Grant Agreement Number: 101101962
Project Acronym: FP6 - FutuRe
Project title: Future of Regional Rail

DELIVERABLE D6.4
Requirements and definitions on Data Bases for Regional Lines (Alpha Release)

Project acronym:	FP6 - FutuRe
Starting date:	01/12/2022
Duration (in months):	48
Call (part) identifier:	Call: EU-RAIL JU Call Proposals 2022-01 (FUTURE-ER-JU-2022-01) Topic FUTURE-ER-JU-2022-FA6-01
Grant agreement no:	101101962
Grant Amendments:	NA
Due date of deliverable:	30-09-2024
Actual submission date:	14-10-2024
Coordinator:	Alessandro Mascis, Wabtec
Lead Beneficiary:	TRV
Version	1.0
Type:	Report
Sensitivity or Dissemination level[1]:	PU
Taxonomy/keywords:	Data mapping, railway demand, railway supply, regional railways

This project has received funding from the Europe's Rail Joint Undertaking (JU) under grant agreement 101101962. The JU receives support from the European Union's Horizon Europe research and innovation programme and the Europe's Rail JU members other than the Union.

[1] PU: Public; SEN: Sensitive, only for members of the consortium (including Commission Services)

Document history

Date	Name	Affiliation	Position/Project Role	Action/ Short Description
19/02/2024	Nils Olsson	NTNU Norway	Professor / task participant, Author	First Draft
15/07/2024	Pranjal Mandhaniya	NTNU Norway	Postdoctoral Fellow / task participant, Author	Second draft
18/07/2024	Matthias Walter	Hacon	WP co-leader, task participant	Review
18/07/2024	Fabian Hois	ÖBB	Task participant	Review
18/07/2024	Malcolm Lundgren	TRV	Task leader	Review
23/07/2024	Pranjal Mandhaniya	NTNU Norway	Postdoctoral Fellow / task participant	Final draft for internal review
08/08/2024	Ira Kataria	Hacon	WP leader, task participant	Review
14/08/2024	Rui Eirinha	GTSP	WP participant	Review
17/08/2024	Amirreza Tahamtan	ÖBB-INFRA	Reviewer within FP6	Review
18/08/2024	José Antonio Giménez Gómez,	Indra	Reviewer within FP6	Review
04/10/2024	Fabrizio Burro	FT	Quality Manager	Quality check and submission Final release
14/10/2024	Fabrizio Burro	FT	Quality Manager	Steering Committee review and submission

Disclaimer

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 author's view – the Europe's Rail Joint Undertaking is not responsible for any use that may be made of the information it contains. The users use the information at their sole risk and liability.

Table of contents

Executive Summary

This report encapsulates the experiences with data collection for Task 6.3 in Work Package 6 (WP6) of Flagship Project 6 FP6-FutuRe and is referred to as Deliverable 6.4. Task 6.3 involves the mapping of databases needed for various use cases within the broader context of WP6. WP6 involves specification of customer services in the context of regional rail, e.g. a multimodal travel solution including occupancy forecasts. The subsequent task corresponding to T6.3 is T11.3. The data mapping is targeted in T6.3 and the mapped data will be made available for the use case implementations in T11.3.

The data on demand and supply related to railway infrastructure is required for proper conduct of operations. For example, one should know the demand and supply statistics before operating new trains or constructing new lines between two locations. The main sources of data for the supply side of railways can be identified as transport operators, infrastructure managers, and ticket vendors. Meanwhile, the demand data can be acquired from the ticket sales and passenger counts. These sources are not limited, as data on demand and supply can be generated by simulation methods, as done by researchers.

Additionally, this report discusses the differences between static and dynamic data streams, focusing on their updating frequency, as well as the format and availability of these data streams. The need for data simulations and data validation is also discussed.

List of abbreviations, acronyms and definitions

Abbreviation / Acronym	Description
AI	Artificial Intelligence
API	Application Programming Interface
APC	Automated Passenger Counting
CCTV	Closed Circuit Television
DRT	Demand Responsive Transport
DSB	Danske Statsbaner
ERA	European Union Agency for Railways
FP	Flagship Project
GDPR	General Data Protection Regulation
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GTFS	General Transit Feed Specification
IM	Infrastructure Manager
MET	Meteorologisk institutt
ML	Machine Learning
NeTEx	Network Timetable Exchange
OJP	Open API for distributed journey planning
PRM	Passengers with Reduced Mobility
SSB	Statistisk Sentralbyrå
SCB	Statistiska Centralbyrån
TMS	Traffic Management System
UC	Use case(s)
WP	Work Package

List of figures

List of tables

1. Introduction

European railways are investing highly in high-speed trains and main lines. Main lines connect big cities, mostly the capitals of EU (European Union) countries. More funding for main lines creates a relative deficit towards the branch or regional lines. Regional lines cover less population density and, thus, are allotted with less budget. These definitions are relative as new construction and speed amendments can convert regional lines to commuter lines by changing population density and travel patterns. However, there is still the question of sidelining branch lines. Assuming the same state of infrastructure and rolling stock on these lines, one option to improve the quality of regional lines is by concentrating on passenger information systems (PIS).
High-quality information on current travel situations can only be provided to travellers if the underlying databases are sufficient and complete. This task will compile the requirements of such databases. Sufficient data is needed, for example, in developing AI-based systems. The goal is to provide travellers with the best possible information. Therefore, user feedback needs to be collected to modify the data supply. The data can be divided into three clusters:

- Data on regional railways: Task 6.3 is most concerned with this type of data. This may include train timetables, information of train demand and the capacity of trains.
- Data for first and last miles: In case of shared modes of transport involving regional rail station, the data for other connecting modes will be gathered.
- User feedback: Integration of user needs in the data supply models will be done via collected user feedback. A broad definition of user feedback is discussed in Section 6.3 of this report.

