CEF2 RailDataFactory

D 1.1 – Use cases, their descriptions and operational scenarios

Version 1.3

Due date of deliverable: 31/03/2023

Actual submission date: 17/05/2023

Resubmission date: 11/08/2023

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

Document status			
Revision		Date	Description
Referring to overall D 1 - Data Factory Concept, Use Cases and Requirements	0.1	09/03/2023	Document template generated
	0.2	24/03/2023	Major parts of content transferred from Confluence
	0.3	29/03/2023	First complete draft
	0.4	04/04/2023	Draft version submitted to advisory board
	0.5	11/04/2023	Use case and requirements sections merged
	0.6	19/04/2023	First review and commenting the advisory board comments
	0.7	24/04/2023	Final version after addressing of all advisory board comments, sent for final consortium approval
	1.0	28/04/2023	Version submitted to the project officer
	1.1	03/05/2023	Correction on formatting and spelling errors
	1.2	16/05/2023	Extract of deliverable D 1 created for D 1.1
	1.3	11/08/2023	Disclaimer updated based on the feedback of the granting authority

Project funded by the European Health and Digital Executive Agency, HADEA, under Connecting Europe Facilities Digital Grant Agreement 101095272

Dissemination Level		
PU	Public	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 deliverables D 1.1, D 1.2 and D 1.3 and input to this work!

Disclaimer

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do 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 can be held responsible for them.

Furthermore, 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 author(s) and project consortium do not 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 a first set of deliverables of the study comprising D 1.1, D 1.2 and D 1.3, the high-level vision of the pan-European Data Factory is introduced, key operational scenarios and use cases are defined, and related requirements in particular on the pan-European Data Factory backbone network and data platforms are derived and complemented with legal, regulatory and Cyber-security related aspects to be considered. Altogether, these requirements serve as a basis for the further work in this study.

ABBREVIATIONS AND ACRONYMS

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

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

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

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

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

1.1 AIM AND SCOPE OF THE CEF2 RAILDATAFACTORY STUDY

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

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

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

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

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

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

1.2 DELINEATION FROM AND RELATION TO OTHER WORKS

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

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

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

The input from the TAURO project is, however, taken into consideration in particular in the derivation of use cases for the Data Factory, as covered in Chapter 4.

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

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

1.3 AIM AND STRUCTURE OF THIS DELIVERABLE

This current document represents deliverable D1.1 of the CEF 2 RailDataFactory project, describing envisioned operational scenarios and use cases related to the pan-European Data Factory.

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

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

The remainder of this document is structured as follows:

- In Chapter 2, a representative operations scenario is depicted, both from the perspective of rail operations, and from a technical perspective;

- In Chapter 3, the envisioned contributor concept of the Data Factory is introduced, and roles are defined that are the basis for the subsequent use case descriptions;
- In Chapter 4, 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;
- Finally, in Chapter 5, this document is concluded with a summary and the expected next steps in the study.

2 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.

2.1 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.

2.2 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 1 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 1.

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.

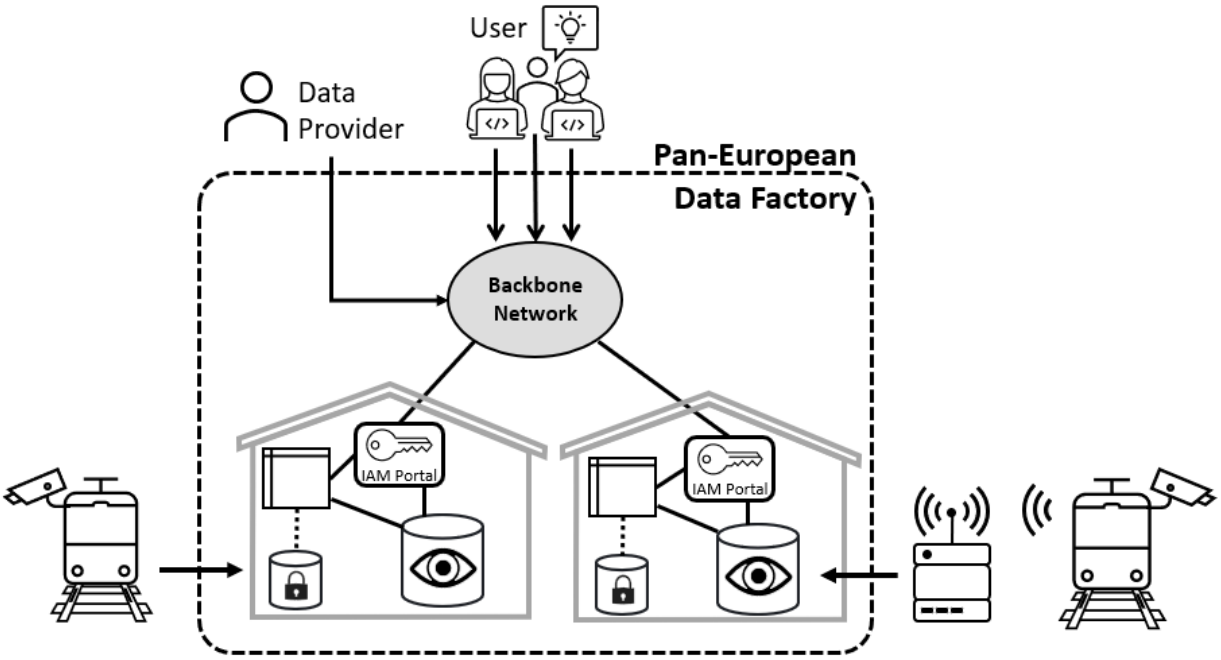

Figure 1. Representative operations scenario of the Data Factory (from technical perspective).

3 DATA FACTORY CONTRIBUTION CONCEPT AND ROLES

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

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

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

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

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

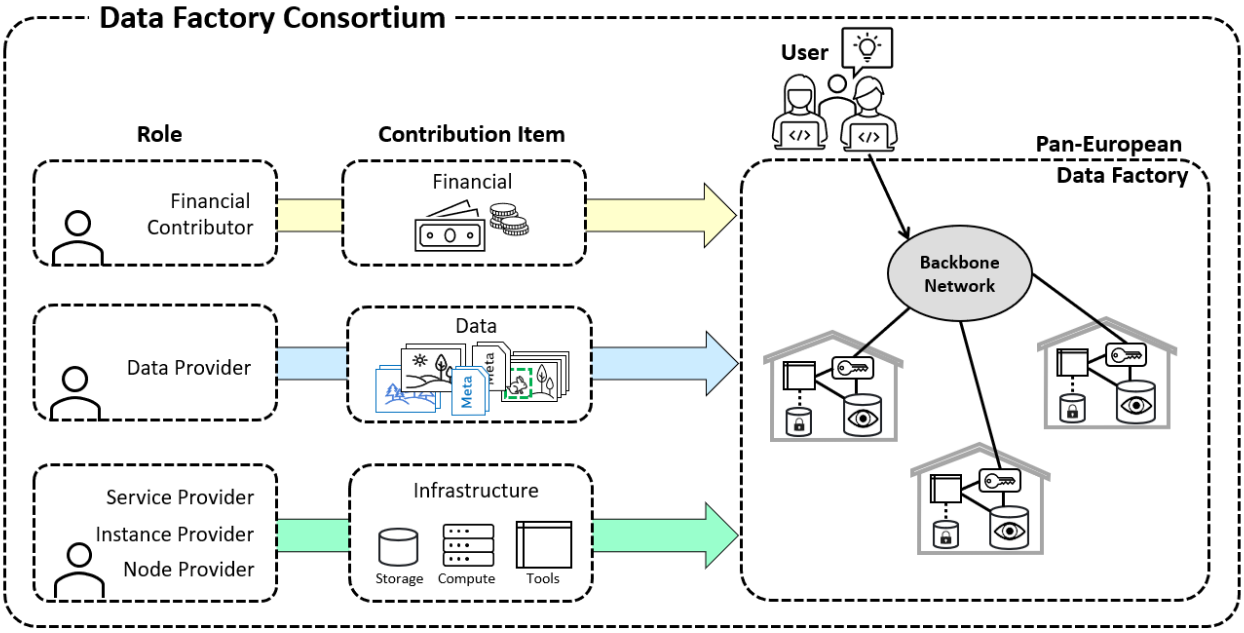

Figure 2. Illustration of the roles involved in the Pan-European Data Factory.

4 DATA FACTORY MAIN USE CASES

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 4.1 have been formulated from the perspective of the user or consumer, and the technical use cases in Section 4.2. address the the required pan-European backbone network and data platforms.

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

Figure 3 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**.

