C4R

Capacity for Rail

Towards an affordable, resilient, innovative and high-capacity European Railway System for 2030/2050

Verified data architecture

Submission date: 22/03/2017

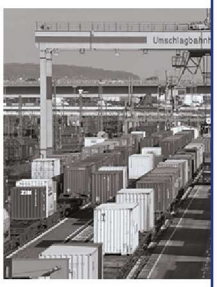
Deliverable 34.2

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 605650







Collaborative project SCP3-GA-2013-60560
Increased Capacity 4 Rail networks through enhanced infrastructure and optimised operations
FP7-SST-2013-RTD-1

Lead contractor for this deliverable:

- University of Birmingham

Project coordinator

- International Union of Railways, UIC

Executive Summary

Deliverable D3.4.2 of the C4R project complements the work done in D3.4.1 by looking at the ICT architectures and data resources the industry may choose to draw on in the short to medium term.

Section 2 of the document presents a candidate software architecture for future TMS platforms, the Enterprise Service Bus. The ESB is becoming the architecture of choice within the software engineering community, and has already been proposed for use in the rail sector by previous projects including InteGRail and ON-TIME.

Section 3 of the document focusses on semantic models for data integration, and the associated supporting architectures. It presents the potential costs and benefits of the semantic approach to data integration, and discussed some of the limitations including the scalability of the models, the challenge of distributing the reasoning architecture, and version management.

Section 4 shows how data external to the railways can be used to support railway operations through the provision of enhanced situational awareness. It showed how geotagged data can be harvested from social media platforms, assigned to train services, and presented to control room staff based on an analysis of the sentiment contained in the message. It also discusses some of the issues of the use of social data in a railway context, including the need to ensure user privacy, and issues of the trustworthiness of the data, particularly if that data is being used as the basis for decision making.

Summary of recommendations:

- That the Enterprise Service Bus architecture is an appropriate foundation for the development of Traffic Management Systems, and that the outcomes from existing research activities using this technology, including the FP7 ON-TIME project, should be adopted by the industry;
- That based on figures from similar industrial sectors, the costs of poor integration of data between ICT systems in the European railways could be as much as 1% - 2% of annual revenue. Continuing with efforts to improve the management and integration of data resources within the industry must be treated as a priority over the next 5 years. The appropriate usage of new technologies, including ontology, has a key role to play in these improvements;
- The harvesting and utilisation of open data from public sources has great potential to provide the industry with enhanced situational awareness during disruptions, but care is needed to ensure user privacy is respected, and that (as far as is practical) the authenticity and provenance of material used for decision making.

Table of contents

Index of figures

Index of tables

Abbreviations and acronyms

Abbreviation / Acronym	Description
API	Application Programming Interface
AWT	All Ways Travelling (EU-founded project)
ERA	European Railway Agency
ERIM	European Railway Infrastructure Masterplan
ESB	Enterprise Service Bus
GML	Geographic Markup Language
GTFS	General Transit Feed Specification
HMI	Human-Machine Interface (train driver display)
ICT	Information and Communications Technology
IDM^{VU}	Infrastruktur-Daten-Management für Verkehrsunternehmen
IEEE	Institute of Electrical and Electronics Engineers
IGRIS	InteGRail Information System
IM	Infrastructure Manager
LOD	Linked Open Data
MAAP	Multi-Annual Action Plan
NaPTAN	National Public Transport Access Node
NASA	National Aeronautics and Space Administration
NeTEx	Network and Timetable Exchange
NRE	National Railway Entity
PIS	Passenger Information System
OSM	OpenStreetMap
OWL	Web Ontology Language
RaCoOn	Railway Core Ontology
railML	Railway Markup Language
RCM	Remote Condition Monitoring
RDF	Resource Description Framework
RINF	Register of Infrastructure
RTPI	Real-Time Passenger Information System
RTTP	Real-Time Traffic Plan
S2R	Shift2Rail
SOA	Service Oriented Architecture
SPARQL	Protocol and RDF Query Language
SWRL	Semantic Web Rule Language
TMS	Traffic Management System
UIC	International Union Of Railways
WP	Work Package
XML	Extensible Markup Language
XSD	XML Schema Definition

1. BACKGROUND

1.1 INTRODUCING CAPACITY4RAIL

Note to readers: section 1.1 of this document provides a brief introduction to the C4R project, and to the aims of the SP3 / WP3.4 work stream. Readers familiar with this background may skip to section 1.2 for a brief review of the storyboards used by the SP3 team to structure the work in this area, or to section 2 for the material of software architectures.

CAPACITY4RAIL (C4R) is an EU FP7 funded project that aims to answer the question *"How can we obtain an affordable, adaptable, automated, resilient and high-capacity railway, for 2020, 2030 and 2050?"*

CAPACITY4RAIL will provide an overall increase in railway capacity through the adoption of a holistic view of the railway as a system of interacting technical components, delivering services that are driven by customer demand. It is structured into sub-projects (SP) with interacting work packages (WP) as presented in Figure 1-1.

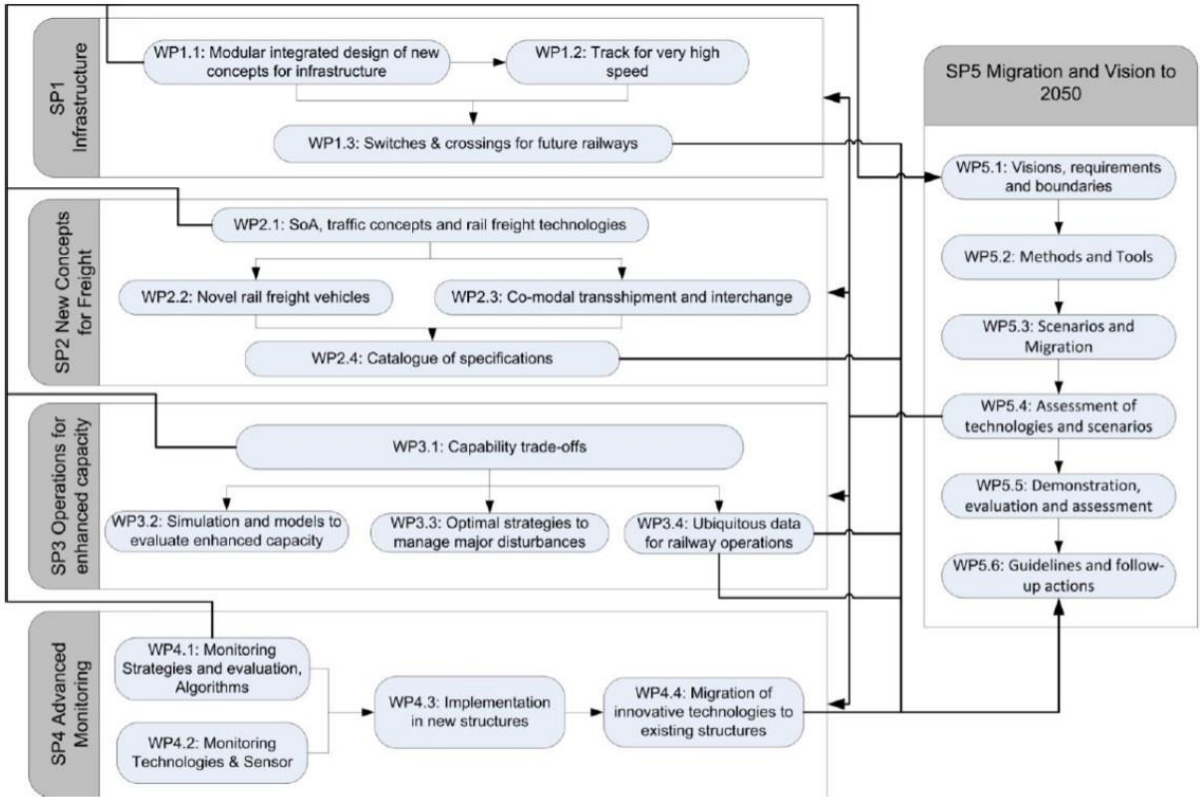


FIGURE 1-1 CAPACITY4RAIL STRUCTURE BREAKDOWN AND INTERACTIONS

Sub-Project 3 (SP3) of CAPACITY4RAIL, “Operations for enhanced capacity”, is focussed on delivering tools that will help planners to understand and prioritise system capabilities and decide on optimal strategies to:

- Increase overall system capability;
- Respond dynamically to planned and unplanned changes; and
- Support real-time punctuality management.

Guidance produced under SP3 will cover topics such as incident and emergency management, and will include (for example) recommendations on the management of extreme weather situations. C4R technology roadmaps, based on the concept shown in Figure 1-2 for automation of traffic management systems, will enable the rail industry to meet the challenges of the future; including the need to accommodate high-speed rail freight solutions on infrastructure that is already capacity limited, and the need for greater levels of transhipment between rail and other modes. (CAPACITY4RAIL, 2013). Further to this, links between the members of the SP3 team from the C4R consortium, and the consortia involved in ongoing projects (e.g. IN2RAIL) will ensure that the outcomes of work in C4R as appropriate.

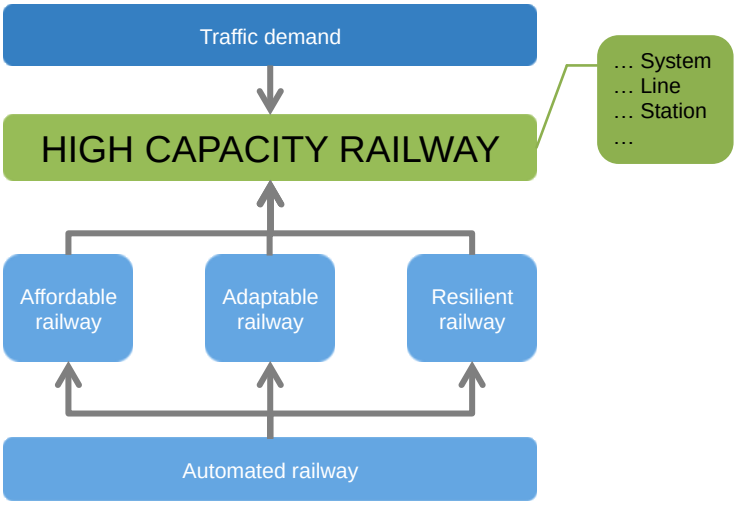


FIGURE 1-2 ROADMAP FOR AUTOMATION OF TRAFFIC MANAGEMENT SYSTEMS

1.1.1 UBIQUITOUS DATA FOR RAILWAY OPERATIONS

WP3.4 of the C4R project, “Ubiquitous data for railway operations”, aims to assist the rail industry with the collection, storage, and effective utilisation of large and diverse sources of data, and the extraction of meaningful information and knowledge to support operational strategies.

The first deliverable from WP3.4, D3.4.1 focussed on the current state of the art in data exchange used in Europe’s railways, and also considered the touch points with other transport modes — a key consideration when optimising the use of available capacity. As the suite of models used internally by

the industry are already comparably well understood by the relevant stakeholders, a particular focus was placed on open sourced, and community-developed data models; these models represent a significant investment of time by their development communities, and as a result of the wide variety of use-cases they're designed to support tend to offer a broader view on a topic than the more focussed, safety-driven in-industry models, making them very suitable for use in non-safety critical tasks beyond the scope of rail, such as managing multimodal interactions.

In this document, D3.4.2 of the C4R project, the team have built on the findings of D3.4.1 to propose how the models and services described might evolve in the near-to-mid timeframe, presenting material on novel data sources, representations, and architectures that support the visions captured in the SP3 storyboards. As was the case in D3.4.1, the team believe that the complex, multi-stakeholder business environment of the multimodal transport system is based served by the sharing of data and models wherever practicable, and as a result the provision and utilisation of open data features as a cross cutting theme both throughout this text, and between the two documents (Figure 1-3).

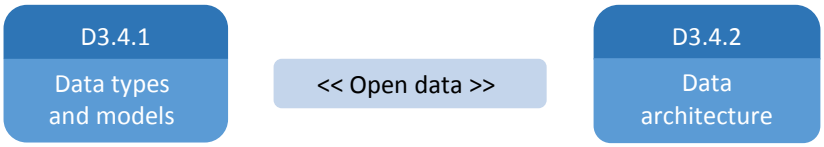


FIGURE 1-3 WP3.4 DELIVERABLES D3.4.1 AND D3.4.2

The remainder of this document is structured as shown in Figure 1-4.

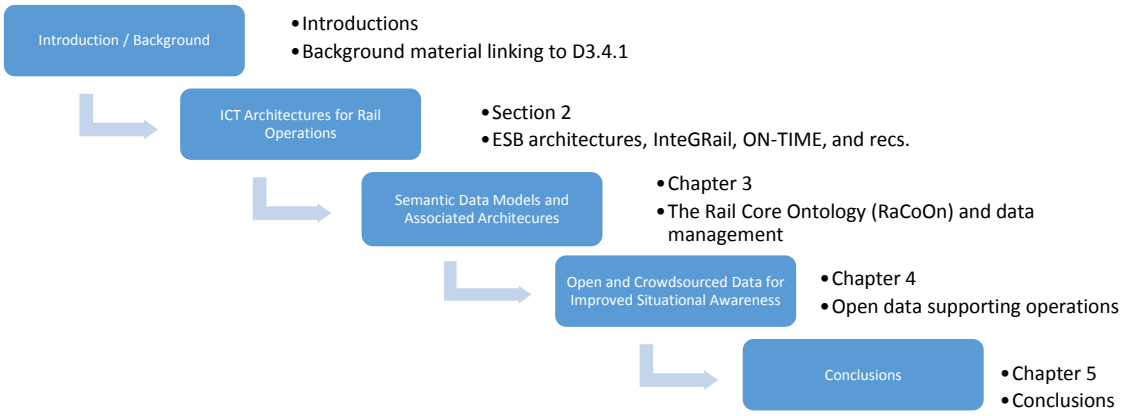


FIGURE 1-4 STRUCTURE OF DELIVERABLE 3.4.2

1.2 Storyboards for Ubiquitous Data in Support of Railway Operations

Note to readers: the content in this section is summarised from D3.4.1, and forms necessary background to the work in this document. Readers familiar with C4R D3.4.1 may skip this section and proceed directly to section 2 “Software Architectures for the Rail Industry”.