Relevant data and databases for these three clusters will be identified and mapped to use cases in WP6. The actual data will be collected and processed for further analysis. If data is presently not available or incomplete, simulated data may be generated.
Norwegian data has been chosen as a pilot case. It is an ambition for future work to expand data coverage for different countries, with a special focus on the regional applications demonstrated in other parts of WP6.

The report consists of 9 sections. Section 0 is the introduction to the report and what it conveys. Section 2 outlines different data streams for regional railways. Section 3 discusses the data sources on train traffic, while Section 4 talks about the data on railway infrastructure and rolling stock. Section 5 delves into other data streams like weather and demography that are helpful for PIS. Section 6 addresses the data requirement for UCs in WP6. Section 7 is the conclusion of the report. Lastly followed by a list of the cited references and sources and useful annexes that supplement the report.

2. Data on regional railways

One way to describe this data is by representing it from a supply and demand perspective. Supply data describes the transportation system and the transport opportunities offered to passengers and freight. Demand data illustrates the desired or actual use of the transport system and how the offered transport options for passengers and freight are used.

2.1. Data based on supply and demand

Supply data include data such as:

- Name of transport operators
- Network topology and routes/lines (topology)
- Timetables
- Connection links where interchanges may be made, default transfer times between modes (same or different) at interchanges
- Planned interchanges between guaranteed scheduled services
- Stop facilities access nodes (including platform information, help desks/information points, ticket booths, lifts, stairs, entrances and exit locations)
- Vehicle's capabilities (low floor, wheelchair accessible, etc.)
- Accessibility of access nodes, and paths within an interchange (such as existence of lifts, escalators, etc.)
- Existence of assistance services
- Real time data: disruptions, delays, access node status
- Fare products, tariff, price
- Services on stations

Supply data is typically obtained from different stakeholders in the transport systems, such as:

- Transport operators
- Infrastructure managers
- Transport agencies, public and private (typically public for public transportation, and private for freight)

Timetables and key information about a train service are widely available. Punctuality and delays are often mentioned as similar terms but have different meanings. Delays are measured in time units, whereas punctuality is expressed through percentages. One must define the threshold for when an arrival is counted as a delay to measure punctuality. These registration limits vary from country to country. It should be noted that timetable and delays are different types of data. Timetables can be considered static while delay data need to be real-time. Thus, timetables and delays can be separately reported as static and dynamic quantities.

On the other end of the spectrum, demand data includes:

- Ticket sales
- Passenger logs on public transport vehicles

- Logs of people observed in public transportation areas (entering and exiting stations, on platforms etc.)
- Freight volumes between destinations
- Trip planning request by passengers

Data from onboard sensors and ticketing systems are both managed by the public transportation providers. By contrast, surveys, payment statistics and mobile phone data may be available to stakeholders outside the public transportation system. This can be an advantage, as access to ridership data can be an issue for business reasons.
Train ridership is influenced by several factors, including fares, transit time, transit comfort characteristics and feeder accessibility of transit, price and service characteristics of the competing modes, seasonal variations, and monthly working day variations, as well as socioeconomic conditions of the service areas in the medium or long term.

Demand data is partly obtained from the same type of stakeholders as supply data but also includes transport models, general demographic data and short-term information like weather data and other specific information such as events. Transport models are a systematic representation of the complex real-world transport and land use (Australian Transport Assessment and Planning, u.d.) These transport models may involve planning train capacity among other processes.

Demand data can be based on revealed preferences showing actual travel and transport or stated preferences that illustrate desires or intentions of travel and transport. Revealed preferences are hard facts (although with several weaknesses), while stated preferences are hypothetical. The two can be cross-referenced for increased accuracy.
Demand data can be represented as origin-destination matrices, which map the travel patterns between different locations. These matrices often incorporate transport mode choices and are frequently based on models using available input data and algorithms for estimation and generalization.

2.2. Data based on update frequency

Data requirements can also be described based on the frequency of updates. The updating frequency is influenced by how frequently the data changes, how fast a change can be measured, and how the new value is distributed. The frequency of change is an important factor to consider. For example, supply type of data, such as railway infrastructure (line capacity, platform, station layout etc.), can remain constant for several years (especially for regional lines). At the other extreme, train delays change by the minute or even second and can be distributed to users continuously via screens at stations/in trains and in apps on personal devices. In between, there are features that may change annually or semi-annually, such as timetables, or more unpredictably, such as the rolling stock used for a particular train departure. Changes in rolling stock circulation plans can occur daily or even hourly.

Even though, this means that the updating frequency in practice is on a continuous scale for different features, it seems practical to distinguish between static and dynamic (real-time) data.

Static data can be uploaded in batches and does not necessarily need to be available online or through APIs. On the other hand, dynamic data should ideally be available in real-time through some sort of API or web interface.

Note that static data still need to be updated based on disruptions and regular intervals. For example, there have been changes in operation due to the collapse of a major bridge on the Dovre line that is the main connection for trains from South Norway to North Norway in August 2023 (Frazer Norwell, 2023). This disruption was eventually resolved in May 2024 after the bridge was reconstructed. This phenomenon shows how a user can get wrong information in case this disruption is not included in the Journey planning applications.

2.3. Data based on availability and readability

Requirements for data can also be described based on its accessibility. For example, some data are publicly available while others are under restricted access. The first idea is to know about the existence of the data, then it can be characterised by its level of accessibility. Data that can be accessed by anyone are the most prioritised. Some agencies are required to share these data. In Norway, for instance, examples include:

1. Historical and 24 hours forecast weather data (MET, 2024).
2. GTFS (General Transit Feed Specification) and NeTEx (Network and Timetable Exchange) data (ENTUR, 2024).

The access to data can also be associated with its readability to machine. Most standard data are in JSON and XML formats. However, valuable data may also be recorded in text message format or other formats that are difficult for machines to read, requiring human supervision. This matrix is illustrated in Figure 1.