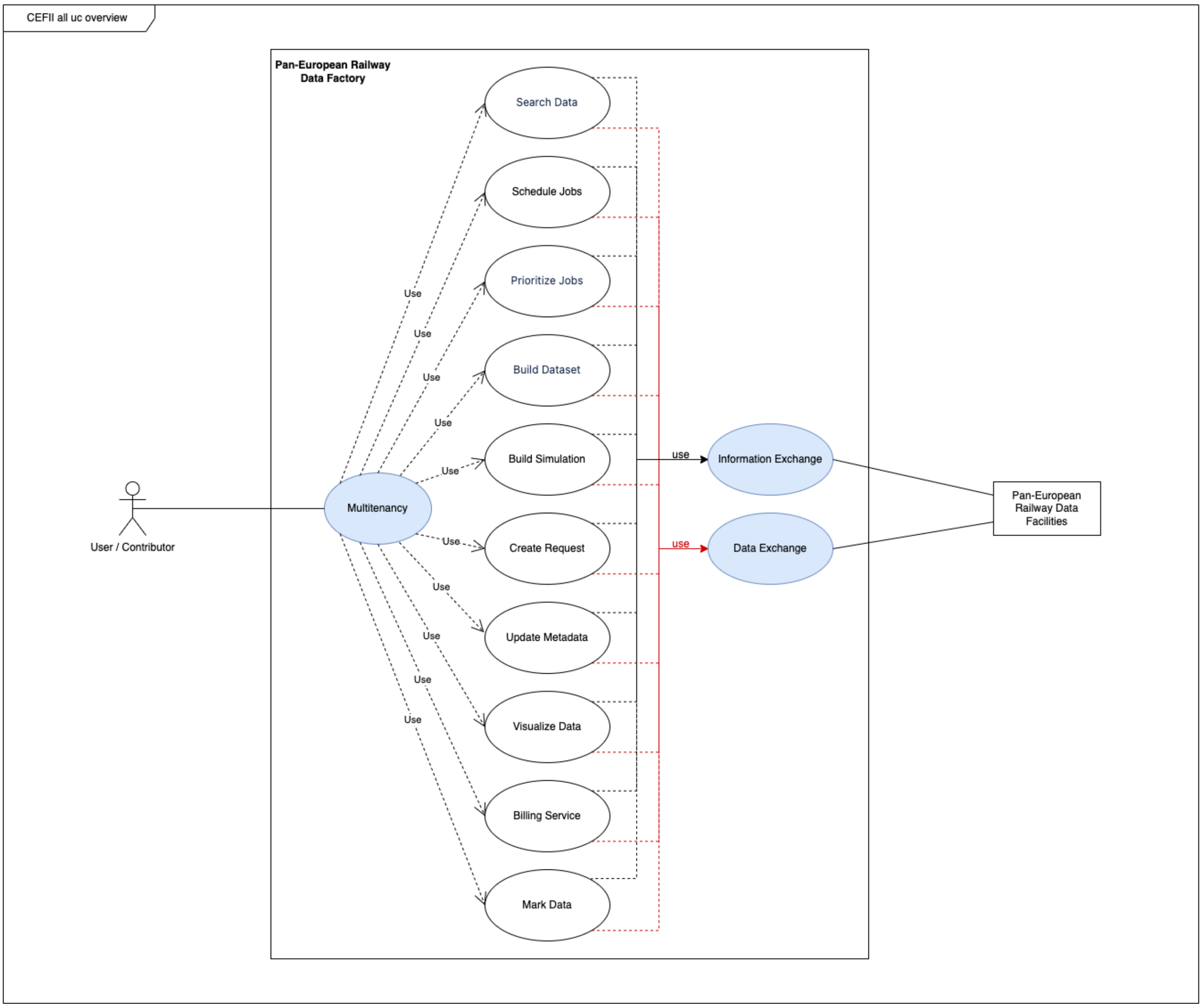

Figure 3. Overview of the general and technical use cases covered in this deliverable.

4.1 General Data Factory Use Cases

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

4.1.1 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.

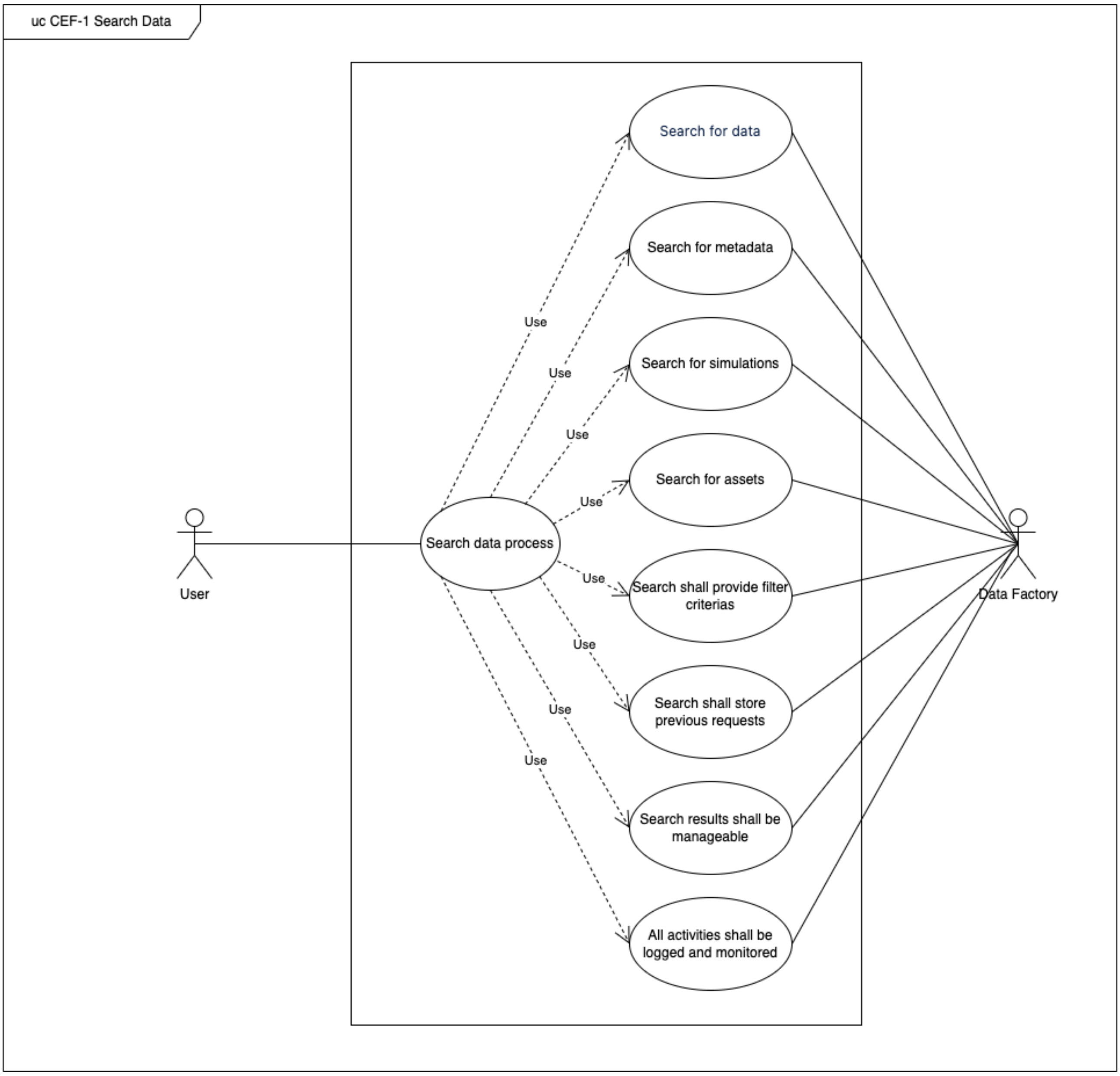
uc CEF-1 Search Data
Search for data
Search for metadata
Search for simulations
Search for assets
Search shall provide filter criterias
Search shall store previous requests
Search results shall be manageable
All activities shall be logged and monitored
Search data process
Use
User
Data Factory

Figure 4. Use Case 1: Search data within the Pan-European Data Factory.

4.1.2 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.