C4R deliverable D3.4.1 presented and analysed a set of three storyboards for ubiquitous data in railway operations, discussing their data requirements, the relationship between those requirements and what can be delivered using existing models, and the potential for supplementing privately-held rail industry data resources with open data from the wider community. The storyboards provided a structure on which discussions could be based, an important reporting device in a domain where, technologically-speaking, any advances will inevitably evolve beyond all recognition in the medium to long term.

The storyboards were created using a backcasting approach, starting from visions of an “ideal” industry in 2050; this was to ensure that, as far as practically possible, they retained a focus on the desired capabilities of the railway system rather than specific technological solutions. The storyboards, which are presented in detail in sections 1.2.1, 1.2.2, and 1.2.3, covered the following topics:

- Consistent Cross Industry Infrastructure Data;
- Multimodal Transport Systems;
- Real-time operational data.

This document will present views on the supporting infrastructure / architectural paradigms that could be used to deliver the types of services presented in the SP3 storyboards, and as such the storyboards themselves have been reintroduced at this stage to ensure that the reader has the background knowledge necessary to understand the motivation for the main topics addressed later in the document: service orientation, the use of ontology and ontological reasoning in high velocity data environments, and harvesting data for increased situational awareness from public sources.

1.2.1 CONSISTENT CROSS INDUSTRY INFRASTRUCTURE DATA IN SUPPORT OF PLANNING, SIMULATION AND OPERATIONS

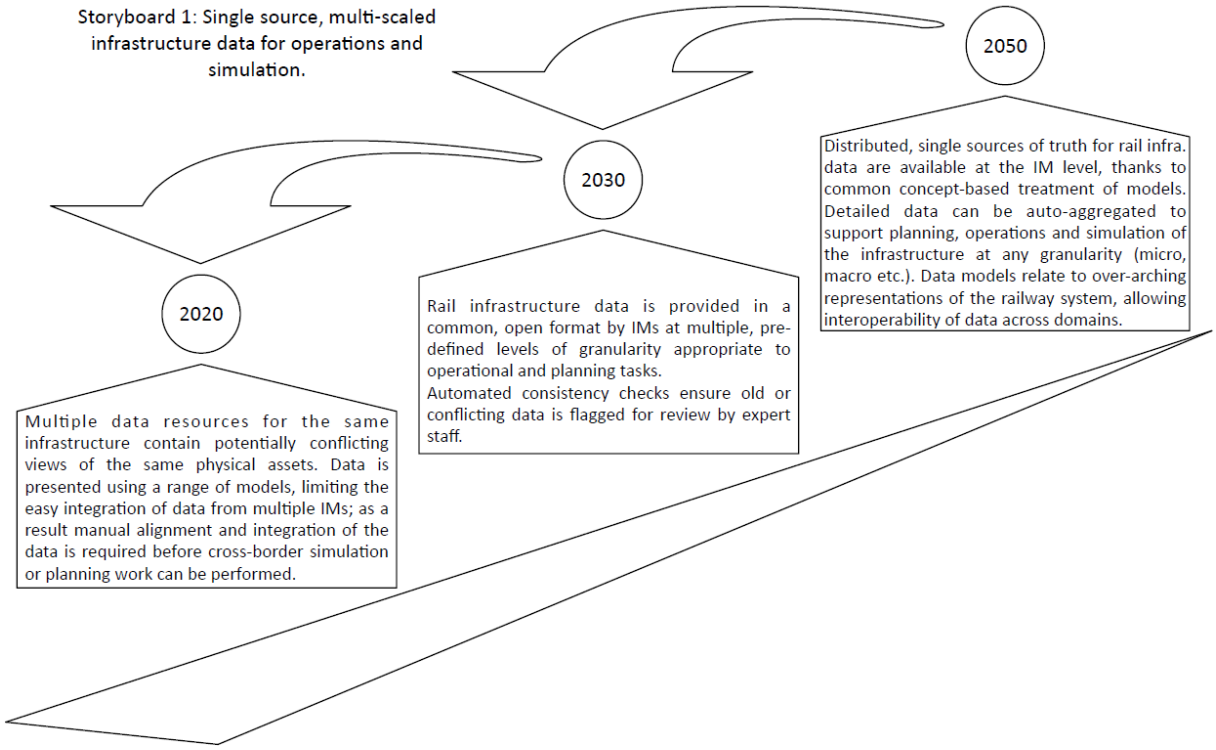


FIGURE 1-5 STORYBOARD 1 – INFRASTRUCTURE DATA FOR OPERATIONS AND SIMULATION

Figure 1-5 is a backcasting diagram showing how improved cross industry information exchange will support the increase of capacity on the existing infrastructure. This will be achieved through the delivery of more timely and accurate information to tools for planning, simulation and operations, both in national and international services. A uniform data exchange format, that is able to reflect the user needs in a future-proof way, builds a solid foundation for reducing implementation costs and enabling a broad acceptance.

Currently, different data formats for railway infrastructure data are used to exchange information between different applications, different companies or even different divisions of the same company. Besides infrastructure manager (IM) specific data formats, there exist some promising initiatives to unify data exchange formats across Europe within the railway sector and in the whole transport sector as well. These initiatives are driven by both national interest groups, like the German format IDM[VU] and by legislation mandated by the European Union, like INSPIRE and RINF. Open source initiatives based on a free cooperation of professional and unpaid developers, like railML and OpenStreetMap, are also important players in this area.

Independent from the current application perspective, infrastructure data sets have to deal with different topological granularities, which are defined in Table 1-1 (ERIM Workgroup, 2014), in order to comply with several requirements from planning, simulation and operations.

TABLE 1-1 TOPOLOGICAL GRANULARITIES – GENERAL ABSTRACTION LEVELS

Level	Scale	Description
Corridor	Very small	Primary routes within a network, e.g. rail freight corridor
Macroscopic	Small	A generalised view of the mesoscopic level, e.g. multiple tracks within a line appear as a single line
Mesoscopic	Intermediate	A generalised view of the microscopic level
Microscopic	Large	Track level information at the highest level of details

Figure 1-6 illustrates these levels of granularity as used in this document:

- National as well as cross-border infrastructure data sets shall seamlessly integrate at each of these levels. IM borders are treated the same way as state borders.
- Corridors mostly correspond to TEN-T corridors for certain services, such as freight and passenger or conventional and high-speed transport, including their important stations for fulfilling the service.
- Macroscopic level equals to the typical national network of lines, where a line comprises of one single track or two parallel tracks for each direction. Operational points are considered from junctions to large stations including smaller stations or just stop points. Connections between lines may be deduced from linking the same operational point ignoring the ability to traverse.
- Mesoscopic level consists of the same operational points as for the macroscopic level. Whereas tracks between operational points are defined instead of lines. Connections between tracks are established through operational points, traversing feasibility shall be provided.
- Microscopic level contains tracks, switches and crossings and more detailed trackside facilities, e.g. signals, platforms.
- Nanoscopic level, which would focus on both rails, is not considered in this document.

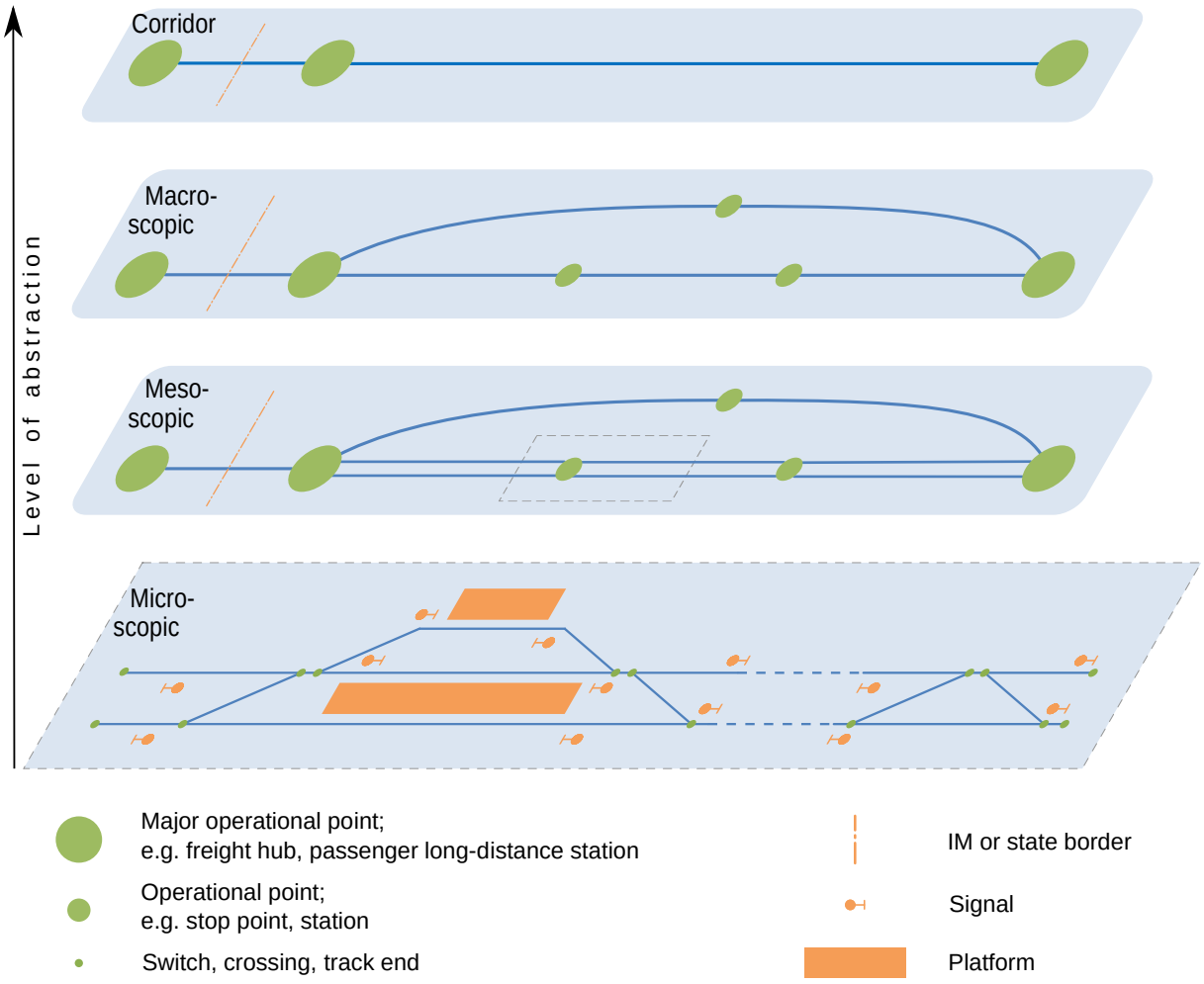


FIGURE 1-6 TOPOLOGICAL GRANULARITIES OF INFRASTRUCTURE

Depending on the nature of the task being performed, input data at any one of these levels of abstraction may be required, as shown in Figure 1-7. Generally speaking, corridor and macroscopic level data provide an adequate base for the planning of long-term and mid-term railway traffic, however mesoscopic data must be considered for more detailed operational planning tasks. Meso- and microscopic levels are appropriate for fine-grained simulation and real-time operation.

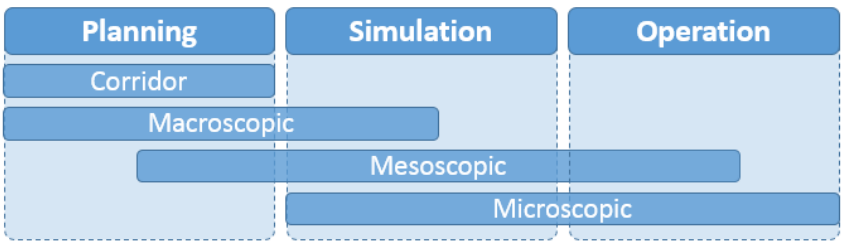


FIGURE 1-7 TOPOLOGICAL GRANULARITIES FOR PLANNING, SIMULATION AND OPERATION

In general, detailed simulations based on microscopic data sets are very time-consuming and therefore not suitable for real-time traffic controller assistance. Approximated simulations based on macroscopic data are fast enough to be used operationally, but lack conflict detection at the level of track vacancy. A compromise solution using elements of both approaches would lead to fast simulations that work well in all but the most complex capacity scenarios at junctions or busy stations.

A lack of infrastructure data at the appropriate level of granularity to support a given scale of simulation is a relatively common problem in this domain, so a further use case in this area can be found in the provision of abstracted / inferred data based on available information at a different scale to provide approximate simulation results where needed.

Additional benefits arise, if the already available level-specific data sets are joined and compared with data sets at other granularities. Storyboard 3 “Real-time operational data across organisational and member state boundaries” is partly based on this approach. The comparison of data sets from different sources may also lead to the detection of inconsistencies, enabling better data qualities.