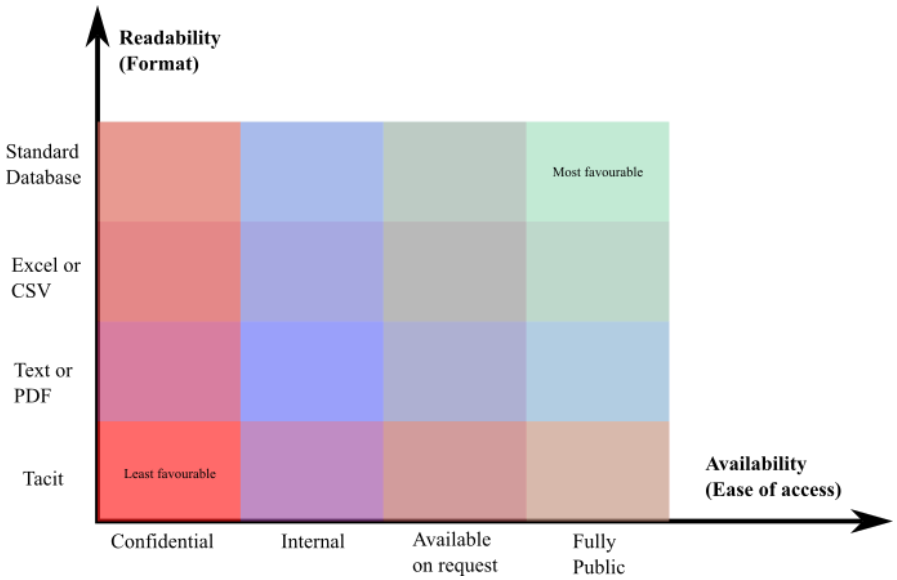

Figure 1. Data spectrum based on readability and availability

3. Data sources for train traffic

3.1. National Access Points

By passing the EU-Regulation 2017/1926 about the provision of EU-wide multimodal travel information services, each member state is obliged to provide a National Access Point (NAP). Regarding public transport, the NAPs are required to provide static travel data which includes:

- Transport operators
- Stop facilities access nodes (including platform information, help desks/information points, ticket booths, lifts/stairs, entrances and exit locations)
- Routes/lines (network topology)
- Timetables
- Planned interchanges between guaranteed scheduled services
- Hours of operation

This information is necessary to provide a multimodal trip planner. By the time of beginning of this project, almost all member states had set up a NAP (NAPCORE, 2024).
Usually, public transportation services and vehicle-sharing services are provided by many different providers. Each provider might independently submit data describing their own services. These providers share resources such as a train station. Services from different providers might be available at this train station. For example, the different bike-sharing services can operate at a railway station, but they might have different codes for locations and services.

If the different providers do not exchange information beforehand and agree on a common reference for the mentioned train station, each provider might use a slightly different name, a different geolocation and different sets of other attributes. For trip planning systems it is then difficult to recognize that all services share the same and identical train station. To solve these issues, standardisation and availability of data needs to be enforced.

3.2. Passenger counting and estimated demand

There are several ways to obtain data on train ridership. Manual technologies have been in practice since the beginning, and are still, surprisingly, the dominant method of gauging passenger count. Some of the ways to get the data on this matter are:

- Passenger counting at platforms.
- Ticketing data
- Manual observations or CCTV counting
- Automatic passenger counting in rolling stock
- Weight sensors on rolling stocks
- Mobile network data analysis

Some of these have been the main data source for the Norwegian railways. In addition, different travel behaviour surveys have been carried out. Fare collection systems are used for collecting passenger fares and controlling access to the transportation service. These systems can be used to track not only the number of passengers, but also the entry and exit points for travels. A closed-

loop fare collection (register access and exits) can provide information about the route and the time spent on the journey. An open loop (only register access) cannot register the route and the exit used, as it does not need to be identified. But, in both cases, the information has to be managed according to GDPR.

Automatic Passenger Counting (APC) is gaining popularity due to its accuracy and ease of instrumentation. An APC is an electronic device, which accurately records boarding and alighting data on transit vehicles such as trains. Sensors are in the doorways to a vehicle. When a person passes, the sensors count movements, and determine if they are entering or exiting the vehicle. APC is used frequently in research as well (Kuipers & Palmqvist, 2022).
Another technology for demand estimation is the use of CCTV and intelligent people counters to log numbers of travellers boarding and alighting from vehicles. On-board closed-circuit television (CCTV) are frequently installed on trains for surveillance and safety. This technology can also be applied for detection for people counting.

Number of passengers can be estimated based on electronic weighing equipment (EWE). EWE is installed in many modern trains because it supplies data for the braking system. This information can be used to estimate the number of passengers in the trains, as the weight of a train is a function of the number of passengers in the train at any time.

In 2013, testing of Automatic Passenger Counting (APC) from the German Dilax (Dilax, 2024) began in Norway on some trains. The APC registers the number of people that embark and disembark through each train door on every station, by means of sensors in the doorways. Logs from one of the lines of Norwegian Railways door counts are shown in Annex 1. The data shown in Annex 1 is just a representative sample of the actual door count data acquired for the progression of this task. The actual data consists of over 30 thousand entries with more than 30 columns (minimum 34 columns). These entries represent the APC records of each door on each train.

There has been an attempt to get mobile data to estimate the traveller count (Sørensen, et al., 2018). These types of data are regulated and requires to be purchased from telecom operators. Although these data are very reliable in what they perceive, they serve as a secondary source to get traveller demand.

3.3. Global Navigation Satellite System (GNSS) and Mapping services

Travel time and ridership can be detected using satellite feeds that include GPS (USA), Galileo (Europe) and other such service providers. Smartphones have built-in sensors, which can be used to extract movement profiles of commuters. These approaches can capture in detail individual travel behaviour but are limited by GDPR standards and may be costly to obtain from commercial actors that hold the data.

The satellite services can be used by navigation system providers. For example, TomTom Traffic Stats is a self-service product that gives access to what is claimed to be the largest historical road

traffic database, including road speeds, travel times and traffic density (TomTom, 2024). Examples of research based on TomTom data are numerous, including studies by (Hamner, 2010) and (Caban, 2021).