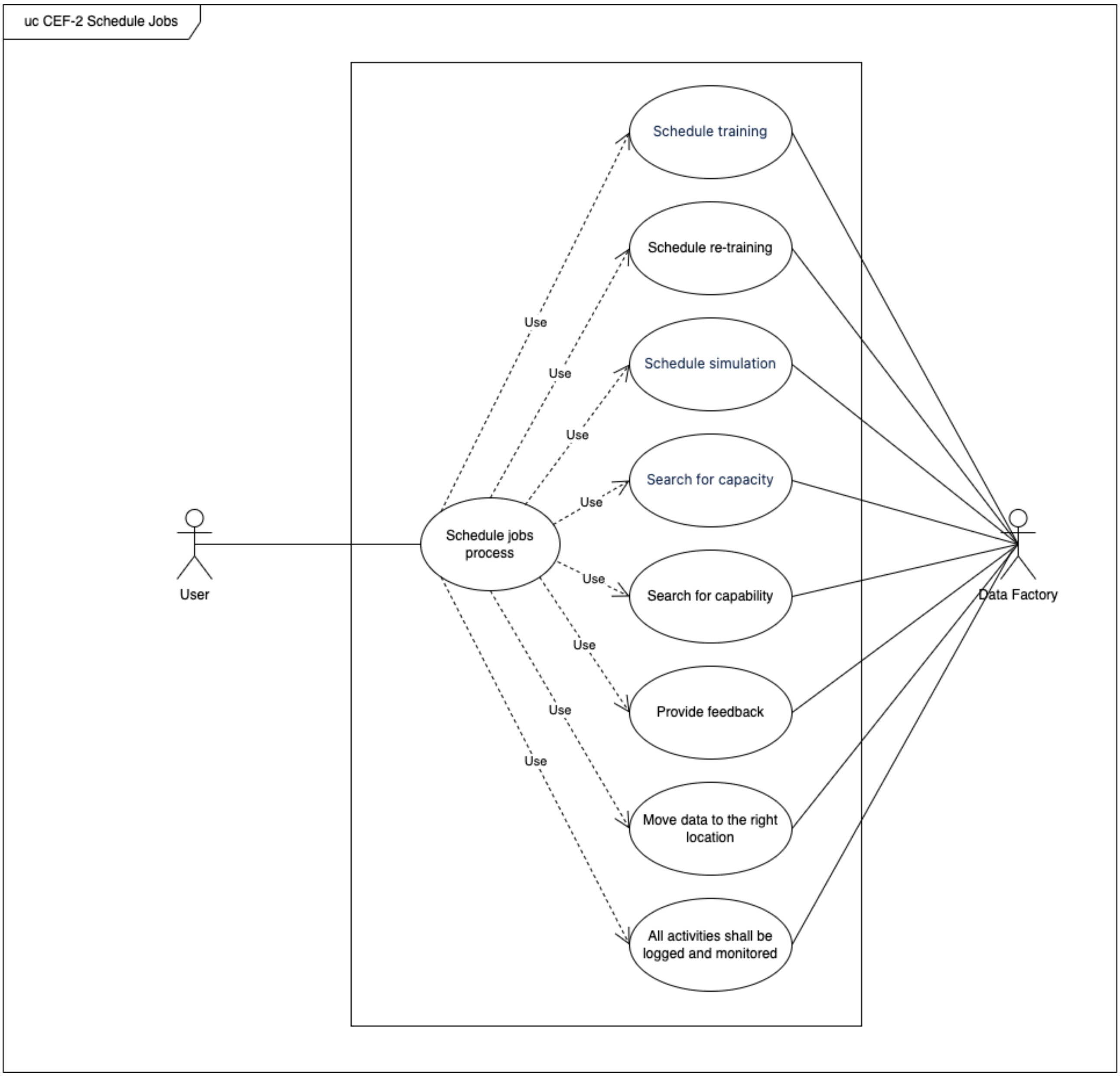

Figure 5. Use Case 2: Schedule jobs within the Pan-European Data Factory.

4.1.3 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.

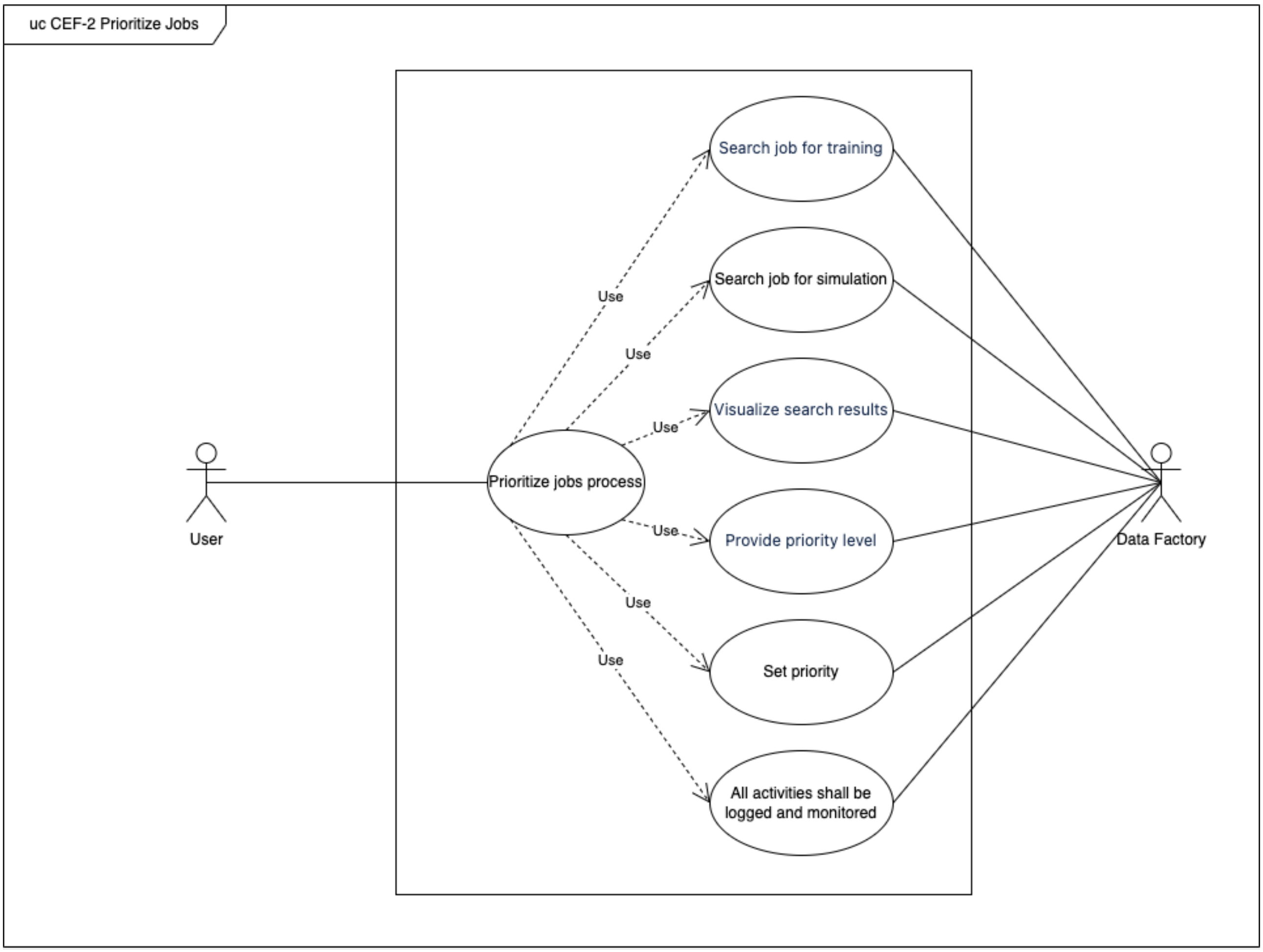

Figure 6. Use Case 3: Prioritize jobs within the pan-European Data Factory.

4.1.4 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.