1.2.2 Effective usage of multimodal transport system capacity

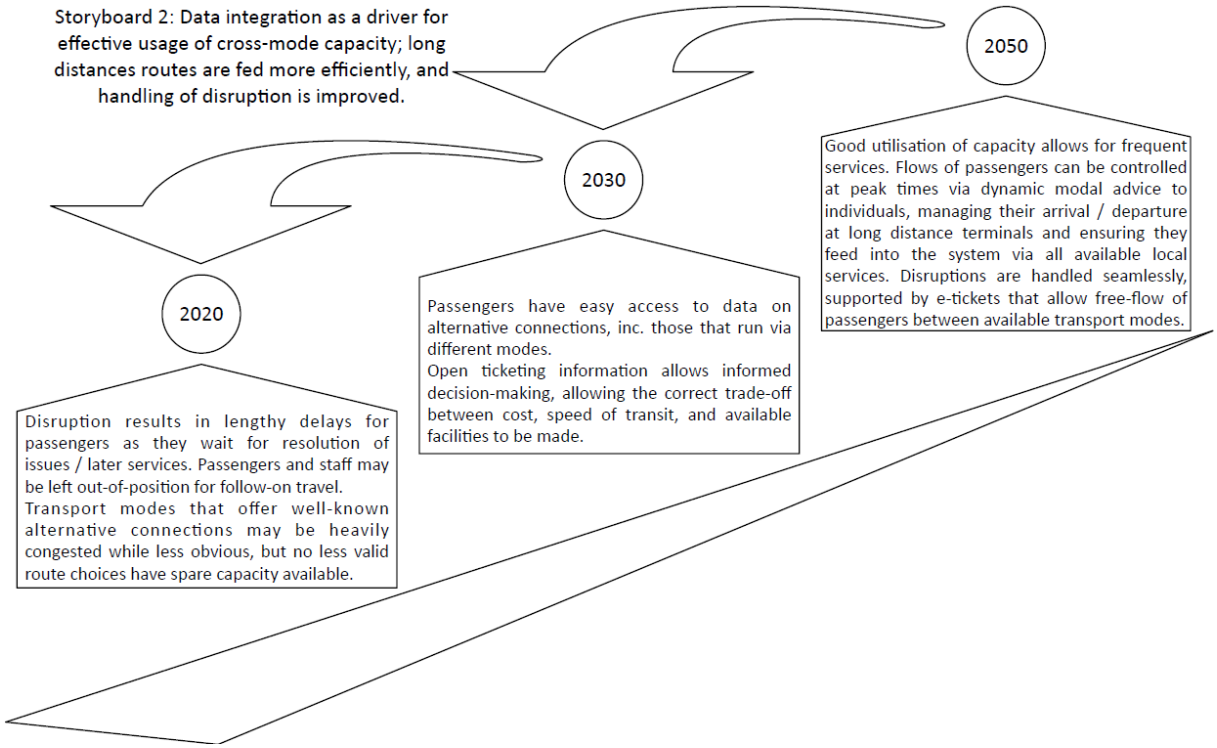


Figure 1-8 Storyboard 2 – Effective usage of cross-mode capacity

The second storyboard focuses on the use of data integration as a driver for more effective use of existing network capacity both within rail and in the wider multimodal transportation system, particularly during disrupted operations (Figure 1-8).

Multimodal transport is characterised by the use of more than one mode within the scope of a single end-to-end journey. By making the best use at available multimodal capacity, C4R will free up rail capacity and encourage modal shift from modes such as shorthand air travel. The set of transport systems considered are shown in Figure 1-9.

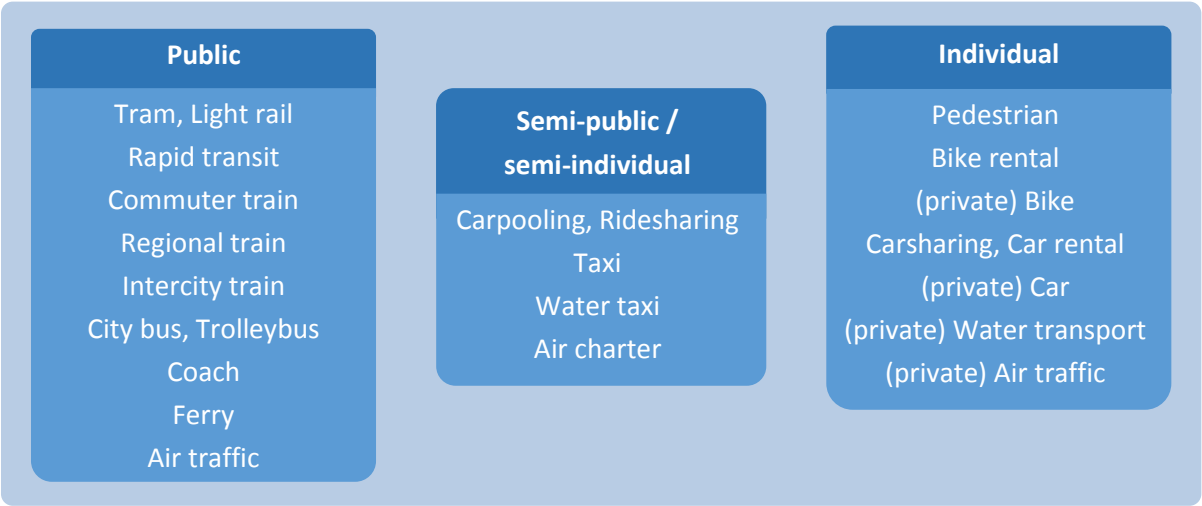


FIGURE 1-9 PASSENGER TRANSPORT MODES

The highest impact from increasing railway capacity within the scope of the second storyboard is expected to focus on public transport modes, keeping in mind that individual and semi-individual transport modes play an important role as feeder systems. Air traffic is only considered as a feeder in case of blackout or irregularities.

Beyond the direct scope of C4R, the smooth interaction of data across the outlined modes is a key outcome for the sector as described in the Whitepaper of the European Commission "Roadmap to a Single European Transport Area – Towards a competitive and resource efficient transport system", which specifies ten goals. One amongst them is "Multimodal information services" (European Commission, 2011):

> *(5) A fully functional and EU-wide multimodal TEN-T 'core network' by 2030, with a high quality and capacity network by 2050 and a corresponding set of information services.*

A similar objective has been investigated by the "*All Ways Travelling (AWT)*" consortium active from 04/2013 to 01/2016. Appointed by the European Commission, AWT will develop and validate a model for a multimodal pan-European passenger transport information and booking system (AWT Consortium, 2013 - 2015). In its first phase, "In-depth study of multimodality", AWT prepared a list of parameters that influence the passengers' decision for use of different transport modes (AWT Consortium, 2014):

- Timetable information – accurate and on short call;
- Station information – including transfer and navigation path information;
- Fare information – individual and cross-mode.

It is hoped that the reliable provision of this information subset will encourage modal shift from individual to public transport modes as driven by the availability of flexible information services.

1.2.3 Real-time operational data across organisational and member state boundaries

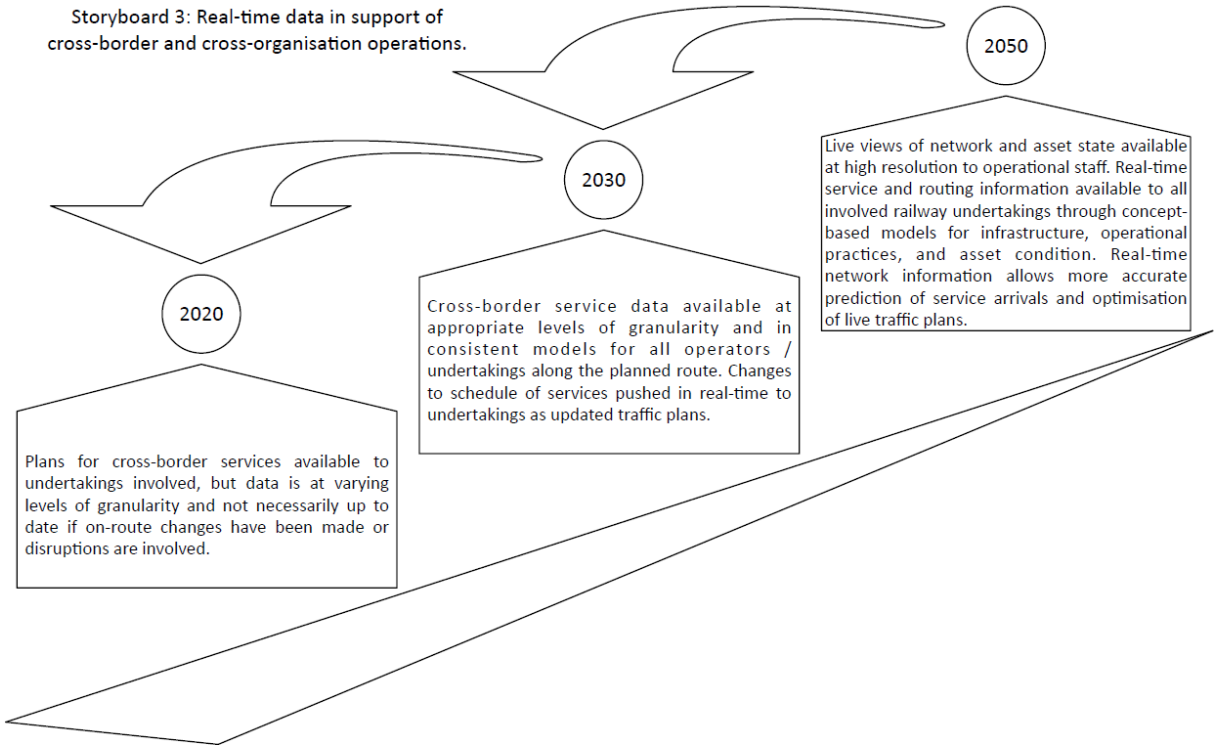


Figure 1-10 Storyboard 3 – Real-time data in support of cross-border / cross-organisation operations

The third storyboard deals with the handover of planned rail services between organisations at operational or state borders and the delivery of timely operational data throughout a journey (Figure 1-10). The availability of accurate, real-time data opens the door for operational optimization and customer information.

In addition to key data from the railway, such as booked train paths and working timetables, data from a number of external stakeholders may also influence the capacity trade-off in the wider multimodal system. In particular a clear understanding of live delays across the system as a whole and the way in which those delays propagate, are of critical importance to the effective use at capacity available.

From the perspective of data modelling, several factors have a heavy influence on this issue. On the one hand, clarifying the level of granularity and standardizing the content and transfer of planned and real-time data will enable the widest possible usage in new services. Where possible, the provided data sets shall be enriched with all available data in a standardized structure.

By providing consistent data that is more easily integrated with information from other transport modes, the rail industry will be able to maximise on its IT investments and deliver capacity improvements beyond the scope of its own assets (i.e. it will drive the use of local modes as feeders for the national rail network). Finally, real-time and planned operational data shall be available for any interested party in the railway domain.

Figure 1-11 illustrates the evolution of data model and recipients exemplarily on traffic disruption on a railway line:

- 2020: IM informs subsequent RU about the incident within his responsibility zone. Data is exchanged in the IM-RU-specific way. Thus, data format and granularity of data is not standardized;
- 2030: Subsequent RUs along the planned route get real-time operational data, even before they enter the responsibility zone of the IM in charge. All involved partners get the same information in the same format and the appropriate level of granularity including quality indication;
- 2050: IM provides information for any interested party, i.e. RUs along the planned route as well as along routes that join the disturbed route behind the disrupted area. Thus, newly available slots can be used to optimise the network capacity. Comprehensive data are provided in a detailed level of granularity.

At the local level, cross border exchange of information may take place between different stakeholders with the same function, or at operational boundaries such as the interface point between two routes. On larger scales, the same software interfaces and governance processes can be used to exchange data between different infrastructure managers (IMs) across member state borders.

Prospective optimization for a cross-border region through re-scheduling of services is based on current and short-term predicted network capacity states, which rely on sound real-time data. Integrating actual, precise, pre-processed sensor data enables more robust estimations and predictions. Research on sensor data and monitoring of railway infrastructure elements is done in the frame of SP4 of CAPACITY4RAIL, which is complementary to the findings of the current WP.