Apart from this, most journey planning applications require a cartographical mapping service for better planning and visualisation. Both commercial and open-source mapping service providers (MSPs) are available. Most common commercial mapping service is provided by Google Maps. On the other hand, the most common open-source mapping is provided by OpenStreetMap. There are different positive and negative aspects of both. Table 1 shows a list of different journey planning apps and their map bases. To clarify, the journey planning apps shown in Table 1 may not be unique to the location and vice versa. For example, AtB app used in Trøndelag region of Norway is best for short distance transport, while long distance trains and buses can be covered by Vy (via website and apps).

Table 1. Examples of different journey planning apps and their map bases

User Location	Name of the app	MSP
Trøndelag, Norway	AtB app	OSM (AtB, 2024) and Google (AtB, 2024)
Dublin, Ireland	TFI live app	OSM (TFI, 2024)
Vienna, Austria	WienMobil app	Google (Wiener Linien GmbH, 2024)
Austria	ÖBB Scotty app	OSM (ÖBB-Personenverkehr AG, 2024)
Sweden	SJ.SE app and SL app	Google (AB Storstockholms Lokaltrafik, 2024) and (SJ AB, 2024))
Cologne, Germany	KVB app	OSM (Kölner Verkehrs-Betriebe AG, 2024)
Cologne-Bonn region	VRS Auskunft app	OSM (Verkehrsverbund Rhein-Sieg GmbH, 2024)
Lisbon, Portugal	Carris app	Google (COMPANHIA CARRIS DE FERRO DE LISBOA, E.M., S.A., 2024)
Luxembourg	CFL mobile app	Google (CFL, 2024)
Various cities	Moovit app	OSM (Moovit, 2024)

The comparison in Table 1 gives a context on journey planning apps using commercial and open-source maps. Google and OSM are just examples of such. There are other commercial mapping bases such as Apple maps and similar other applications in open-source domain.

4. Data sources for infrastructure and rolling stock

4.1. Rolling Stock Register

According to Commission Regulation (EU) No 1300/2014, there should be a rolling stock register in each member state to provide detailed information regarding PRM capabilities (ERA, 2024). This information should follow the geometric specifications of entry points and availability of ramps for wheelchairs.

A summary of key information of Norwegian rolling stock is available online (NorskeTog, 2024). This summary can be created manually or possibly through web scraping. An example of such is shown in Annex 2. The table shown in Annex 2 is a representative sample of the rolling stock properties. The properties of a rolling stock are much more elaborate and can be openly viewed on (NorskeTog, 2024).

4.2. Infrastructure data - Network statement

According to RailNetEurope (2024) Article 27 of Directive 2012/34/EU (RailNetEurope, 2024) on establishing a single European railway area, each rail Infrastructure Manager (IM) shall publish a network statement. This shall describe infrastructure capacity and commercial, technical, and legal access conditions for existing and potential railway operators. Such network statements are a single source of up-to-date, relevant information in a transparent and non-discriminatory way. This means that network statements should be useful data sources, and they are. However, much of the information is available in non-structured formats, such as text on websites and PDF files. Network statements are updated yearly and are thus considered as static data.

For example, Belgium IM Infrabel has it in a tabular format (Infrabel, 2024). In Norway, the network statement is set-up by BaneNor (BaneNor, 2024). An example of such information is shown in Annex 3. Data can be collected manually or automatically (web scraping) extracting and summarising key data from the network statements. Data on station facilities, including those for Passengers with Reduced Mobility (PRM), are typically published in the network statement (An example is shown in Annex 3). While much of this information is available online in a relatively structured format, making it suitable for web scraping, it can also be summarized in other formats, such as an Excel sheet.

5. External factors and related data

There are several factors that may influence travel demand, including weather as a dynamic factor and demography as a relatively static factor.

5.1. Weather data

Weather data, which is significantly correlated with the demand, is widely available on websites and APIs. Weather data can be collected from the respective country's meteorological institute. For example, Norwegian Metrological Institute provides such services. Within their services, it is possible to extract data about temperature and wind (min/mean/max), snow (depth), and precipitation. By choosing a customized range of dates and specified weather stations, it is possible to collect data for most locations in Norway. These APIs can provide both historical data based on location and date range, as well as daily forecasts for up to 24 hours. The overall structure of the data flow is simplified in the form of APIs as shown in Annex 4.

5.2. Demographical data

As a proxy for the number of people at a station, the population of a municipality or other demographical data of higher resolution, may serve as an indicator of the number of inhabitants in the vicinity of a station or public transportation hub. It should be noted that this proxy may not be very accurate, for example, when considering stations in business or recreational areas, or those connected to airports. The theory is that the population of an area will, on average, be a decent proxy for the number of people at a train station in the same area. To find the number of people in an area, public demography statistics can be used (SSB, 2024), (SCB, 2024) as they have a lot of statistics on inhabitants in general. For some stations, the station name is the same as the municipality name, while for others, this is not the case. However, this can be an approach for data validation.

5.3. Databases of databases

Collections that compile various transport data streams are valuable resources for validating existing data and filling in gaps. They can cover websites with summaries of available data and links to more information. Some examples include:

1. Overview of Norwegian data streams: https://developer.entur.org/pages-real-time-intro, https://api.banenor.no/customer-info/realtime/v2.1/rest/et,
2. Overview of Swedish data streams: https://www.trafiklab.se/api/

Multiple data streams for one type of object can be fruitful, for two reasons:

1. These databases can be used to validate the already acquired data. For example, Denmark State Railways (DSB) use weight sensors on the top of APC at doors to supply information about occupancy of trains to passengers (International Association of Public Transport, 2022).
2. A secondary database can be used to fill missing data in the primary database.

6. An illustration of data needs for UCs in WP6

Based on the discussion and use cases defined in WP6, an analytical survey was conducted among the task leaders. The purposes of the survey were:

- Mapping the data requirements of different use cases in WP6.
- Ranking the most required data categories.
- Identifying the need of data simulation based on the mapped data streams (sources).