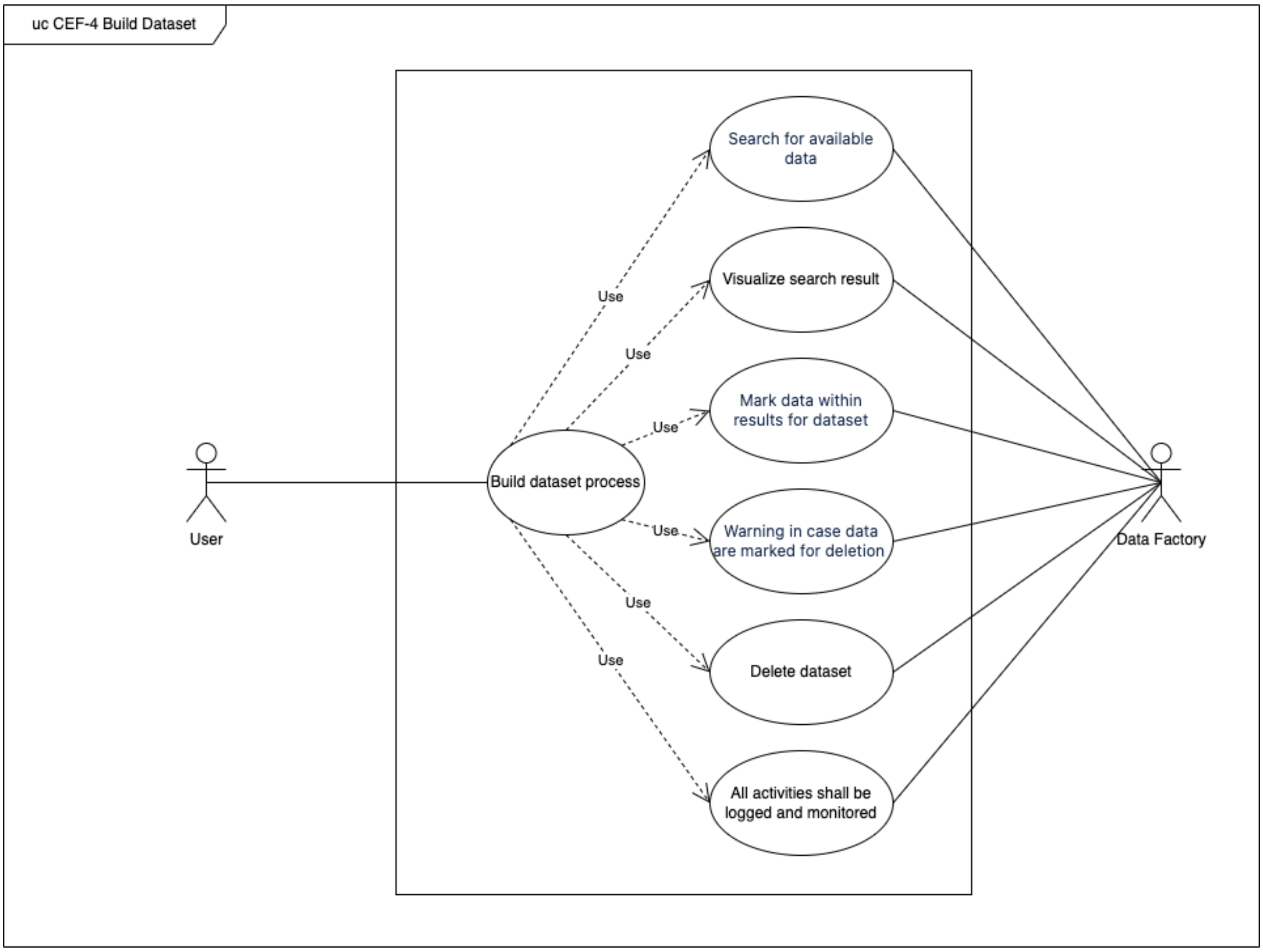
uc CEF-4 Build Dataset
Search for available data
Visualize search result
Mark data within results for dataset
Build dataset process
Warning in case data are marked for deletion
Delete dataset
All activities shall be logged and monitored
Use
Use
Use
Use
Use
Use
User
Data Factory

Figure 7. Use Case 4: Build dataset within the Pan-European Data Factory.

4.1.5 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.

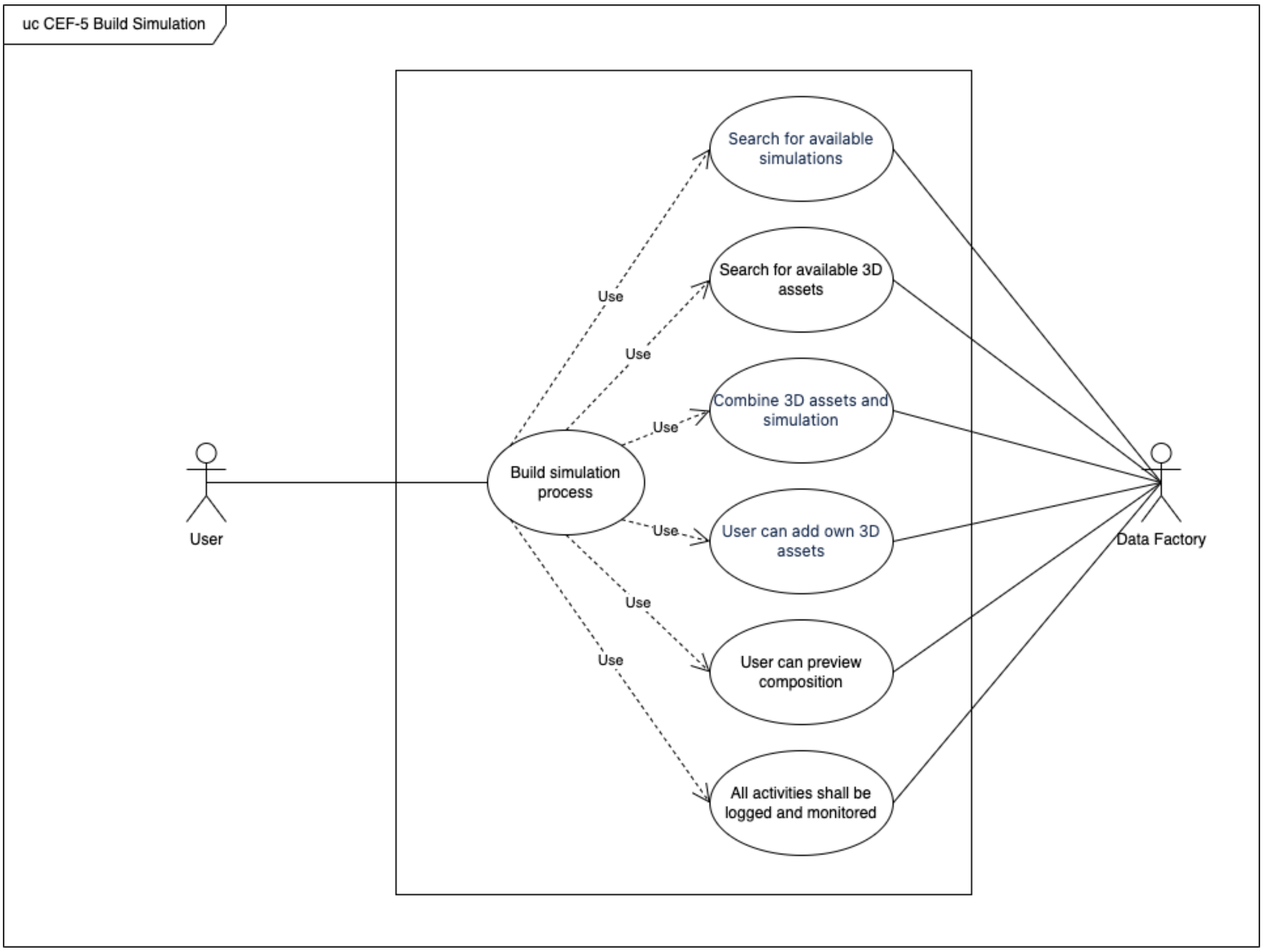
uc CEF-5 Build Simulation
User
Build simulation process
Use
Use
Use
Use
Use
Use
Search for available simulations
Search for available 3D assets
Combine 3D assets and simulation
User can add own 3D assets
User can preview composition
All activities shall be logged and monitored
Data Factory

Figure 8. Use Case 5: Build simulation within the Pan-European Data Factory.

4.1.6 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.