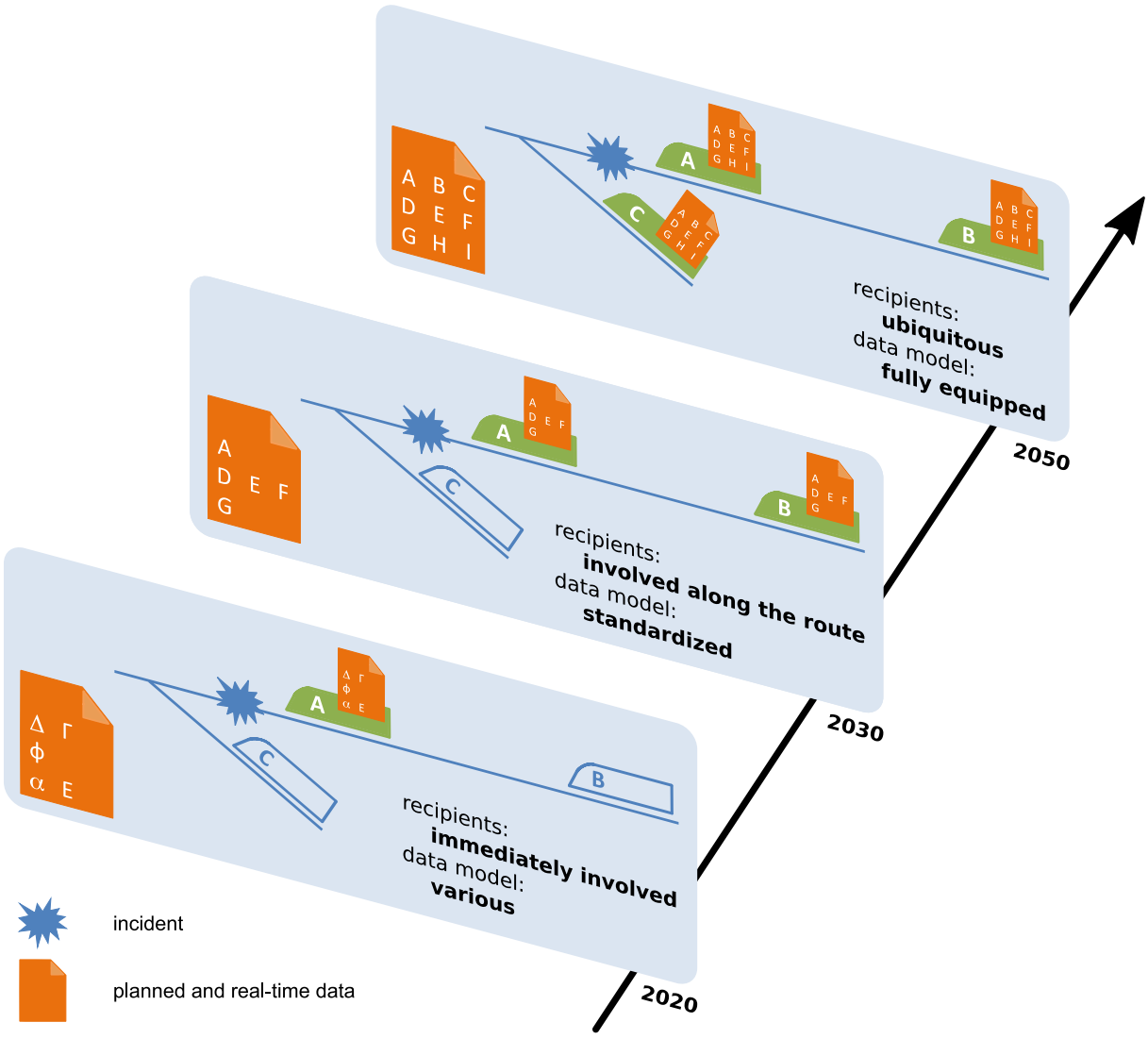


FIGURE 1-11 EVOLUTION OF AVAILABLE DATA FOR INVOLVED UNDERTAKINGS

1.3 SUMMARY OF CONCLUSIONS FROM D3.4.1 IMPACTING ON THIS DOCUMENT

The discussions in deliverable led to a number of key conclusions relevant to the scope of this document. Firstly, it provided suggestions for models and data resources that could be used to meet the needs of the storyboards (see Table 1-2), some of which have architectural requirements beyond those that are in common usage within the rail industry at this time. In particular, the linked open data architectures that underpin semantic data models are an outstanding requirement, and these will be discussed further in section 3, along with associated cost/benefit estimates (section 3.3), and limitations of the approach (section 3.4).

TABLE 1-2 DATA MODEL RECOMMENDATIONS FOR KEY CONCEPTS IDENTIFIED IN D3.4.1

Concept	Model
Network topology	RailTopoModel/railML3, OpenStreetmap/OpenRailwayMap for degraded mode
Topography	railML, OpenRailwayMap, RaCoOn or similar
Fixed asset configuration	railML
Sensor data (asset status etc.)	L.O.D. format – Semantic Sensor Network or similar
P.I.S. (paths, fare models, ticketing)	NeTEx
Timetables	railML (rail), NeTEx (multimodal)

Also in section 3, this document will discuss the use of linked open data in the context of sensors and sensor-driven operational responses. This topic will provide a bridge to the types of data needed to support novel technologies being produced in SP4 of the C4R project, and offer a loosely-coupled approach to the management of the interface between the physical world of Remote Condition Monitoring (RCM), and the largely virtual space occupied by traffic management.

The wider use of open and crowd-sourced datasets in support of railway operations, in particular to deliver improved situational awareness where industry-owned data does not exist or systems are operating in degraded modes, was also an important theme. The use of this data could, quite naturally, take many forms, and so while the most immediate possibilities include support for infrastructure mapping (particularly across transport modes) via OpenRailwayMap, or end-to-end ticketing (a topic already being addressed under Shift2Rail, and initially in the IT2RAIL lighthouse project), this document is focussing on the use of social media data to support operations, and this is considered in section 4.

2. Software Architectures for the Rail Industry

The design of software systems and associated architectures used within the railways of Europe have, in common with many other infrastructure-led sectors, been dominated by the changing business environment the industry has faced in the last 30 years. The nationalised railways of the 1960s, 70s, and 80s, meant that the very early ICT systems used to manage railway operations were, quite naturally, developed to work in the context of vertically-integrated railway systems. As the railways began to become privatised in the early 1990s, these vertically integrated architectural models persisted for legacy systems (and updates to them), but in general new ICT, and particularly systems being developed for information exchange over the web, began to be specified in-house by individual stakeholders; a move that led to a proliferation of similar ICT platforms, and the formation of a complex web of interconnected platforms.

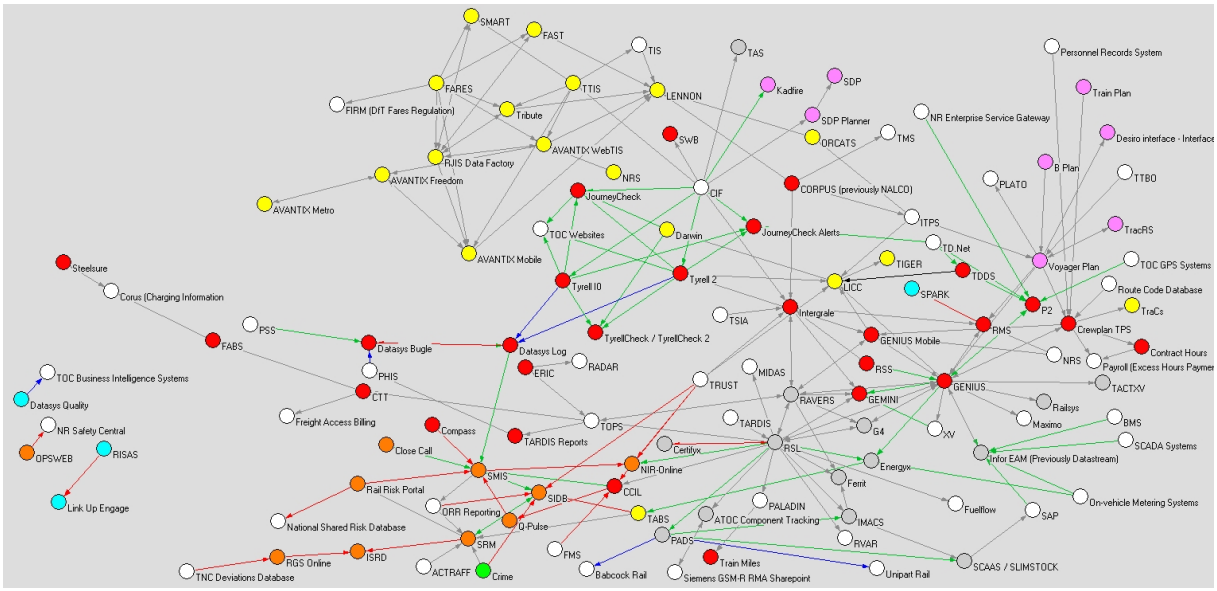


Figure 2-1 Shared ICT platforms, not operated by the infrastructure manager, in use in a typical European railway

Figure 2-1, developed by the C4R SP3 team, shows the (data) interactions between a set of ICT systems within a typical European railway system, in this case the UK railways. The systems and interfaces shown were drawn from a curated catalogue of shared ICT systems within the UK rail industry, which is managed by the Rail Safety and Standards Board (RSSB)(RSSB, 2011); specifically, the systems included are those that are shared between stakeholders but are not owned / managed by Network Rail, this means that very few of the systems that existed under the vertically-integrated railway of the late 80s are shown (this greatly simplifies the number of interconnections, and makes the diagram more tractable when presented in 2D). The colour-coding is captured in Table 2-1.

TABLE 2-1 KEY TO COLOUR-CODING IN FIGURE 2-1

Vertices		Arcs	
Crime and security	Green	**Manual data input interface**	Red
Customer / commercial inc. PIS and ticketing	Yellow	**File or report download / upload interface**	Blue
Operations inc. train performance, control logs etc.	Red	**Electronic data interface**	Green
Planning	Purple	**Interface type unspecified**	Grey
Rolling stock systems	Grey		
Safety systems	Orange		
Supply chain, strategy and R&D	Cyan		

The figure illustrates the extent of the architectural issues facing the European rail industry, in that even where systems are largely integrated via electronic interfaces (as is the case between most of the operational systems, shown as red vertices), a large number of bespoke end-to-end interfaces exist within the system. Over time, this type of architectural arrangement becomes increasingly inefficient, with the order of n^2 unidirectional interfaces being needed to completely link n systems, resulting in high maintenance costs, and a poor level of understanding of 2^{nd} degree impacts of changes to the system. The issue of interface proliferation in large scale systems is well known in the ICT industry, and as a result an alternative paradigm to the use of end-to-end interfaces has been under development for a number of years; this will be discussed in the following section.

2.1 SERVICE ORIENTATION AND THE ENTERPRISE SERVICE BUS MODEL

2.1.1 AN INTRODUCTION TO SERVICE ORIENTATION

Service Orientation is the idea that the ICT systems used by a business should be designed in a way that reflect the processes used within an organisation, rather than the organisation having to change the way it does business in order to fit better with the ICT infrastructure it operates. The use of ICT architectures based on this concept (and hence known as Service Oriented Architectures (SOA)) promotes the easy reuse of code, straightforward software maintenance, and use of predictable, repeatable ICT practices across a business.

Hurwitz et al. (Hurwitz, Bloor, Kaufman, & Halper, 2009) define SOA as “a software architecture for building applications that implement business processes or services by using a set of loosely coupled, black-box components orchestrated to deliver a well-defined level of service.” It’s important to note that this description is very much from a business, rather than a technical perspective. While in implementation a SOA consists of a messaging bus that passes information between a collection of

software services as defined by a workflow, the most effective way to think about SOA is as a tool that enables an organisation to perform its business processes. Adopting SOA within a business is about far more than buying into a new set of software tools; it's about rethinking the core functions of the business and how it should perform them, it's about governance and putting policies in place that define how information resources within the company should be used and by whom, it's about understanding the business's software and information assets and how added value could be generated from them, and once all those things are in place it's then about engineering a technical solution.

At the heart of any SOA implementation is the idea of a business service; a complete, self-contained set of processes and information that result in the performance of a business function. In the case of a bank for example, business services may include the opening of new accounts, the selling of mortgages, the calculation of monthly interest, and determining when to call the bailiffs in if it's all gone wrong. By structuring an organisation's ICT around business functions rather than other more "convenient" software patterns, SOA allows complex business applications to be created quickly and easily by coupling individual business services together; the coupling may be based either on predefined workflows or be performed dynamically based on the task the user is performing. As business practices change over time, the SOA applications can be easily changed by swapping old business services for new ones, and existing business services can be updated with new, better-performing code without having to rewrite whole applications.

2.1.2 KEY COMPONENTS OF A SERVICE ORIENTED ARCHITECTURE

By definition, most components of a SOA are loosely-coupled and able to exist largely independently of each other. There are, however, a number of key components that should be part of any SOA if it is to actually work effectively as a complete system.

The central component of any SOA is the Enterprise Service Bus (ESB); the ESB is a collection of software services that facilitate communication between the other components of the SOA, monitor the performance of the SOA with respect to any service level agreements that are in place, provide appropriate security measures to prevent unauthorised access to data or services, and log SOA activity for auditing purposes. Supporting the ESB in this task are two other components, the Registry and the Repository; these keep track of the identity and functionality each service, its data needs and outputs, the rules associated with its use, and the version history. In general terms, the registry can be thought of as the "phone book" for the SOA, enabling service discovery. The repository on the other hand is more like a file room; it contains the metadata necessary for describing key business processes, information on how SOA components link together to perform those functions, the rules governing the use of services within the SOA, the levels of performance to be delivered by each component, and details of who is responsible for the change management of a component or process.

While the ESB, registry and repository make it possible for services to be found in the SOA and facilitate communication between them, additional functionality is needed to execute the services that make up a complete business process / workflow and orchestrate their activities. The Business Process Orchestration Manager, Service Broker and Service Manager fulfil this role, enabling end-to-end ordering / execution of services, identification of service instances via the registry, and quality of service (QoS) / tracking of failures for the workflow respectively. In practice, many SOA implementations integrate the functionality of some or all of these components into the ESB.

2.1.3 SERVICE ORIENTATION IN THE RAILWAYS

InteGRail and ON-TIME

Service Orientation is not a new concept in the railway industry. Early in 2005, the EU Framework 6 project InteGRail proposed the use of an ESB called the InteGRail Service Grid (see Figure 2-2) as a core component of their data integration platform for the railways. The InteGRail Information System (IGRIS) worked in conjunction with a primitive ontology to deliver a flexible platform for information management across systems and stakeholders within the industry, and across a range of domains including operations, planning, and RCM.