6.1. Data requirements for UCs in WP6

Table 2 outlines the use cases from WP6 which require data from T6.3. These use cases are directly adopted from each deliverable and corroborated with D6.9.

Table 2. Use cases identified for data requirements mapping

Task	UC Id	UC name
6.1.1	UC-FP6-WP6-1.1.1	Travel planning for regional lines including a DRT service for first/last mile (demand-responsive transport such as taxi or ridesharing services)
	UC-FP6-WP6-1.1.2	Travel planning for regional lines taking into account rules of competition for Public Transit and DRT
	UC-FP6-WP6-1.1.3	Simulation of required DRT capacity based on predicted travel demand
	UC-FP6-WP6-1.1.4	Support OJP (Open API for distributed journey planning) trip search requests and include DRT in the response
6.1.2	UC-FP6-WP6-1.2.1	Synchronization of operational processes among regional rail operators to adjust ad-hoc timetables
	UC-FP6-WP6-1.2.2	Synchronization of operational process among regional rail operators and other services to adjust ad-hoc timetables
	UC-FP6-WP6-1.2.3	Trip search based on the ad-hoc timetable
	UC-FP6-WP6-1.2.4	Passenger information portal providing personalized details about regional connections and services at stations
	UC-FP6-WP6-1.2.5	Passenger information portal provides a map showing Points of Interest that can be individually filtered by category
	UC-FP6-WP6-1.2.6	Travel planning for specific user groups with reduced mobility (Selection of a default profile)
	UC-FP6-WP6-1.2.7	Journey planning for passengers with reduced mobility with a personalised profile (Adjustment of a default profile)
	UC-FP6-WP6-1.2.8	Using pareto-search to minimize walking distance
6.2	UC-FP6-WP6-2.01	Sending updated operational plan and calculated forecast provided by the TMS to passenger information services/systems (PIS)

Task	UC Id	UC name
	UC-FP6-WP6-2.02	Usage of the number of expected travellers for timetable planning or traffic dispatching
	UC-FP6-WP6-2.03	Receiving and using the number of expected travellers between subsequent stops of a given train for timetable planning or traffic dispatching
	UC-FP6-WP6-2.04	Receiving and using the number of expected travellers embarking/disembarking at the stations for timetable planning or traffic dispatching
6.4	UC-FP6-WP6-4.01	Forecast Occupancy of Vehicles using Journey Planning Requests Data
	UC-FP6-WP6-4.02	Display Forecasted Occupancy Information to Travelers when Planning Trips
	UC-FP6-WP6-4.03	Estimation of Mobility Demand beyond Rail (First/Last Mile Analysis)
	UC-FP6-WP6-4.04	Detection and Characterization of Abnormal Train Usage Peaks
6.5	UC-FP6-WP6-5.01	Impact of Weather and Train Composition on Train Schedules and Delays
	UC-FP6-WP6-5.02	Synchronization Between Train and Regional Bus Schedules
	UC-FP6-WP6-5.03	Traveller feedback for congestion analysis
	UC-FP6-WP6-5.04	Train Platform Allocation Problem
6.6	UC-FP6-WP6-6.01	User plans to ship a single parcel from station A to station B without transfer possibility (must stay on the train) and with mandatory personal drop off and pick-up
	UC-FP6-WP6-6.02	User plans to ship a single parcel within a region from address A to address B via parcel lockers
	UC-FP6-WP6-6.03	CEP company plans to install a regional collection- and distribution network for parcels. This use case includes the shipment of single parcels within the region (see UC-FP6-WP6-6.2)
6.8	UC-FP6-WP6-8.01	Business Intelligence Analysis
	UC-FP6-WP6-8.02	Demand based adjustment to train schedules
	UC-FP6-WP6-8.03	Service feedback from customers
	UC-FP6-WP6-8.04	Documentation of software configurations, procedures, and changes
	UC-FP6-WP6-8.05	Data backup
	UC-FP6-WP6-8.06	Data retrieval from backup

Task	UC Id	UC name
	UC-FP6-WP6-8.07	Logging, Auditing and Compliance
	UC-FP6-WP6-8.08	Cross-border journey planning
	UC-FP6-WP6-8.09	Performance requirements and concurrency
	UC-FP6-WP6-8.10	Software availability
	UC-FP6-WP6-8.11	Redundancy of system components
	UC-FP6-WP6-8.12	Analysing the acceptance of the service
	UC-FP6-WP6-8.13	Analysis for infrastructure expansion or adoption by the network planner
	UC-FP6-WP6-8.14	Staff demand planning based on forecasted customer demand
	UC-FP6-WP6-8.15	Forecasted demand for train fleet size
	UC-FP6-WP6-8.16	Travel service notifications
	UC-FP6-WP6-8.17	DRT service provider demonstrates interest to join ecosystem

Based on an internal survey from WP6 participants, another table was generated. Table 3 outlines the data which will be required by the use cases mentioned in Table 2. Codes are assigned to data types to make the table simpler and easy for referencing.

Table 3. Data identified for use cases

Code	Data
D01	Journey planning requests (from journey planning app)
D02	Historical vehicle occupancy (e.g. door counts)
D03	Forecasted vehicle occupancy (from historical data or journey planning requests)
D04	Train schedules /timetables (static and real-time)
D05	Train composition (includes capacity, PRM access)
D06	Weather data (historical and forecasted)
D07	Peak hour estimation (maximum demand in a day)
D08	Simulated passenger demand (from historical demand or literature)
D09	Station infrastructure data (platforms, PRM access)
D10	User Feedback (from journey planning app)
D11	Bus schedules /timetables
D12	Parking information (spaces, PRM parking)
D13	Information on public/disruptive events

Finally, the data were mapped against the use cases and the combinations are shown in Table 4. Table 4 shows which different data types are required by each use case in WP6. Initial identifier 'UC-FP6-WP6' was removed to fit the table in the document.