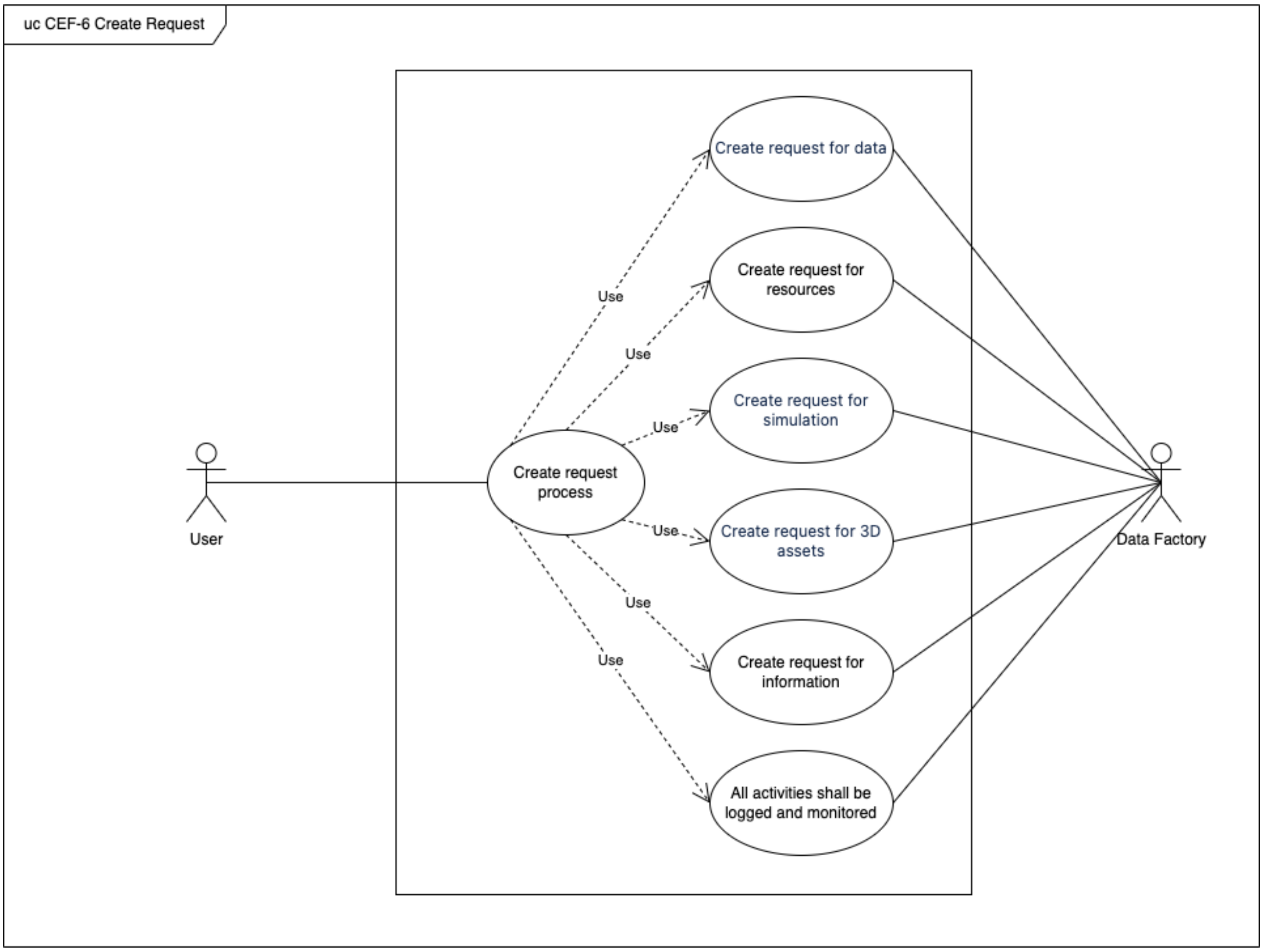
uc CEF-6 Create Request
User
Create request process
Use
Create request for data
Create request for resources
Create request for simulation
Create request for 3D assets
Create request for information
All activities shall be logged and monitored
Data Factory

Figure 9. Use Case 6: Create request within the Pan-European Data Factory.

4.1.7 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.

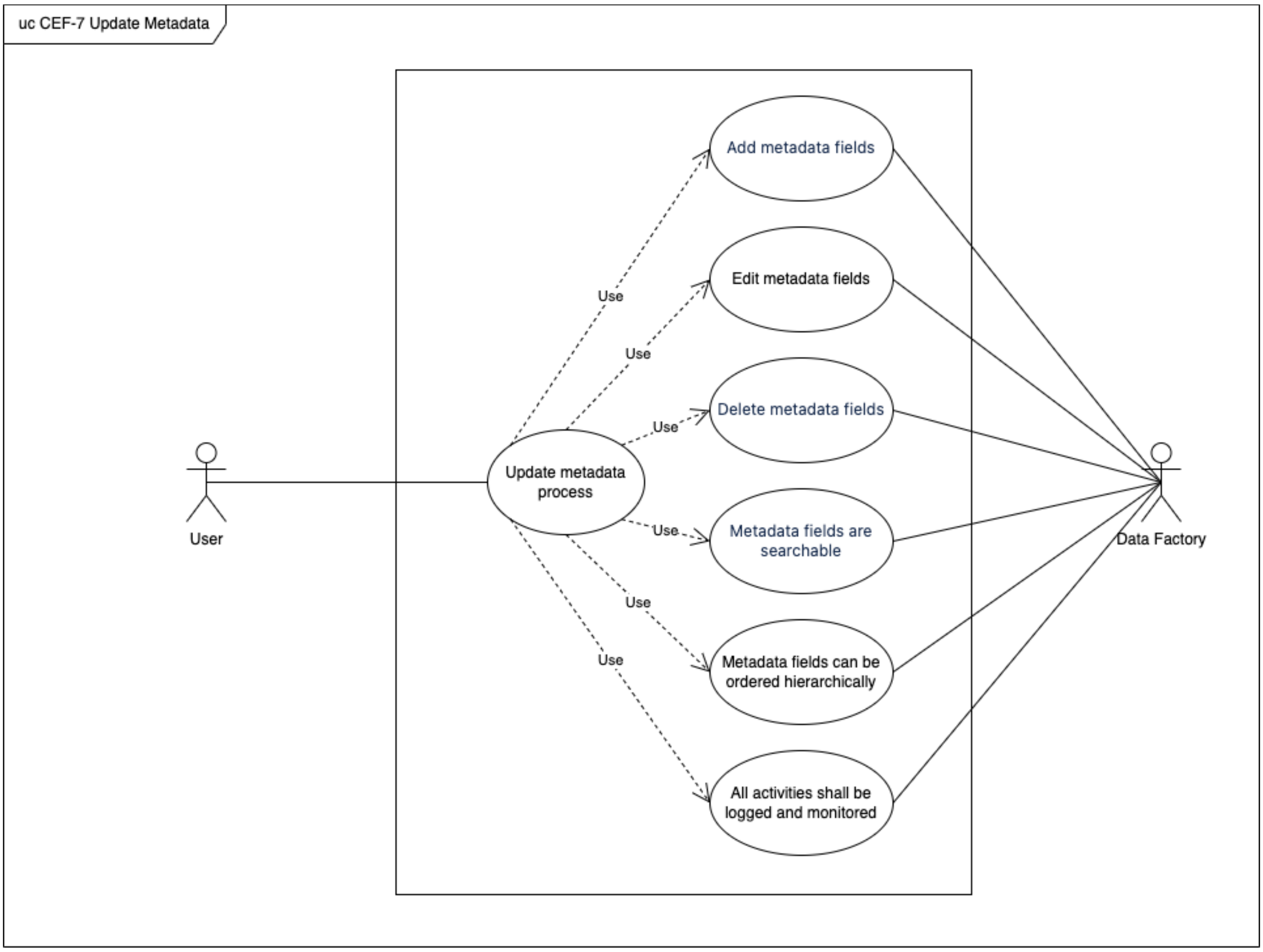

Figure 10. Use Case 7: Update metadata within the Pan-European Data Factory.

4.1.8 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.

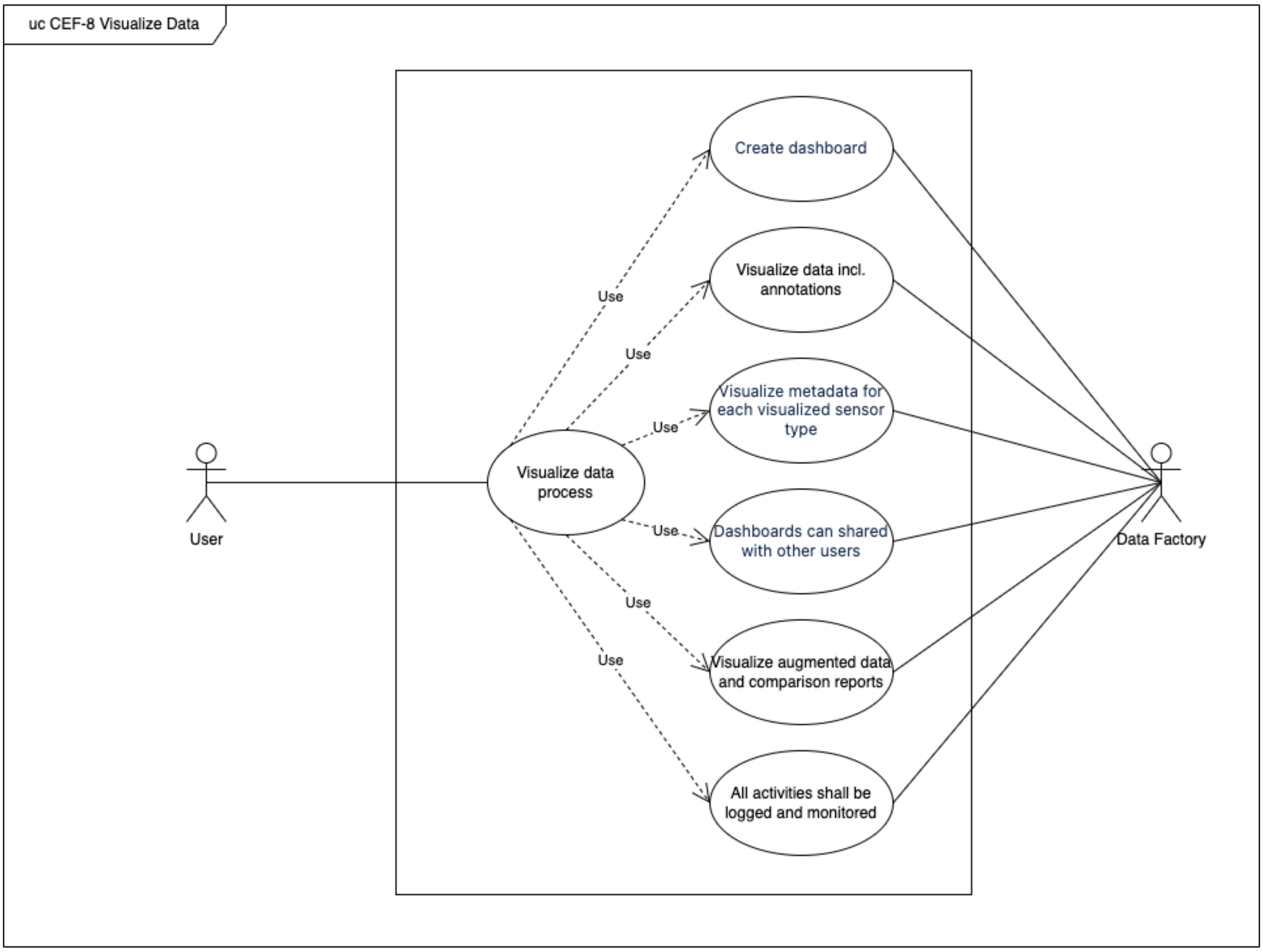
uc CEF-8 Visualize Data
User
Visualize data process
Use
Use
Use
Use
Use
Use
Create dashboard
Visualize data incl. annotations
Visualize metadata for each visualized sensor type
Dashboards can shared with other users
Visualize augmented data and comparison reports
All activities shall be logged and monitored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