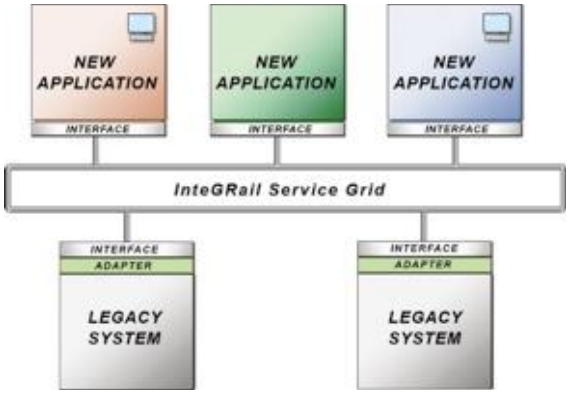


FIGURE 2-2 THE SERVICE GRID PROPOSED BY THE INTEGRAIL CONSORTIUM (INTEGRAIL CONSORTIUM, 2009)

While the InteGRail consortium's work demonstrated the potential for both the application of ESB and semantic data models within the industry, the technological solutions developed were ahead of their time, and difficult to implement in an industrial context. Despite this, the project was successful in providing proof-of-concept, and in recent years the industry has been slowly migrating towards this type of architectural solution, particularly in the research arena.

Between 2011 and 2014, the EU Framework 7 project ON-TIME picked up the core ideas of service orientation from InteGRail, and built an SOA for railway operations, with specific demonstration scenarios in the area of disruption management. The technical details of the architecture produced

can be found in ON-TIME project deliverable D7.2 (ON-TIME Consortium, 2013b), but are summarised here for completeness.

The overall structure of the ON-TIME integration layer (also known as the event-based architecture) can be seen in Figure 2-3, along with the connecting services provided by each of the work package teams. The aim of the integration layer was to provide a decoupled mechanism for linking traffic management algorithms (shown towards the bottom of the figure) to a Traffic Management System / Train Control System; this would enable the decision-making modules to be updated frequently, while the safety-critical TMS and its associated connections to actuators in the physical infrastructure could remain untouched.

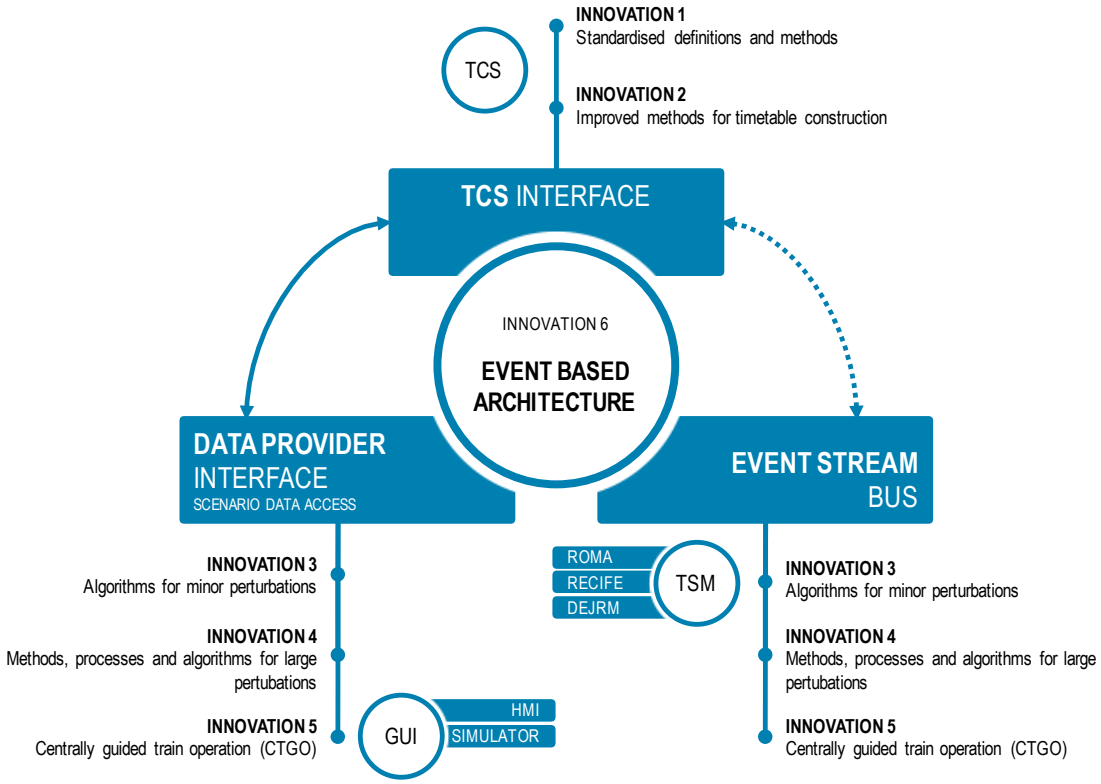


FIGURE 2-3 THE ON-TIME PROJECT INTEGRATION FRAMEWORK (ON-TIME CONSORTIUM, 2014)

The ON-TIME architecture, produced by WP7 of the project, was based around two core elements. An ESB, created by NTT Data and based on the well-known RabbitMQ messaging framework, and the project data model, an extended version of railML 2 with extensions to handle interlockings and Real Time Traffic Plans (RTTP). The message bus system used by the project can be seen in detail in Figure 2-4, it consisted of a straight-forward publish / subscribe system for message handling, which allowed any number of modules to receive data published to the bus, and also to contribute new results / traffic management decisions as available (and critically without blocking other functionality).

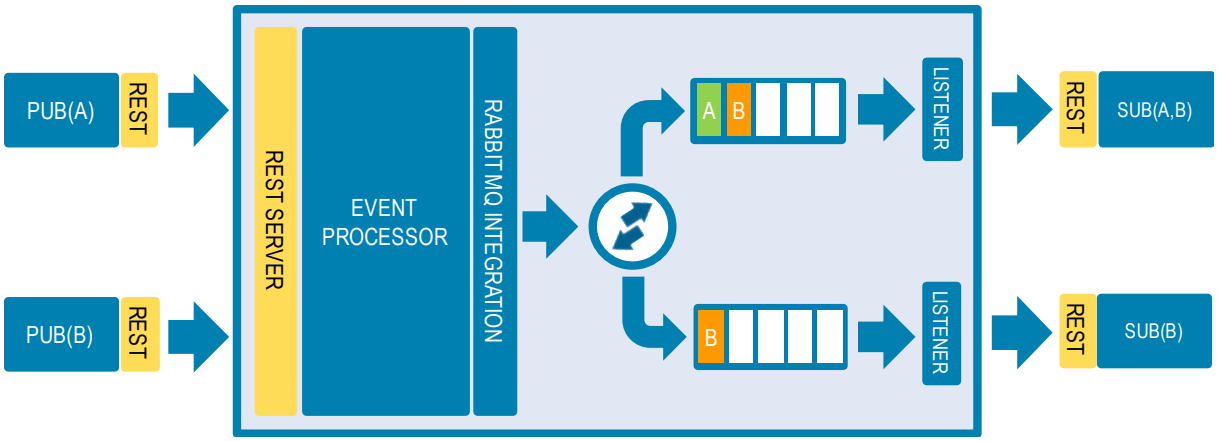


FIGURE 2-4 THE ON-TIME ESB AND DATA SUBSCRIPTION SYSTEM

The ESB was supported by an extended version of the railML 2 model, custom-created for the project. The data concepts required for traffic management were mapped out in a data dictionary, and this was then annotated to indicate whether the concepts were present in the standard railML 2 models. An illustration of this is shown in Figure 2-5. Individual concepts were annotated and cross-linked to indicate their hierarchical / compositional structure. This dictionary was then used to prioritise topics for the railML extension work with the ON-TIME project, and to inform the modelling process as it developed.

FIGURE 2-5 ILLUSTRATION OF THE ON-TIME DATA CONCEPTS (ON-TIME CONSORTIUM, 2013A)

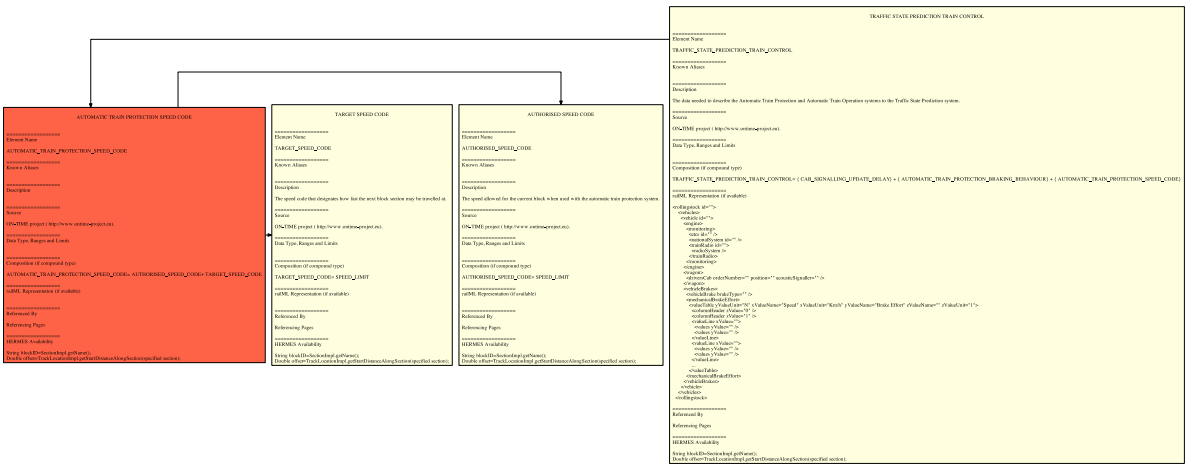


FIGURE 2-6 EXAMPLE OF COMPOUND DATA TYPE FOR THE "AUTOMATIC TRAIN SPEED PROTECTION CODE"

The ON-TIME data dictionary work was later updated by C4R team members to include ETCS messages, a concept not captured during ON-TIME. An example of a compound data type from the updated dictionary is shown in Figure 2-6. Further details on the ON-TIME integration layer and choice of data models can be found in project deliverables D7.2 (ON-TIME Consortium, 2013b) and D7.1 (ON-TIME Consortium, 2013a) respectively.

ESBs in the Industry's Research Vision

The ESB architectural model is becoming common in the software industry, and in the time since the completion of ON-TIME it is increasingly being proposed for large-scale railway projects. Of particular importance is the connection between this type of architecture and the Shift2Rail (S2R) Joint Undertaking's vision for the future of the industry, as captured by the Multi-Annual Action Plan (MAAP) (S2R, 2015), which cites the outcomes of the ON-TIME project as being of importance in this area. The S2R group includes many of the key suppliers in the European railway industry, and their support of the paradigm means that it is likely to be the basis for future standardisation efforts.

Other examples can be found in the national railway systems themselves, and in the case of the UK within the Digital Railway programme, which is responsible for the national roll-out of next-generation signalling and traffic management technologies. Over the past 2 years, the Digital Railway team have been developing an Enterprise Architecture that will serve as the basis for the functional vision of the railway moving forwards. A key component of that vision is a service bus for the industry, that will link both the individual ICT functions within the industry, and serve as a conduit to services from connecting modes as needed.

2.1.4 Software Architectures in SP3

Many of the members of the C4R SP3 team were previously involved with the ON-TIME project, and therefore, unsurprisingly feel that the architectural innovations developed during the project still represent the correct paradigm for systems of this type in the rail industry. This coupled with the industry support for the ESB model (as illustrated by the commitment to it in the S2R MAAP) means that the SP3 team have chosen to recommend that the core architectural component of an ICT platform for railway operations, should follow the ESB template presented by the ON-TIME consortium. However, in the short time since ON-TIME was active, a number of additional developments in the ICT domain have taken place, which must now be reconciled with the architecture proposed by the project team. It is these additional requirements that will be discussed in the rest of this document. The SP3 team are particularly keen to address the needs for semantic data integration, with a view towards interactions between the railways and multimodal transport systems, and to the inclusion of non-TMS data within the traffic management system. This work, and a proposal for a fact-based data management architecture to work alongside it (in conjunction with the ESB handling the core traffic), is presented in section 3 of the report. The team were also keen to investigate how open data resources, either coupled to the railway ICT architectures via linked open data approaches or by more traditional means, could add value to the industry. This work is presented in section 4 of the report. An indicative relationship between the elements is shown in Figure 2-7

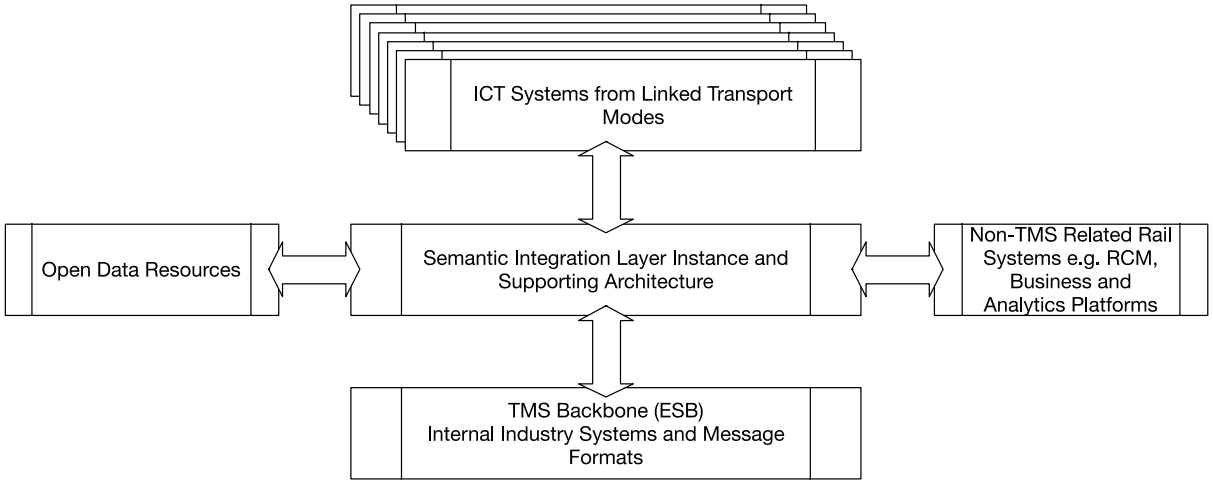


Figure 2-7 Proposed relationship between ESB / rail critical systems, semantic integration layer, and non-critical / third party systems

3. ONTOLOGY AND THE RETENTION OF DATA CONTEXT IN DECOUPLED ARCHITECTURES

The architectures presented in the previous section greatly improve the maintainability, extensibility, and lifecycle costs of large ICT deployments, and have the potential to bring major benefits to the industry; however, reducing coupling between data stores and the points of usage of data also generates new challenges. In traditional ICT architectures, the context surrounding an item of data is inherently captured by the manner in which it is stored, be that its position in a database table or file, or the software package from which it has originated. Because all the links in these systems are defined end-to-end by developers, there is a level of assurance that the data is being used correctly, the developer has looked at some documentation and decided, based on the provenance of the data, that it is appropriate for the task at hand.