Table 4. Data required by different use cases in WP6

UCs	D01	D02	D03	D04	D05	D06	D07	D08	D09	D10	D11	D12	D13
1.1.1				x							x		
1.1.2				x							x		
1.1.3		x	x		x			x					
1.1.4				x							x		
1.2.1				x									
1.2.2				x							x		
1.2.3				x							x		
1.2.4				x	x				x		x		
1.2.5									x				
1.2.6				x	x				x		x	x	
1.2.7				x	x				x		x	x	
1.2.8				x					x		x		
2.01				x									
2.02		x	x	x			x						
2.03		x	x	x			x						
2.04		x	x	x			x						
4.01	x	x		x	x								
4.02			x										
4.03	x												

UCs	D01	D02	D03	D04	D05	D06	D07	D08	D09	D10	D11	D12	D13
4.04			x	x	x	x	x						x
5.01		x	x	x	x	x							
5.02				x							x		
5.03				x						x			
5.04	x	x	x	x	x			x	x				
6.01			x	x									
6.02			x	x									
6.03			x	x									
8.01	x	x	x	x	x	x	x	x	x	x	x		
8.02	x	x	x	x	x			x					
8.03										x			
8.04													
8.05													
8.06													
8.07													
8.08													
8.09													
8.10													
8.11													
8.12	x	x	x				x	x		x			

UCs	D01	D02	D03	D04	D05	D06	D07	D08	D09	D10	D11	D12	D13
8.13	x	x	x	x	x		x	x	x				
8.14	x	x	x	x	x		x	x					
8.15	x	x	x	x			x	x					
8.16				x							x		
8.17													
Sum	9	13	17	28	12	3	8	8	8	4	11	2	1

The data which are most important i.e., which are required by majority of the use cases are:

- D04: Train schedules /timetables (scheduled and delays); it is required by 28 use cases.
- D03: Forecasted vehicle occupancy (from historical data or journey planning requests); it is required by 17 use cases.
- D02: Historical vehicle occupancy (e.g. door counts); it is required by 13 use cases.
- D11: Bus schedules /timetables; it is required by 11 use cases.

6.2. Data gaps and simulated data

The aim of the present task (T6.3) is to map the data required for the UCs in the WP6. Data access can be described as a spectrum of availability, ranging from data which is expected to exist to data which is readily available to everyone. The matrix illustrating data availability and readability is already outlined in Figure 1. However, Table 5 presents a more elaborate version of this matrix. The data sources given in Table 3 will be filled in Table 5 as the project progresses.

Table 5. Matrix to evaluate the need of data simulation

Type	Accessibility	Availability in T6.3/11.3			
		Expected to be available	Sample available	Large data available	Real-time data available
System internal	Restricted	D01, D10		D02	
API	Open				D06
	Restricted				D04
Databases	Open			D04	D11
	Restricted				

Type	Accessibility	Availability in T6.3/11.3			
		Expected to be available	Sample available	Large data available	Real-time data available
Tables, spread sheets	Open			D05, D09, D12	
	Restricted				
Text	Open				
	Restricted				D13

The development in other WP6 tasks largely depends on data, as illustrated in tables above.
To not delay such developments, the temporary use of simulated data can be an approach, especially for the 'Sample available' category in Table 5. This way, a small dataset can be enlarged to make it suitable for testing and development and later replaced with real data in the same format. This approach may not be necessary for the most important data but can be useful for other types of data.

It is presently an open question to what extent the same type of data that has been found in Norway is available for other selected countries. If that proves to be difficult, one alternative is to use data from Norway, or data generated based on the characteristics of Norwegian data as proxy for data in other countries. The examples in Table 5 regarding the mapped data will be extended in T11.3. Thus, this table is left open for discussion in future objectives of T11.3.

6.3. User feedback as a data

User feedback is a particular type of data that is difficult to obtain in early phases of this type of developments. Even if agile methods are applied to generate continuous software deliveries during development, a certain critical mass of functionality is needed to obtain meaningful user feedback. In addition, it is challenging to engage 'average' users at this stage (specification stage). User feedback will therefore be based on the stated expectations from researchers involved in the project, who will have role as users, in addition to being developers. Thus, the definition of 'user' in user feedback will evolve during the specification and demonstration phases of the tasks. At this stage of development, user feedback is mainly related to feedback on the usability of the provided data for use in other WP6/11 tasks. Thus, user in user feedback can be termed as anyone who uses data and provide feedback on it to improve the journey planning. Such users at this stage can be software developers.

7. Conclusions

The aim of this report is to map the data required for UCs in WP6 of FP6 of Europe's Rail Joint Undertaking. This deliverable provides:

- Review and description of existing data bases.
- Descriptions of requirements with a specific focus on regional lines and definitions.

To support the analysis of relevant data, the report proposes novel models for illustrating different levels of data availability. One model uses two dimensions: ease of data access, and readability. Ease of access shows to what extent data is practically available for researchers in general, and specifically for the use cases in WP6. The ease of access ranges from confidential or internal, to available upon request to the public. The readability scale represents how much preprocessing is needed to use the data. The least readable data are tacit and can only be obtainable directly from resource persons in the industry. The next level is text on websites, PDF files and similar typically unstructured formats. This data requires either manual or automatic preprocessing. Depending on the level of structure, such preprocessing has the potential for automation but still needs human quality assurance. Data in Excel and CSV files represent the next level of readability, which typically requires less preprocessing than the lower ones. Finally, standardized databases, for example, available through APIs, represent the highest readability and should not require preprocessing. The model shows that the ideal data comes from fully public standardised databases. As far as possible, it is desirable to use such data.

The report has reviewed the availability of high-quality information and whether such data is available in sufficient and complete databases. It is challenging to obtain such data, even though many EU directives require its availability. Data tends to be in different formats in different countries. Even when data is available, there are issues regarding business confidentiality and GDPR that limit the practical sharing of data. Scalability is a key challenge that requires additional focus. Expansion to Europe will require extensive coordination with data providers and transport authorities. EU-wide standardization efforts, such as through the EU-RAIL initiative, will ensure data interoperability and availability.