Figure 11. Use Case 8: Visualize data within the Pan-European Data Factory.

4.1.9 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.

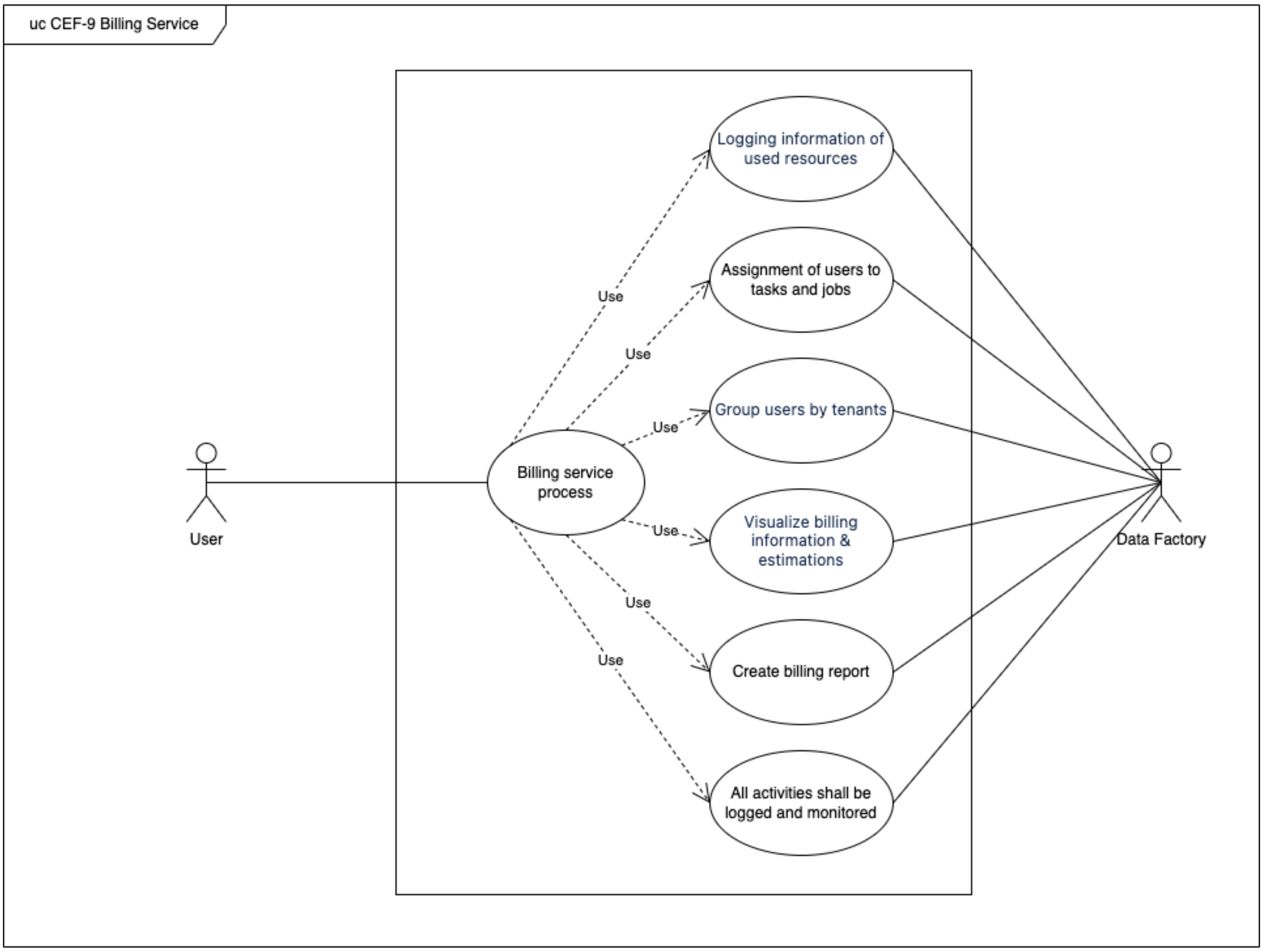
uc CEF-9 Billing Service
User
Billing service process
Use
Use
Use
Use
Use
Use
Logging information of used resources
Assignment of users to tasks and jobs
Group users by tenants
Visualize billing information & estimations
Create billing report
All activities shall be logged and monitored
Data Factory

Figure 12. Use Case 9: Billing service within the Pan-European Data Factory.

4.1.10 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.

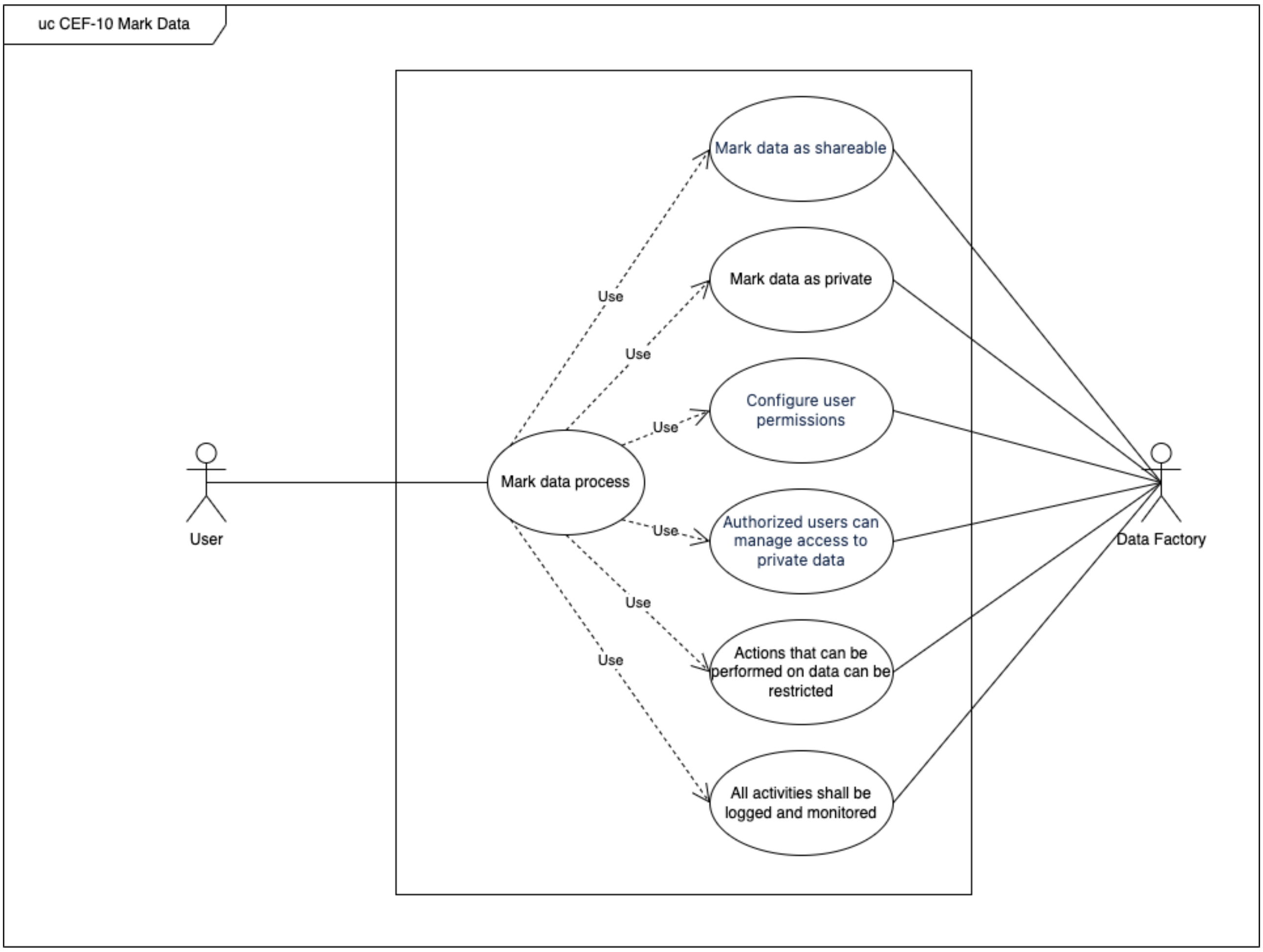
uc CEF-10 Mark Data
User
Mark data process
Use
Use
Use
Use
Use
Use
Mark data as shareable
Mark data as private
Configure user permissions
Authorized users can manage access to private data
Actions that can be performed on data can be restricted
All activities shall be logged and monitored
Data Factory

Figure 13. Use Case 10: Mark data within the Pan-European Data Factory.