As data becomes decoupled from its originating system however, be that via message bus type architectures, or because it is stored in a NoSQL-type data store (e.g. in Hadoop), the contextual information becomes increasingly hard to establish, with implications for the trustworthiness of the data (a lack of understanding of the provenance), as well as the meaning of the data (its semantic).

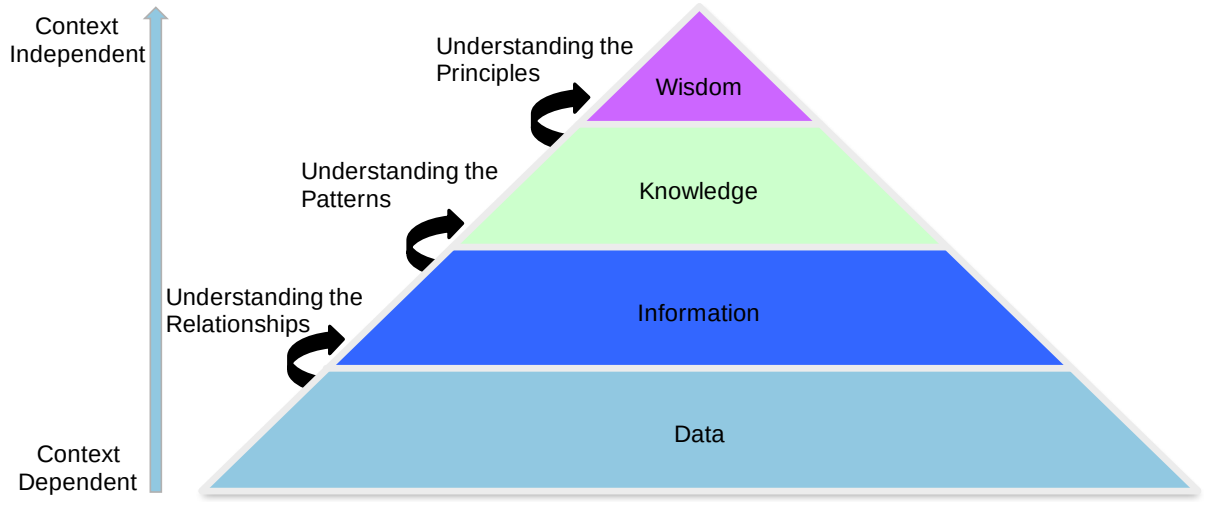


FIGURE 3-1 DIAGRAM SHOWING THE IMPORTANCE OF CONTEXT TO THE INTERPRETATION AND USAGE OF DATA

Figure 3-1 shows the importance of data context to the interpretation / integration of information and the inference of higher level relationships. At the data level, the context in which data is captured is crucial – it matters, for example, if a person's temperature has been taken using a thermometer or just estimated using a hand on the forehead, when deciding if that fact should be used in forming a diagnosis for their condition. As you progress up through the layers explicit context becomes less

important, as the patterns become transferrable between contexts, but it is vital that context is retained in the early stages.

As discussed in D3.4.1, ontology allows the context of an item of data, its meaning, to be maintained outside the scope of the originating system in a machine-interpretable form. Data is marked-up using a set of tags from a published, extensible view of the world. All models extend the same root concepts, allowing a common understanding of the tags to be shared at a high level. Stakeholders can create their own private models by extending concepts in public models, enabling the sharing of data at a high level, without necessarily requiring full disclosure of sensitive details. By drawing on other public ontologies, such as the W3C's PROV-O model, data provenance can be captured, enabling audit trails for individual data items to be created and carried with them across data stores, including information such as the origin of the data, the reason it was collected, the processes that have been applied to it, and any licensing restrictions that have been placed on it during its lifecycle.

The explicit inclusion of data context, along with the data itself, has a number of benefits besides ensuring that the data itself is properly understood. By including rules in the ontology model, it becomes possible to reason over the information, and to automatically infer new facts based on the information available, for example to convert between coordinate systems. By embedding this functionality in the model, developers can both ensure that the correct facts / processes are being used with the correct data, and enable easy updates to the rules used to infer the new information in a single location (the model) rather than in multiple, separate applications (see the railway case study in section 3.2.2 for an example of this in action).

The inherent capture of context allows a new flexibility in the way software interacts with data, Linked Data architectures allow data from one source to reference data from another, enriching the original records and allowing single sources of truth to develop. In the rail industry, such an architecture could draw on station identifiers in OSM data, bus stops in linked NaPTAN, vehicle movements from NR, passenger schedules from ATOC, bus times from National Express West Midlands to form an architecturally decoupled view of a complete journey from a mainline train to the passenger's destination.

In this section of D3.4.2, the SP3 team will show how ontology and the use of linked data offers value to the rail industry, ensuring that the data context that is lost as we move towards ESB and Big Data architectures can be retained, and how data can be moved out of application-defined silos. We will discuss the RaCoOn ontologies, a set of models developed specifically for the rail industry initially discussed in D3.4.1, and in particular consider an important practical aspect of the use of ontologies in industry – the establishment of a methodology by which the models may be created and curated. Next, we will discuss models for sharing sensor data using semantic markup, the W3C's semantic sensor network model, and its role as a bridge between RaCoOn and the monitoring technologies being developed in SP4. Finally, we'll look at a case study that shows a representative usage of an ontology

in the context of railway operations, and how the model can be used to avoid changes to legacy software systems during the transition between two different physical systems or during degrade mode operations.

3.1 Enabling ontology usage in a railway environment

Although the FP6 InteGRail project produced a valid proof of concept for the value of semantic models in the railway domain, the specialist knowledge required to use the complex, new (at the time) models effectively was a serious barrier to industrial implementation. If the technology was to be viable in a railway context, it needed to be demystified so that non-specialist ICT staff could use the models. In the period following the end of the InteGRail project a number of initiatives in the Oil and Gas domain, and several other non-industrial sectors, had shown that it was quite possible to build ontologies in a reliable, reproducible manner, and to use the resultant models in production systems. The following section, which builds on elements of those solutions, describes the process used to build the RaCoOn ontologies.

3.1.1 Design and implementation of domain models

A novel ontology engineering technique based on the NeON methodology (Suárez-Figueroa et al., 2012) was employed in designing the RaCoOn ontologies, based around extracting knowledge from existing railway models and domain experts to inform and validate design decisions. This technique comprised three major steps:

- **Specification**: High level requirements were defined, as well as the scope and content specification of system. Several individual ontology modules were defined according to reusability and level of domain detail: an “upper” module for domain-agnostic concepts, a “core” module for railway knowledge, and several subdomain-specific vocabularies including “infrastructure” and “rolling stock”;
- **Conceptualisation, formalisation and implementation**: Both top-down and re-use oriented approaches were taken in eliciting knowledge for the RaCoOn ontologies, as detailed below;
- **Evaluation and documentation**: Ontology modules were evaluated throughout the design process and then validated at the end of the design process.

Specification and modularisation of domain ontologies

Design of application-specific data models is usually driven by a set of functional and non-functional requirements that can be derived from the established needs of the system. Domain models such as the RaCoOn ontologies, however, are intentionally abstracted from any one particular application and are expected to allow representation of concepts without assuming how the data will later be used. The scope of the RaCoOn ontologies was dictated by three initial use cases: an infrastructure visualisation tool, a railway maintenance application, and a signalling design interchange tool.

Requirements for these use cases were considered in conjunction with applications and data requirements elicited from recent rail industry data workshops (Roberts et al., 2011), and a high level specification for the RaCoOn ontologies created emphasising commonalities between these use cases.

Conceptualisation and formalisation of RaCoOn ontologies

Each ontology module was created by repeatedly iterating over two approaches to model creation: a "top down" method that draws upon expert knowledge to build a hierarchical model of a domain, and a "reuse-oriented" method where existing knowledge was extracted from models such as RailML, Network Rail's Signalling Data Exchange Format (SDEF), and Siemens Rail Automation's Layout Description Language (LDL). In both cases, ontology implementation was performed by defining **ontology design patterns** (ODPs): sets of concepts, relationships, and documentation that define how a particular concept should be encoded in the semantic data model. Figure 3-2 shows steps through the iterative process.

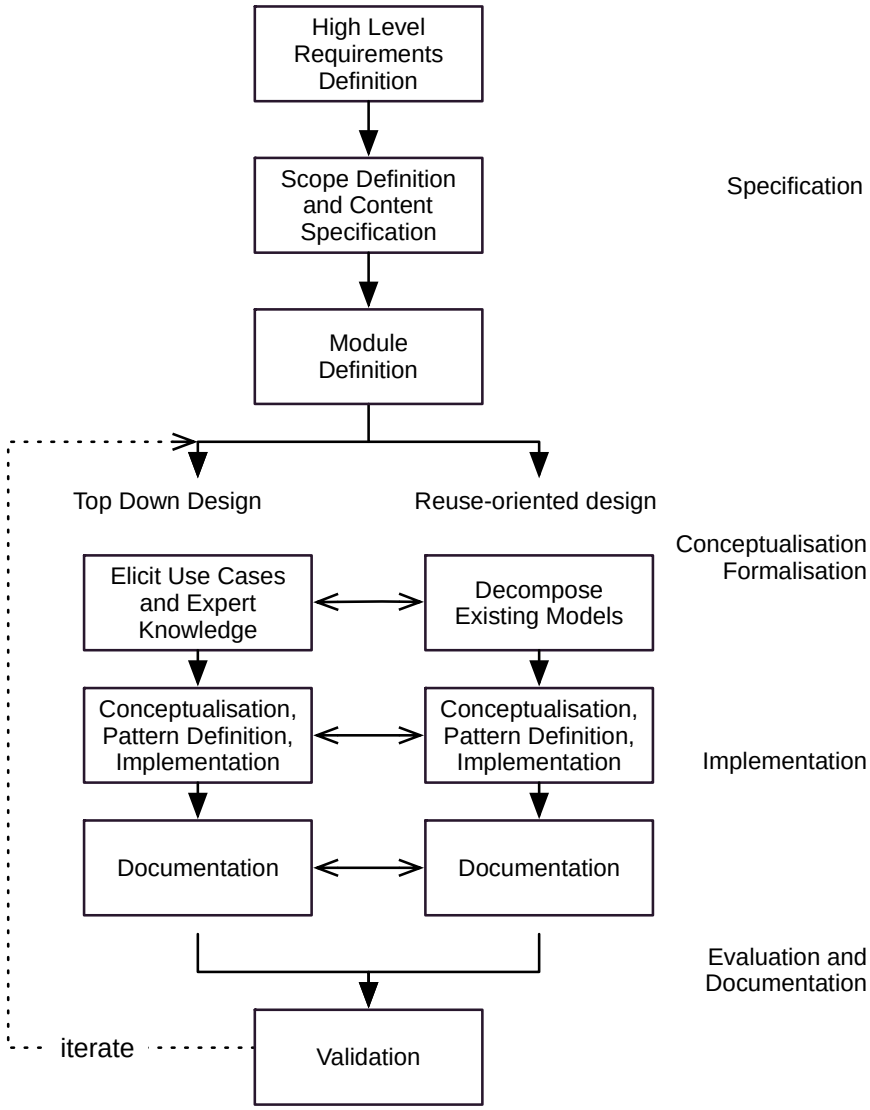


FIGURE 3-2 BLOCK DIAGRAM SHOWING KEY FEATURES OF DESIGN PROCESS

The top-down approach aimed to establish a high quality meta-model structure for railway domain knowledge, and to fill gaps in knowledge that may be present when re-using other models. The process performed was as follows:

- **Review scope of initial ontology (or changes for review)**;
- **Decompose concepts into subcategories, and create competency questions (CQs) around new concepts**. For example, when considering a "railway track" entity, a competency question may be: "How can we establish whether a piece of railway track is electrified, and what type of electrification does it provide?";

- **Consider scope of new CQs**. A decision on whether they are in or out of scope for the current module is made, and in scope CQs are either implemented or constructed using the reuse-oriented approach;
- **Re-engineer concept into OWL design pattern** if appropriate.