Data regarding railway infrastructure is available in a large array of distinctions. From static to dynamic and from standard format to rough recorded logs, these data arrays are described. The report provides examples of data from suitable databases for train traffic. Such data have been identified, exported and pre-processed for further analyses, with data from Norway acting as a pilot case. Data needs for the different WP6 demo cases have been collected and structured. The specific data formats and larger volumes have mainly been obtained from Norway.

The present impression is that there is a limited need to use simulated data, at least for the data types needed for most use cases. It may, however, be needed for simulations of some types of data. As for data requirements and data sets for first and last miles, requirements have been defined on a general level and examples provided. This is, however, a type of data that railway organisations, such as operators and infrastructure managers, typically do not have in-house. It is, therefore, necessary to use openly available data such as demographic statistics and Open Street

Maps.

It has been experienced that user feedback is difficult to obtain in the early phases of this type of development, but efforts are being made to get such feedback as early and as representative as possible.

References

AB Storstockholms Lokaltrafik, 2024. [Internett]
Available at: https://play.google.com/store/apps/details?id=com.sl.SLBiljetter

AtB, 2024. [Internett]
Available at: https://play.google.com/store/apps/details?id=no.mittatb.store

AtB, 2024. [Internett]
Available at: https://www.atb.no/reiseplanlegger

Australian Transport Assessment and Planning, u.d. *Overview of transport modelling.* [Internett]
Available at: https://www.atap.gov.au/tools-techniques/travel-demand-modelling/2-overview
[Funnet 2024].

BaneNor, 2024. *Infrastructure (Network Statement).* [Internett]
Available at: https://networkstatement.banenor.no/doku.php?id=infrastructure

Caban, J., 2021. Traffic congestion level in 10 selected cities of Poland. *Scientific Journal of Silesian University of Technology. Series Transport,* pp. 17-31.

CFL, 2024. [Internett]
Available at: https://play.google.com/store/apps/details?id=de.hafas.android.cfl

COMPANHIA CARRIS DE FERRO DE LISBOA, E.M., S.A., 2024. [Internett]
Available at: https://play.google.com/store/apps/details?id=pt.carris.tecmic

Dilax, 2024. *Automatic Passanger Counting (APC).* [Internett]
Available at: https://www.dilax.com/en/products/automatic-passenger-counting

ENTUR, 2024. *Stops- and Timetable data.* [Internett]
Available at: https://developer.entur.org/stops-and-timetable-data

ERA, 2024. *Rolling Stock - Locomotives and Passengers TSI.* [Internett]
Available at: https://www.era.europa.eu/domains/technical-specifications-interoperability/rolling-stock-locomotives-and-passengers-tsi_en

Frazer Norwell, 2023. *What the collapse of Randklev Bridge means for rail travel in Norway.* [Internett]
Available at: https://www.thelocal.no/20230815/what-the-collapse-of-randklev-bridge-means-for-rail-travel-in-norway

Hamner, B., 2010. *Predicting Travel Times with Context-Dependent Random Forests by Modeling Local and Aggregate Traffic Flow.* s.l., s.n., pp. 1357-1359.

Infrabel, 2024. *Platform heights in stations.* [Internett]
Available at: https://opendata.infrabel.be/explore/dataset/perronhoogten-in-stations/table/?disjunctive.longnamefrench&disjunctive.platform_type&disjunctive.hauteur&disjunctive.type_stopping_point&disjunctive.area&disjunctive.arrondissement&sort=longnamedutch

International Association of Public Transport, 2022. *Improving passenger flow and crowd management through technology and innovation,* s.l.: International Association of Public Transport.

Kölner Verkehrs-Betriebe AG, 2024. [Internett]
Available at: https://apps.apple.com/de/app/kvb-app/id1441639226

Kuipers, R. & Palmqvist, C.-W., 2022. Passenger Volumes and Dwell Times for Commuter Trains: A Case Study Using Automatic Passenger Count Data in Stockholm. *Applied Sciences.*

MET, 2024. *Frost API.* [Internett]
Available at: https://frost.met.no/index.html
Moovit, 2024. [Internett]
Available at: https://play.google.com/store/apps/details?id=com.tranzmate
NAPCORE, 2024. *National Access Point.* [Internett]
Available at: https://napcore.eu/description-naps/national-access-point/
NorskeTog, 2024. [Internett]
Available at: https://www.norsketog.no/en/trains/motorvognsett
ÖBB-Personenverkehr AG, 2024. [Internett]
Available at: https://apps.apple.com/at/app/%C3%B6bb-scotty/id315497345
RailNetEurope, 2024. *RNE Network Statement Common Structure.* [Internett]
Available at: https://rne.eu/organisation/network-statements/
SCB, 2024. [Internett]
Available at: scb.se
SJ AB, 2024. [Internett]
Available at: https://play.google.com/store/apps/details?id=se.sj.android
Sørensen, A. Ø. et al., 2018. Use of mobile phone data for analysis of number of train travellers. *Journal of Rail Transport Planning & Management,* pp. 123-144.
SSB, 2024. [Internett]
Available at: ssb.no
TFI, 2024. [Internett]
Available at: https://play.google.com/store/apps/details?id=com.trapezegroup.TFILive.nta
TomTom, 2024. *TomTom.* [Internett]
Available at: https://www.tomtom.com/
Verkehrsverbund Rhein-Sieg GmbH, 2024. [Internett]
Available at: https://apps.apple.com/de/app/vrs-auskunft/id398472681
Wiener Linien GmbH, 2024. [Internett]
Available at: https://apps.apple.com/at/app/wienmobil/id1107918142