4.2 Technical Data Factory Use Cases

In this section, as mentioned earlier, the focus is on the technical use cases. 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 14, an overview on the technical use cases is given, before these are detailed in the subsequent sections.

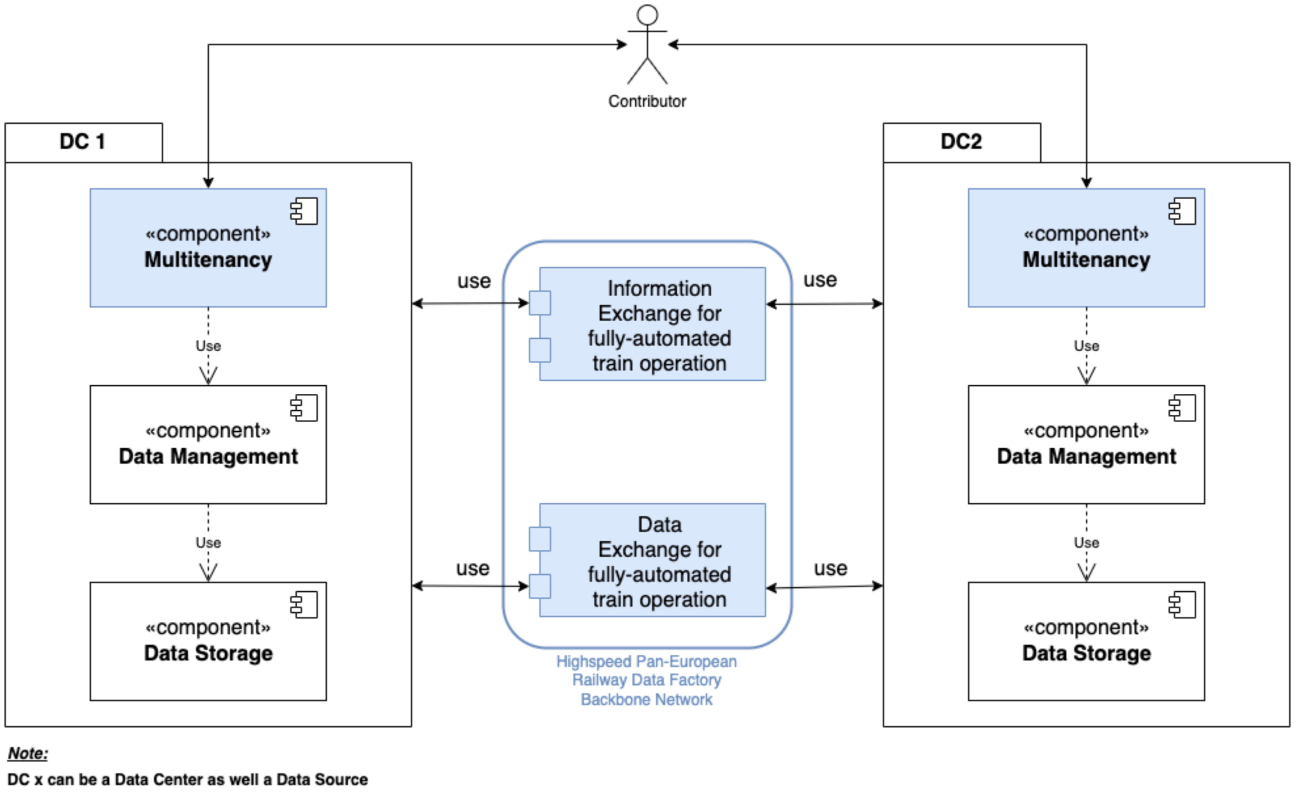
Contributor
DC 1
«component» Multitenancy
Use
«component» Data Management
Use
«component» Data Storage
use
Information Exchange for fully-automated train operation
use
use
Data Exchange for fully-automated train operation
use
Highspeed Pan-European Railway Data Factory Backbone Network
DC2
«component» Multitenancy
Use
«component» Data Management
Use
«component» Data Storage
Note:
DC x can be a Data Center as well a Data Source

Figure 14. Overview on the identified technical use cases.

4.2.1 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).

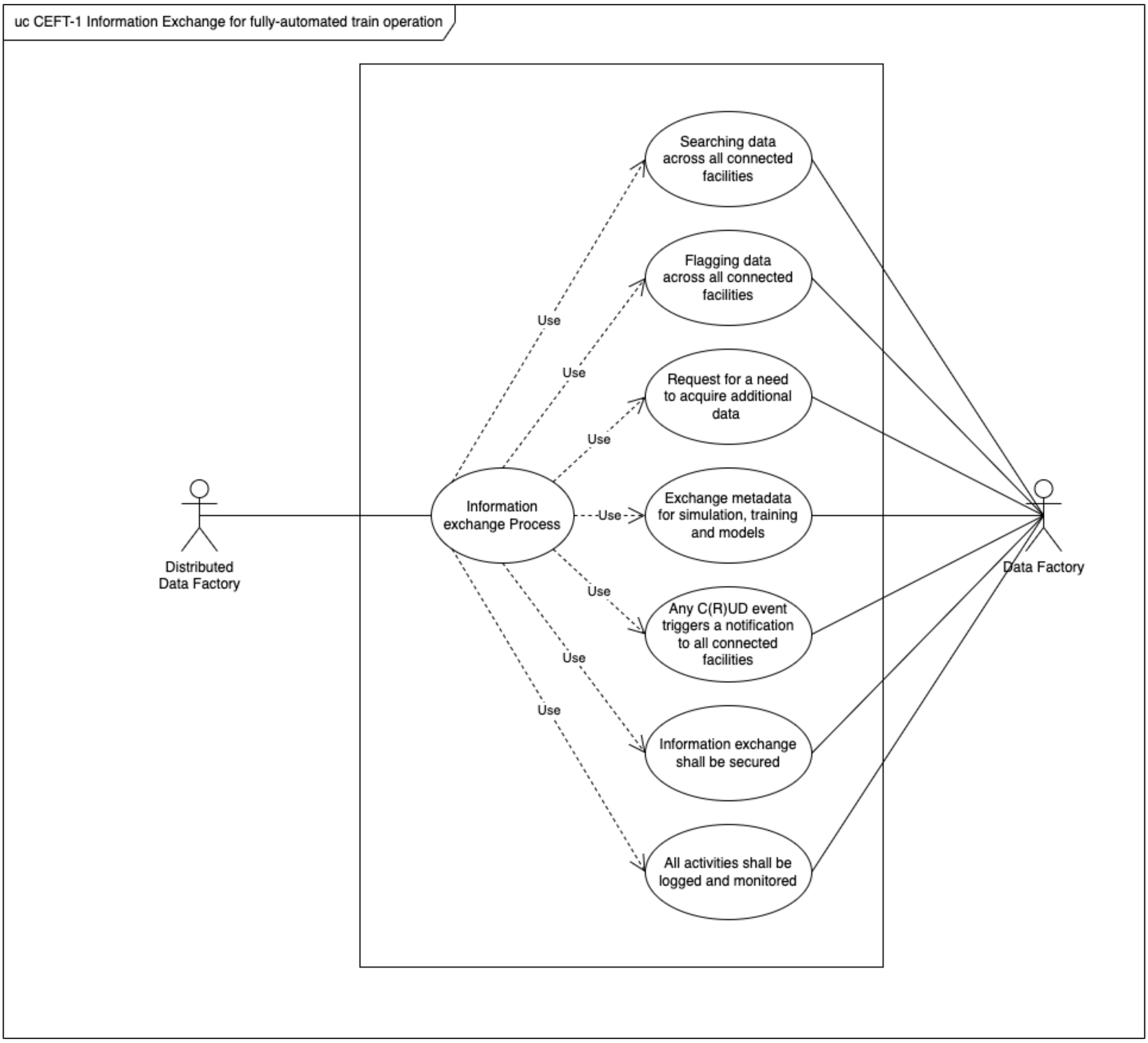
uc CEFT-1 Information Exchange for fully-automated train operation
Distributed Data Factory
Information exchange Process
Use
Searching data across all connected facilities
Flagging data across all connected facilities
Request for a need to acquire additional data
Exchange metadata for simulation, training 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 15. Use Case T1: Technical information exchange within the Pan-European Data Factory.

