain transformation applications while protecting the privacy and integrity of the data

Data Consumer: Includes end users or other systems who use the results of the data application provider

Data Movement: Data movement is handled by transfer and application programming interface (API) technologies

Data Residency (according to Object Management Group - Data Residency Initiative): Data residency is the set of issues and practices related to the location of data and metadata, the movement of (meta)data across geographies and jurisdictions, and the protection of that (meta)data against unintended access and other location-related risks.

Data Privacy: The assured, proper, and consistent collection, processing, communication, use and disposition of data associated with personal information and Personally Identifiable Information (PII) throughout its life cycle

Data Security: Protecting data, information, and systems from unauthorized access, use, disclosure, 2097 disruption, modification, or destruction in order to provide: Integrity, Confidentiality and Availability

6.2 LAWS AND REGULATIONS

Laws and regulations that relate to data storage, processing and transfer may be fragmented and inconsistent across countries and regions.

In particular the following laws and regulations appear essential for the pan-European Data Factory and will hence be analyzed in detail in the remainder of the study:

- NIS2 Directive “Directive of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union” [7];
- GDPR (General Data Protection Regulation) [8];
- BSI IT-Grundschutz-Kompendium (Version: 2023) [9];
- BSI Act (Last update: June 2021) [10].

Data, including aggregate results delivered to data consumers must preserve **privacy**. In particular, data accessed by third parties or other entities (such as foreign railway operators) must follow legal regulations such as GDPR for the European market, and Personally Identifiable Information (PII) such as biometric data, tracking or location data, etc., must be handled and protected in accordance with this regulation.

Particular attention must be brought to **sensitive data** that can be classified under three main categories:

- **Personal data**. Specific examples in the context of the Data Factory would be video or other sensor imagery that would allow to identify individuals, or from which incorrect or inappropriate behavior of rail operations personnel could be derived;
- **Data subject to trade controls** (e.g., technology information, intellectual property);
- **Data subject to industry and government regulations** (e.g., military information). Examples in the context of the Data Factory could be sensor data through which details on military bases could be determined, or other information that could be critical when obtained by a hostile entity.

In this context, the pan-European Data Factory must enable **regular audits and assessments**

- to ensure that the data handling conforms to applicable legal, regulatory and policy requirements regarding privacy;
- to examine and evaluate protection and alternative processes for handling information to mitigate potential privacy risks.

Further, legal aspects such as **contractual enforcement** will have to be agreed between stakeholders (third-party data consumers, data providers, etc.) prior to cross-border data flows to maintain the privacy, security and compliance with the law in particular in the context of data moving and leaving its location / country of origin.

In the remainder of the study, working groups composed of engineering, compliance and legal experts will analyze the pan-European Data Factory in particular w.r.t. cross-organizational and cross-border data flow aspects.

6.3 CYBER-SECURITY

The following security standards have already been identified as relevant in the context of the pan-European Data Factory and will hence be the basis of the subsequent work in the study (noting that the list is non-exhaustive and will be continuously expanded):

- ISO/IEC 27001: Information Security Management System;
- ISO/IEC 27005: Information Security Risk Management;
- ISO/IEC 27017: Code of practice for protection of personally identifiable information (PII);
- NIST Cybersecurity Framework (CSF);
- NIST SP 1500: Big Data Interoperability Framework (volume 4 is related to data security & privacy);
- NIST SP 800-53: Security and Privacy Controls for Information Systems and Organizations;
- DB internal IT-Security Policies.

A special emphasis will be put on the following identified fields:

Risk and Accountability

Business risk level: An impact analysis / risk assessment shall be continuously performed to understand the consequence in the event of a loss, destruction, manipulation or theft of the data. As a result of this impact analysis, cyber security requirements (technical, organizational and procedural) can be determined. The following risks are likely specifically applicable when exchanging data across borders:

- violation of a government law or regulation leading to penalties;
- industrial spying by a foreign company or foreign government;
- increased risks of cyberattacks due to the interconnection and federation of locally managed data centers;
- delays in business transformation and technology modernization due to the multiplicity of data locations and disproportionate fears of non-compliance.

Accountability: The EU General Data Protection Regulation integrates accountability as a principle which requires that organizations put in place appropriate technical and organizational measures and be able to demonstrate what they did and its effectiveness when requested.

Authentication, Integrity and Confidentiality

Authentication and input validation: A mechanism must be employed to validate whether input data is coming from an authenticated source, such as digital signatures.

Integrity and confidentiality: Integrity and confidentiality of data must be enforced, for instance by using encryption (e.g., Transport Layer Security, TLS). One challenge to be solved around encryption techniques is related to the management of the encryption keys incl. distribution, revocation, renewal, recovery, etc.

Access Control and Security Controls

Access Control: There are multiple factors for access control, such as mandates, policies and laws that govern the access of data. Data classification governs the protection of the data and access should be granted only on a need-to-know/-use basis that is reviewed periodically in order to control the access. Visibility and traceability as to who is accessing the data is critical in protecting the data.

The following other security controls are to be addressed in the remainder of the study:

Incident Response, Physical security, Awareness training, Monitoring, Intrusion Detection, Audit...

7 SUMMARY AND NEXT STEPS

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

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

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

REFERENCES

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

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

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

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

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

[6] NIST SP 1500

[7] (EU) 2022/2555 NIS 2 Directive

[8] (EU) 2016/679 General Data Protection Regulation

[9] BSI (Bundesamt für Sicherheit in der Informationstechnik)

[10] BSI Act (Last update: June 2021)

[11] 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