Annex 1

Door count sample on a train running from Oslo S to Gjøvik in Norway

Date	Train number	Station Name	Sum Boardings	Sum Alightings	Sum Passengers
01.10.2022	201	Oslo S	11	0	11
01.10.2022	201	Grefsen	0	0	11
01.10.2022	201	Kjelsås	2	0	13
01.10.2022	201	Nittedal	1	1	13
01.10.2022	201	Harestua	0	0	13
01.10.2022	201	Grua	0	0	13
01.10.2022	201	Roa	1	2	12
01.10.2022	201	Lunner	0	0	12
01.10.2022	201	Gran	0	2	10
01.10.2022	201	Jaren	0	1	9
01.10.2022	201	Bleiken	2	0	11
01.10.2022	201	Eina	1	1	11
01.10.2022	201	Reinsvoll	1	1	11
01.10.2022	201	Raufoss	2	2	11
01.10.2022	201	Gjøvik	0	11	0

Annex 2

Rolling stock specification for different class of trains in Norway

Specifications	**Class-73A train**	**Class-72 train**	**Class-70 train**	**Class-69C train**	**Class-76 train**	**Class-75 train**
Comfort seats/1st class	56	0	30	0	0	0
Standard seats/2nd class	145	305	200	252	196	235
Standing spaces (folding seats in use)	73	218	101	297	96	266
Standing spaces (folding seats not in use)	37	55	50	74	385	83
Folding seats	3	5	3	34	45	60
Wheelchairs spaces	2	1	1	1	4	4
Wheelchair elevator	0	0	0	2	2	2
Bicycle spaces	Yes	0	0	0	3	5
Sleeping spaces	0	0	0	0	0	0
Sleeping compartments	0	0	0	0	0	0
Toilets (closed systems)	5	1	3	1	2	0
Toilet (open systems)	0	0	0	0	0	0
Handicap toilets	1	1	1	1	1	1
Family area, number of seats	16	0	0	0	0	0
Restaurant (number of seats)	19	0	0	0	0	0
Serviced kiosk	0	0	0	0	0	0
Vending machine	1	0	4	0	Yes	3

Annex 3

Network statement summary sample for some train stations in Norway

Station services	**Gjøvik station**	**Raufoss station**	**Reinsvoll station**	**Eina station**	**Roa station**
Parking	Yes	Yes	Yes	Yes	Yes
Pick-up/drop-off point	✓	✓	✓	✓	✓
Parking for travellers	✓	✓	✓	✓	✓
Disabled parking	✓	✓			✓
Free Parking	✓	✓	✓	✓	✓
Pay and display machines	✓				
Commuter parking with app	✓				
Bike racks with roof	✓			✓	✓
Bike racks without roof		✓			
Bike hotel					
Availability	Yes	Yes	Yes	Yes	Yes
Disabled access to platform	✓	✓			
Disabled toilets	✓			✓	
Mobile ramp on platform (entry and exit)					✓
Travel info	Yes	Yes	Yes	Yes	Yes
Static information; Posters					
Timetables	✓	✓	✓	✓	✓
Line map	✓	✓	✓	✓	✓
Informational posters	✓	✓	✓	✓	✓
Monitor with train times	✓	✓	✓	✓	✓
Speaker	✓	✓	✓	✓	✓

Station services	**Gjøvik station**	**Raufoss station**	**Reinsvoll station**	**Eina station**	**Roa station**
Facilities	Yes	Yes	Yes	Yes	Yes
Waiting room	✓	✓	✓	✓	✓
Weather protection (Sheds/Platform roof)					
Non-smoking platforms and stations	✓	✓	✓	✓	✓
Luggage storage	✓				
Newsstands	✓				
Transportation/Communication within walking distance	Yes	Yes	Yes	Yes	Yes
Bus	✓	✓	✓	✓	✓
Taxi	✓	✓			
Airport outside walking distance	✓				

Annex 4

APIs and functions on weather data from Norwegian Metrological Institute

These Python functions call APIs for

1. Geolocation: This code outputs the latitude and longitude based on search locations (similar to Google maps but free)

```
def geolocation(search_term): # get coordinates of place based on a search term for station
    parameters = { 'text': search_term+' Stasjon', 'lang': 'en', 'boundary.country':'NO'}
    location = requests.get('https://api.entur.io/geocoder/v1/autocomplete', parameters)
    try:
        return location.json()['features'][0]['geometry']['coordinates']
    except KeyError:
        return geolocation(search_term)
```

2. Source: This code outputs the weather station nearest the geolocation which records a required weather element.

```
def source(latitude, longitude, weather_type):  # get the nearest weather station code from the station location
    parameters1 = { 'geometry': 'nearest(POINT('+str(longitude)+' '+str(latitude)+'))','elements': weather_type,}
    try:
        source = requests.get('https://frost.met.no/sources/v0.jsonld', parameters1,
auth=(new_frost_key(),'')).json()['data'][0]['id']
        return source
    except KeyError:
        return source(latitude, longitude, weather_type)
```

3. Weather_element: This code outputs the historical weather element from the source on a specific date.

```
def weather_element(source, date, weather_type): # get the weather data from the weather station nearest to the station
location
    parameters = {'sources': source,'elements': weather_type,  'referencetime': date,}
    try:
        reqst = requests.get('https://frost.met.no/observations/v0.jsonld', parameters, auth=(new_frost_key(),''))
        elmnt = reqst.json()['data'] #[0]['observations'][0]['value']  #[0]['value']
        return elmnt
    except KeyError:
        return weather_element(source,date, weather_type)
```

4. Weather_forcast_temp: This code outputs the forecasted weather in 24-hour window.

```
def weather_forcast_temp(location): # get the weather forecast from the weather station nearest to the station location
    session = requests.Session()
    session.headers['User-Agent'] = 'Chrome/120.0.0.0'
    session.params['lon'], session.params['lat']  = geolocation(location)
    reqst = session.get('https://api.met.no/weatherapi/locationforecast/2.0/complete.json')
    daily_forcast = reqst.json()['properties']['timeseries'][0]['data']['instant']['details']['air_temperature']
    return daily_forcast
```