4.2.2 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.

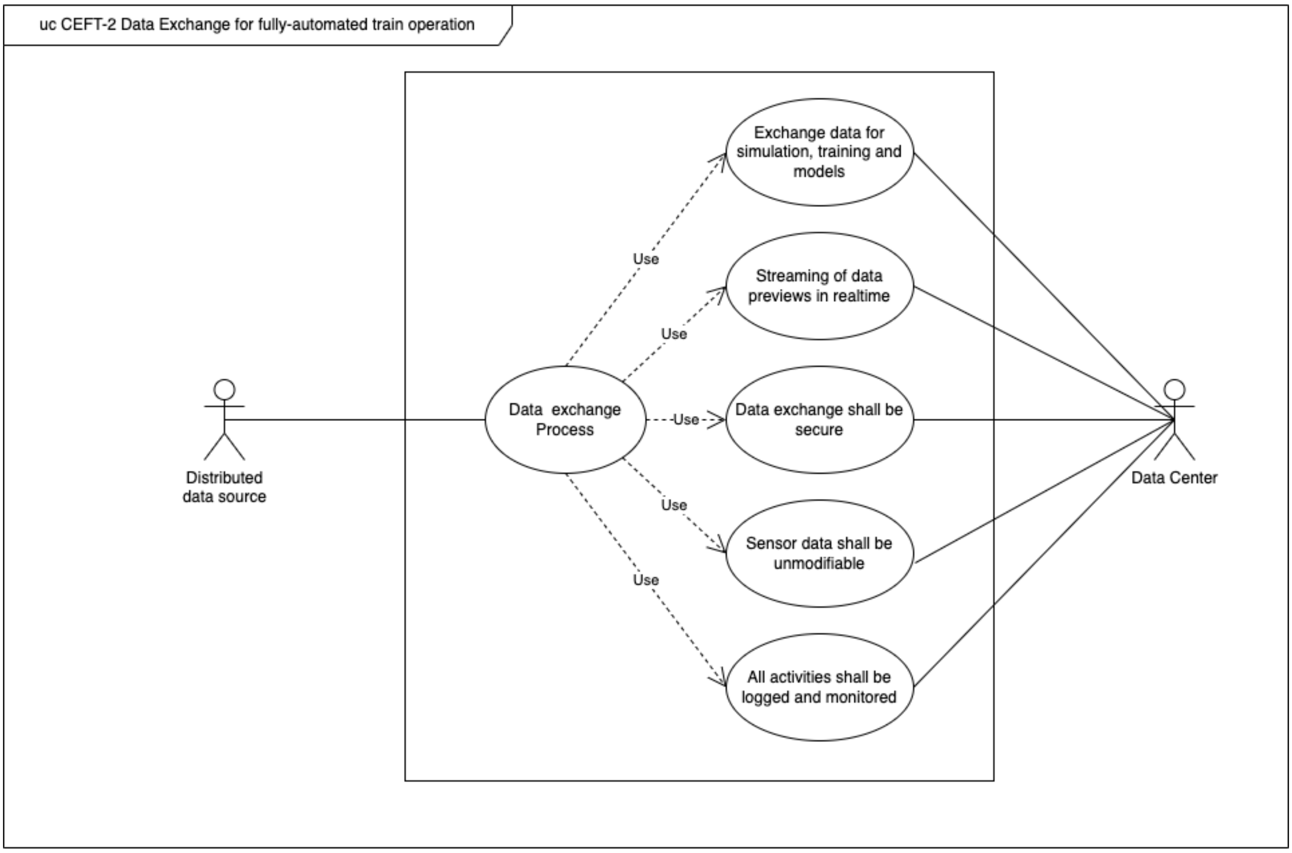
uc CEFT-2 Data Exchange for fully-automated train operation
Exchange data for simulation, training and models
Streaming of data previews in realtime
Data exchange shall be secure
Sensor data shall be unmodifiable
All activities shall be logged and monitored
Data exchange Process
Use
Use
Use
Use
Use
Distributed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

Figure 16. Use case T2: Technical data exchange within the Pan-European Data Factory.

4.2.3 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.

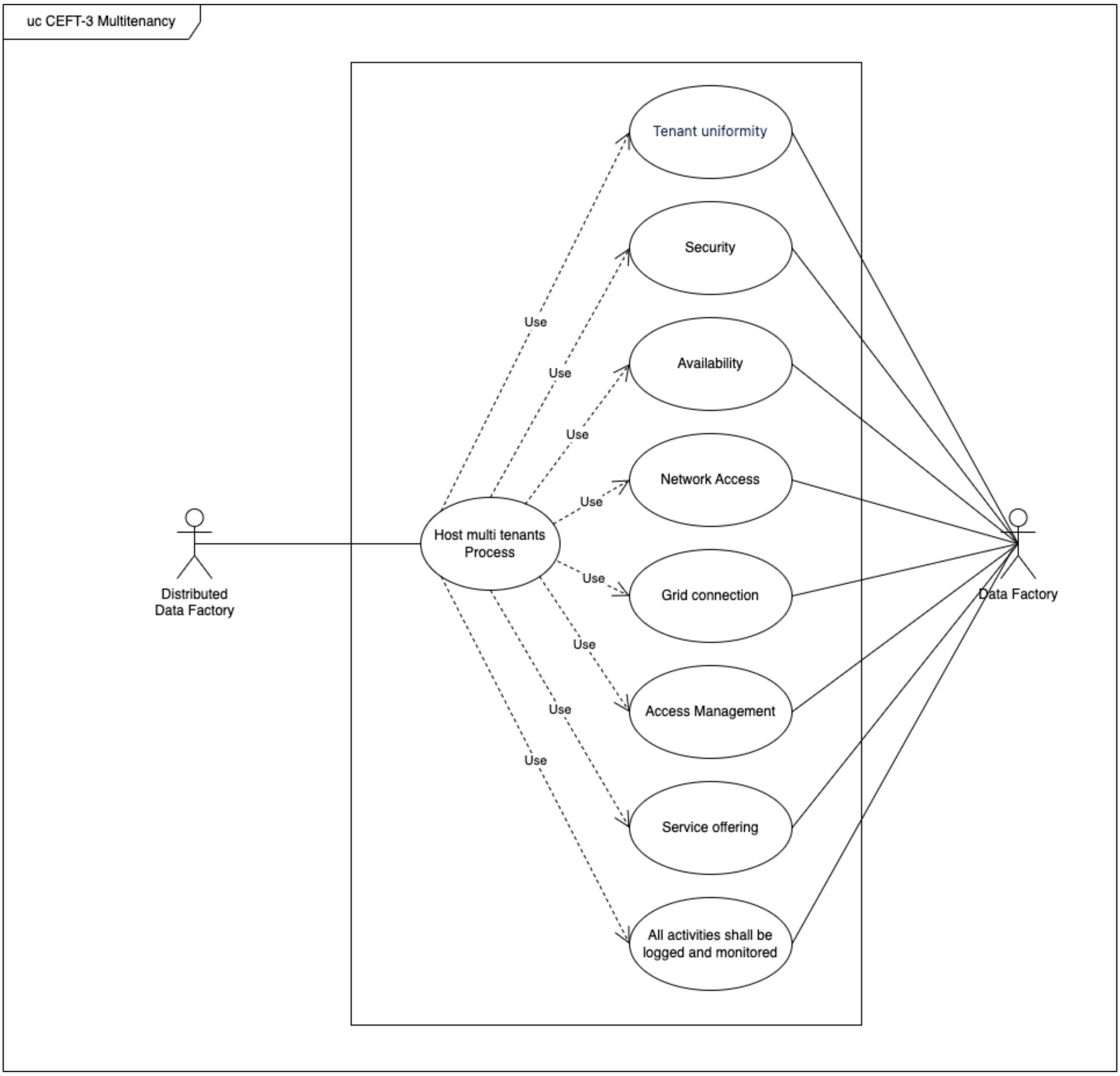

Figure 17. Use case T3: Multitenancy.

5 SUMMARY AND NEXT STEPS

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

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

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

REFERENCES

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

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

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

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

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

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