The reuse-oriented approach was undertaken to map existing domain knowledge from non-ontological sources into the ontology:

- **Identify terms for reuse** through prompts from previous iterations of this or the top-down process;
- **Analyse term semantics** by reviewing documentation and use of a term in the existing model;
- **Re-engineer term into OWL design pattern** by either reusing or extending an existing pattern, or creating a new one;
- **Consider new competency questions** based on term and design pattern.

Figure 3-3 shows decisions made in the creation of an example ontology using this process. New ODPs are shown in the diagram as red stars.

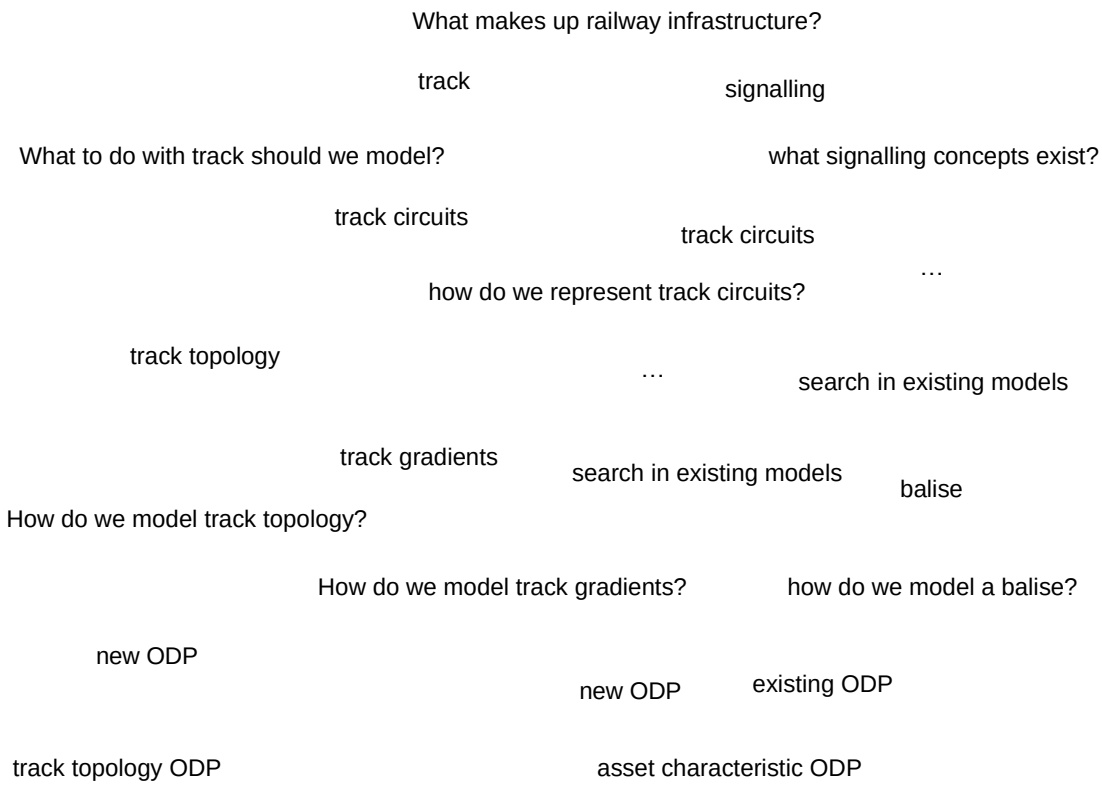


Figure 3-3 Example competency questions and paths to ontology creation

3.1.2 Upper level concepts and extensibility of models

The RaCoOn upper level ontology contains knowledge of generic upper level concepts that transcend the railway domain. Such concepts include space and time, and are mostly reused from existing "gold standard" vocabularies, including:

- The W3C Time Ontology (W3C, 2006), which provides ways of representing instants, intervals, and Allen time relations (Allen, 1984). Entities are labelled with start and end times where required, allowing data to be queried based on the time period in which it occurred.
- The W3C Geo (Brickley, 2003) and Ordnance Survey (OS) Spatial Relations (Ordnance Survey, 2014) ontologies for location positioning.
- The National Aeronautics and Space Administration (NASA) Quantities, Units, Dimensions and Types (QUDT) ontology (Hodgson and Keller, 2011) provides an exhaustive list of quantities, units, dimensions and datatypes. These are used in the upper ontology in conjunction with an appropriate design pattern to represent measurements and datatypes.
- ISO15926:2 (ISO, 2003b), which provides a meta-model for entity types. The ontology classifies objects into independent (can exist in their own right), and dependent (existence depends on another entity, such as in the case of a measurement), which is useful in defining acceptable ranges and domains for properties.

The rail core vocabulary ontology is a result of work carried out manually constructing and curating knowledge from other domain models and from UK industry experts. The vocabulary and its sub-modules predominantly draw upon corresponding elements in railML 2.2, relying on both its XML syntax and human-readable documentation in building an equivalent semantic data model.

3.1.3 Examples of open ontology models - observing the railway in real-time

An obvious example of where linked data can support the rail industry operationally, can be found in the linking of RCM data with traffic management processes. The use of sensor data, and in its more general sense asset status, within the field of railway operations / traffic management is still very much in its infancy. At present, the rail industry installs and uses sensors on assets primarily in relation to remote condition monitoring, and the models used in those systems reflect that role (for example the use of the Mimosa OSA-CBM / ISO 13374 standard (ISO, 2003a) by Network Rail within Intelligent Infrastructure). As automated assessment of asset status becomes a more important element of traffic management moving forwards, it will be necessary for increasingly context-rich descriptions of sensors to be used, enabling the software systems to use data from a wide range of assets and sensors with the appropriate business logic to derive actionable asset status information. This section of the document, will attempt to show how context-rich data models can bridge this gap, enabling data from

a diverse set of sensors to be handled in the same way by a receiving system (such as a traffic management platform).

Sensor Web Enablement

The Open Geospatial Consortium's (OGC) Sensor Web Enablement (SWE) framework (Botts et al., 2006) is designed to enable developers to make sensors and sensor data repositories available online. The framework, which is backed by over 300 companies and research organisations worldwide, covers all the main aspects of sensor data collection and delivery, including specifications for open interfaces, sensor service descriptions, feasibility planning for sensor installations, and driver management, however it is the sensor data processing and observation models that are most relevant to operational applications. The SWE framework is also compatible with a range of other models that are made available by the same consortium, including the Geography Markup Language (GML) and IndoorGML specifications, the Location Services (OpenLS) specifications for developing location-based software applications, and GeoSPARQL, a query language for accessing geospatial data via the Semantic Web. Within the SWE framework, the description of sensors and sensor data are handled by the SensorML and Observation and Measurement XML models, although the O&M model is of limited applicability if using simple encodings (e.g. comma separated text or similar).

The SensorML model is designed to allow the description of the processes of data collection and transformation, including the description of any sensors and actuators that are involved in the process. From a railway perspective, this can be thought of as a description of a crossing barrier, sensors attached to the asset such as a current clamp on the barrier motor, and the processing performed on the current waveform such as down sampling of the data and baseline adjustment. As with many XML-based standards, SensorML includes a certain amount of flexibility in terms of the information that must be included in a valid file, and as a result can be expressed in a very compact form if required, although this comes at the loss of contextual information about the data.

One very convenient feature of the SensorML model is that it includes native support for describing sensor data that is accessed via resources remote to the file, via web services or similar), this provides a useful way of dividing raw sensor data values from the description of the sensor system that is generating them, and hence (with an appropriate choice of encodings) allows very effective use of bandwidth once the system configuration itself has been described.

Semantic Web for Earth and Environmental Terminology

The Semantic Web for Earth and Environmental Terminology (SWEET) ontologies (Raskin and Pan, 2003) are designed to provide an upper-level ontology model for earth and environmental science work. Developed by the NASA Jet Propulsion Lab in Pasadena, the models enable the representation of natural phenomena, human activities, and most importantly from the perspective of the In2Rail project, the data that is used to describe them (processes, states, and observations).

SWEET is based around ontologies, conceptual models of a world-space that were developed to inherently capture the context of data items for use on the Semantic Web. Ontologies are most easily thought of as data models that can allow a machine to make reasoned inferences in much the same way as a human, in the rail domain early examples of this could be seen in the framework 6 INTEGRAIL project (InteGRail Consortium, 2011), which used ontology as the basis for automatic network statement checking, as well as inference of vehicle status in condition monitoring (e.g. inferring that a vehicle was faulty because one of its axles had a hot box etc. etc.). The additional contextual information also makes ontology-type models popular choices for representing metadata used by simpler, less expressive models – SWE for example uses elements both of its own ontology repository, and SWEET's, for metadata in its XML model.

As with SWE, SWEET can capture all the details of sensor configuration etc. that might be needed for decision making in an operational context. As an ontology, its use would also enable the use of automated reasoning approaches to infer the asset status directly from the sensor data, delivering a very clean architectural model in which the business rules were encapsulated in the data model rather than in the application logic.

Semantic Sensor Network

The Semantic Sensor Network (SSN) ontology (Compton et al., 2012) is the World Wide Web Consortium's (W3C) suggestion of how sensor data should be made available on the web, and as with SWEET is based around the context-rich ontology model family. The SSN model enables developers to describe sensors, the data they produce and how it is measured, and to identify the assets the sensors are installed on, as well as enabling the description of key supporting data, such as the remaining useful battery life of a remote sensor.

The SSN differs from SWEET in that the developers have adopted a more open root to the model, choosing to extend open concepts in other public models (notably the widely used DOLCE Ultralite model) rather than create their own representations of time etc. The use of common "upper level" ontologies means that data exchanged and integrated by the users of different domain models (e.g. ontologies for the rail and highway domains) are more likely to be directly comparable, however this is at the expense of a loss of "control" of the root concepts by the maintainers of the SSN ontology themselves.

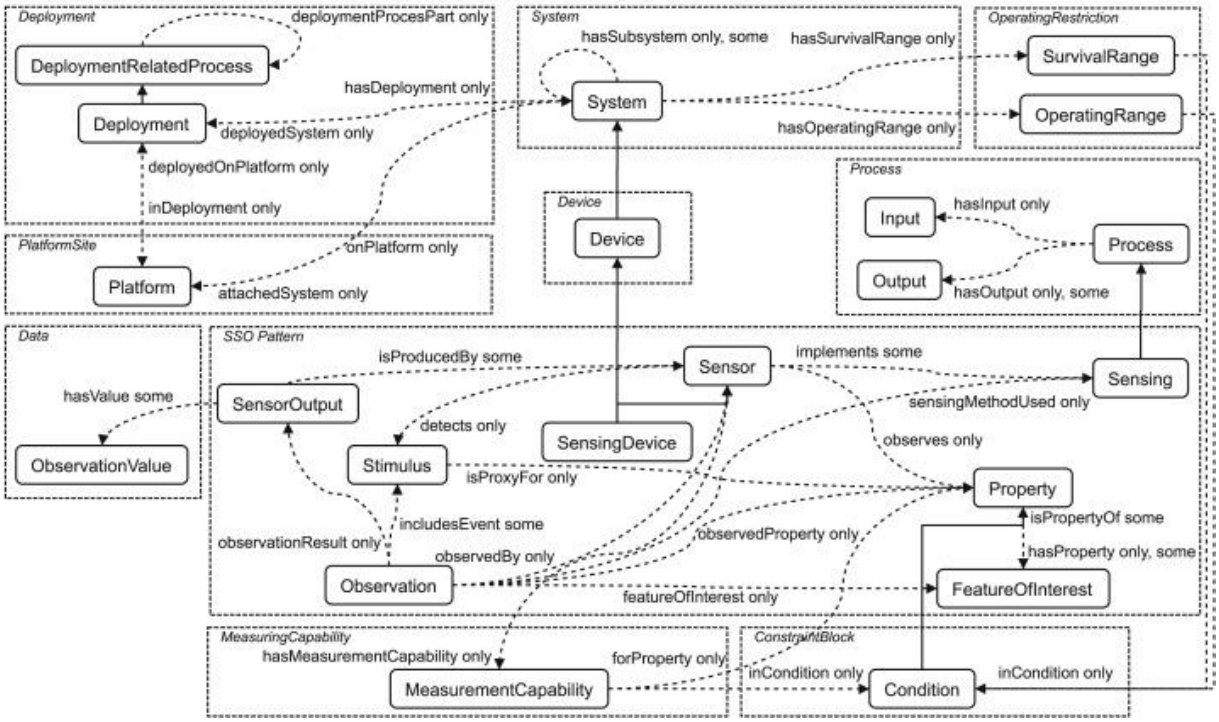


FIGURE 3-4 SENSOR DATA CONCEPTS IN THE SSN MODEL (COMPTON ET AL., 2012)

3.2 AN ONTOLOGY-BASED ARCHITECTURE FOR UBIQUITOUS RAIL DATA

In this section, we present how an ontology-based architecture can be used in a representative rail industry case-study. We present a candidate architecture for the management of data as facts (in the form of RDF triples) that can then be fed to / from an ESB as needed, in section 3.2.1 we then introduce examples of how ontologies and ontology reasoning can be used to automatically infer key pieces of information used in the operational domain, finally, in section 3.2.2 we demonstrate how the architecture can be used to deal with the “dynamic data” environment that can be found in railways operating in degraded mode or during transition periods between technologies.

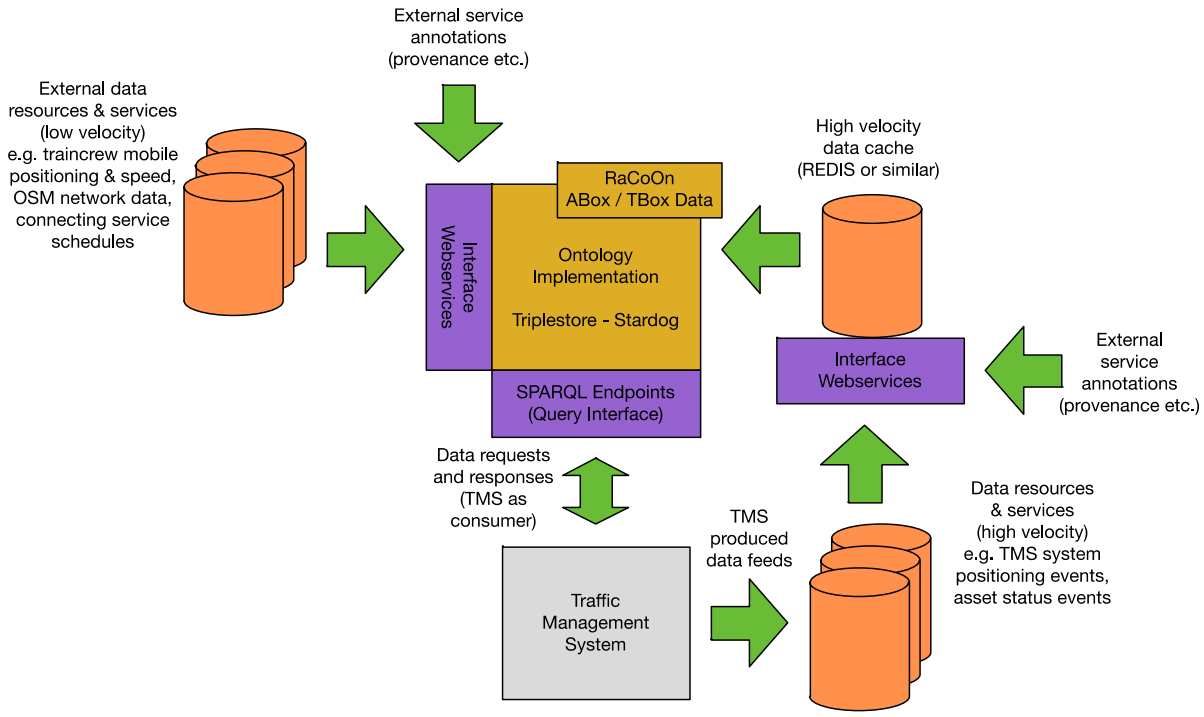


FIGURE 3-5 CANDIDATE ARCHITECTURE FOR ONTOLOGY DEPLOYMENT IN THE CONTEXT OF RAILWAY OPERATIONS

Figure 3-5 shows how RaCoOn might act as an interface between a TMS and other data resources. The RaCoOn ontology model is a set of concepts that describe the railway, relationships that link them to each other, and rules that describe transformations between data types. The implementation of RaCoOn, and the data it carries, is stored as triples – facts of the form subject-predicate-object e.g. "377 401" isA "Class 377" – that support computational reasoning, and these are held in a type of graph-based database known as a triple store. Data that is directly provided to the RaCoOn instance as one of these facts is referred to as T-Box data, while information that RaCoOn has inferred based on the facts it has available, and the rules and relationships contained within its model, are known as "asserted" or A-Box data. As new facts are added to the RaCoOn instance, COTS programmes called reasoners apply the rules within the model to the data available, and update the set of asserted facts accordingly. It is this ability, coupled with the inherent capture of metadata by the ontology, that allows the system to adapt to the data available at a given time, and to choose the "most appropriate" response to a query.

The RaCoOn instance itself is supported by a Linked Data Architecture, which from the perspective of TMS fulfils two main functions:

- Firstly, it helps to ease the integration of data external to TMS into the system, and to ensure that the most up-to-date information published by the data provider is being used as the basis for decision making. In a similar vein, by using links to data directly published by the provider,

rather than working on local copies, the architecture will helps ensure data consistency between the TMS and other tools / applications in use within the industry;

- Secondly, the architecture allows the ontology to refer to facts as needed rather than working directly with every update, which the use case being considered may not require. This is important because if the facts themselves are stored as part of the RaCoOn instance, every time a value changes the reasoner must be re-run across the whole set of facts available; this is a computationally-intensive process, particularly if the values change on a frequent basis. Through the use of a linked data architecture, the system will use a key-value pair database, which is optimised for high-throughput data, as a buffer / cache – enabling "just in time" updates to the values used for reasoning in the ontology triggered by data requests from the user, and greatly reducing the computational load on the system as a whole.

3.2.1 THEORETICAL EXAMPLES OF ONTOLOGY APPLICATION IN THE CONTEXT OF DEGRADED MODE RAILWAY OPERATIONS

From a service viewpoint, RaCoOn and a supporting Linked Data architecture could provide a number of vital services to a TMS, while simultaneously acting as a conduit for data resources with the potential to contribute to the tactical picture of the railway system available during degraded mode operations. These are the resolution of various types of train descriptors, and the provision of vehicle completeness checking without the need for physical devices running the length of trains.

Provision of vehicle completeness checking

Rather than providing physical vehicle completeness checks as a fallback in non-track circuit based systems, a combination of cab positions obtained from GPS, and a RaCoOn instance could be used to infer vehicle completeness. The ontology rules used for this process would be based on known unit lengths and the distances between the reported GPS fixes for the two cabs, and these could then be encoded in the RaCoOn instance, allowing them to be adjusted without impacting dependent applications. In situations where only one cab is present in a train, e.g. for freight services, a second position reporting unit be incorporated in the train tail light, allowing the same function to be used.

Where multiple units are coupled to form a single train, the system could make a further inference of connection based on cabs moving off in close proximity from known operational locations (e.g. station platforms) where coupling is likely to have taken place. These locations could be defined using rules in the RaCoOn instance, and would be modifiable, ensuring that these capabilities can be updated as needed without affecting the functionality of applications dependent on the vehicle completeness checks in any way.

Resolution of train describer data

Understanding the formation of existing trains upon entry into degraded mode operations, and being able to resolve the various service and vehicle identifiers used to describe them in data resources such as the working timetable is a serious operational issue in many railway systems. In an ontology-based system this reconciliation between the identification systems could be handled using rules in an always-on instance of the ontology model, which would draw together information on the days services, to relate them to the physical vehicles on the network. The rules included in the RaCoOn instance that will drive this process could, at least in the first instance, be based on the process used by when resolving service identifiers with GSM-R equipment installed in vehicle cabs.

Improved tactical awareness

During degraded mode operations, where a variety of data resources of differing levels of trustworthiness may be in use to provide information on the state of the railway network, it is advantageous to be able to draw on inferred information resources that draw together multiple data streams to create a single “answer” with a greater confidence than any of the individual resources alone. Furthermore, the rapid pace of technology change seen in the consumer-led markets beyond the scope of traditional railway information systems means that a degree of flexibility is required in terms of the types of external information resources that may be available to support railway decision making when conventional data is unavailable or incomplete.

An example of this type of externally-driven function could lie in the use of positional data from train crew mobile phones as an additional verification of vehicle location – this could be used as a cross-check against the GPS-derived cab equipment positions (the train crew should not lie outside the envelope defined by these positions, given accuracy estimations based on the number of satellites used to obtain the position) during degraded mode. Furthermore, by drawing this information into the system during normal (rather than degraded mode) operation, the ontology instance could produce a more accurate vehicle position for use with passenger information systems in areas of the network protected by low-resolution technologies such as track circuits.

3.2.2 Representative example of ontology use in a dynamic data landscape

This section of the report aims to demonstrate the value of ontology in a dynamic data environment, such as would be experienced during degraded modes of operation. Specifically, it will show how applications using data via an ontology can continue to function despite particular data resources becoming unavailable. The work is presented in the context of a theoretical Real Time Passenger Information (RTPI) system.

Aim

The case study aimed to show how the use of ontology and linked data can help the industry maximise on investment in existing information systems despite changes elsewhere in an increasingly technology-driven railway system. In particular, it set out to show how the use of ontology can provide a bridge between legacy systems and newer replacement services without sacrificing functionality, and how interfaces between such legacy systems and more contemporary linked data-based systems can be set up. As the volumes and variety of data gathered in new information systems on the railway continue to increase, this demonstrator seeks to illustrate the practical uses of semantic data models in simplifying interfaces and applications, and enriching content.

Scenario

The storyboard for the system was as follows:

1. Imagine a railway network equipped with legacy, low resolution train positioning systems, such as track circuits and axle counters. These devices are placed close enough together to drive signalling systems but only provide a low resolution view of where trains are located across a network.
2. The data produced by the train positioning systems is used to (amongst other things) power a number of passenger information systems, including platform boards and third-party applications for mobile devices.
3. As part of an upgrade programme, for example a migration to European Rail Traffic Management System (ERTMS), existing low resolution train positioning equipment on a line is replaced by a more accurate system. Future passenger information systems can be designed to operate using the higher resolution positions from the new system, but existing passenger information systems, that require positional data to be at track circuit level, will all need updating - a costly process that involves many stakeholders if third-party applications are included.
4. In an information landscape utilising ontology, the data being delivered by the positioning systems, and being used by the passenger information systems, is described unambiguously; the computer “knows” exactly what data is available and what is needed by the applications. Rules can be added to the data model describing how data in one form is converted to the other, allowing the system to deliver inferred track circuit-level data to legacy systems based solely on the new, high resolution location data.
5. By using the combination of ontology, rules, and reasoning, it becomes possible to maintain the functionality of existing applications, despite changes elsewhere in the rail system, without altering the applications' codebase. Ontology will allow the industry to design and implement information systems only once in a changing technological landscape. Old and new applications will be able to co-exist and can be driven by the same underlying data resources.

Technology choices

The key implementation technologies used in in the demonstration are from the same stack as the example in Figure 3-5. Specifically, these are:

- Stardog, an RDF triple store used to store all data (ontology and resources). Stardog is a scalable off-the-shelf product that provides several ontology reasoners selected for a range of task types. It conforms to W3C standards on linked data storage and presentation, allowing a future-proof, generic interface between the application and data store to be created;
- REDIS, a key-value database optimised for use with high-velocity data. The REDIS component of the stack serves as a buffer for the RCO, preventing unnecessary re-running of the reasoners and reducing the computational load on the system;
- SPARQL, the SPARQL Protocol and RDF Query Language can be thought of as an SQL equivalent for linked data applications. It enables query of the RCO instance held in the triple store, and is used for operations on the RCO-side of the interface web services.

User interfaces were created using standard web development technologies – the Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), and Javascript. As part of a proving activity three scenario-driven interfaces were created:

- Legacy Departure Board System (Using Track Circuit Data). In this scenario, a user can select a train station and view a very basic simulation of a platform-based passenger information board, including departure point, destination location, scheduled, and expected times. Expected times are calculated based on the position of trains on a track circuit (such as would be provided by a train describer system), which is queried directly from the triple store. The current track circuit of each train can also be displayed for exploratory purposes;
- Train Position Map (Using Mileage Data). The train position map shows the “live” locations of each train on the network. The system queries the ontology for mileage location, and displays it in line with the train's route through the network. Through rule reasoning, the ontology provides the train position map with the most relevant data should both be available;
- Entity Information View (Using Linked Data and Inference). The final view is provided should a user want more information on a particular train, station, or location. The application requests information from the ontology about the location in question, and returns useful information. In the case of train services, inference provides information about the rolling stock itself as well as the train service; for locations, reasoning provides additional information such as touching/neighbouring entities and line reference information.

The information usage by the first two views is summarised in Table 3-1:

TABLE 3-1 INFORMATION USED IN SCENARIOS / VIEWS OF DATA LANDSCAPE

	Track Circuit Data	Mileage (Moving Block) Data

Legacy Departure Board System	Directly supplied "real" track circuit data used.	Inferred track circuits based on train mileage data used.
Train Position Map	Inferred (approximate) train location as mileage used generated based on known track circuit locations.	
Train Position Map (Both Location Sources Available, i.e. Redundancy of Resources at Different Levels of Granularity)	Rule reasoning used to choose the "preferred" location object for the task being performed.	

Design patterns and application of reasoning

Infrastructure and Location Storage Design Pattern

Infrastructure and location data is stored in the train locator demonstration model as linked data, following patterns designed by the core ontology. Data taken from ATOC working timetable files was used as a base for modelling train movements, and track circuits were added manually, using simulated track circuit distances. Each "Track Circuit" object included a start location and end location, and each of these locations had an associated mileage and Global Positioning System (GPS) co-ordinates, and these track circuits are aggregated into ":ServiceNode" objects that are referenced in timetable data. Figure 3-6 shows an example Service Node associated with a track circuit, which is in turn associated with maximum and minimum locations at points along the track infrastructure.

By linking track circuits to mileages and known pieces of infrastructure, inference can provide train services associated with them with further information. For example, in the case of a train stoppage or cancellation, passengers using linked-data based applications could check the next station's facilities and connections based on the train they are currently on, although this functionality is not shown in the demonstrator.

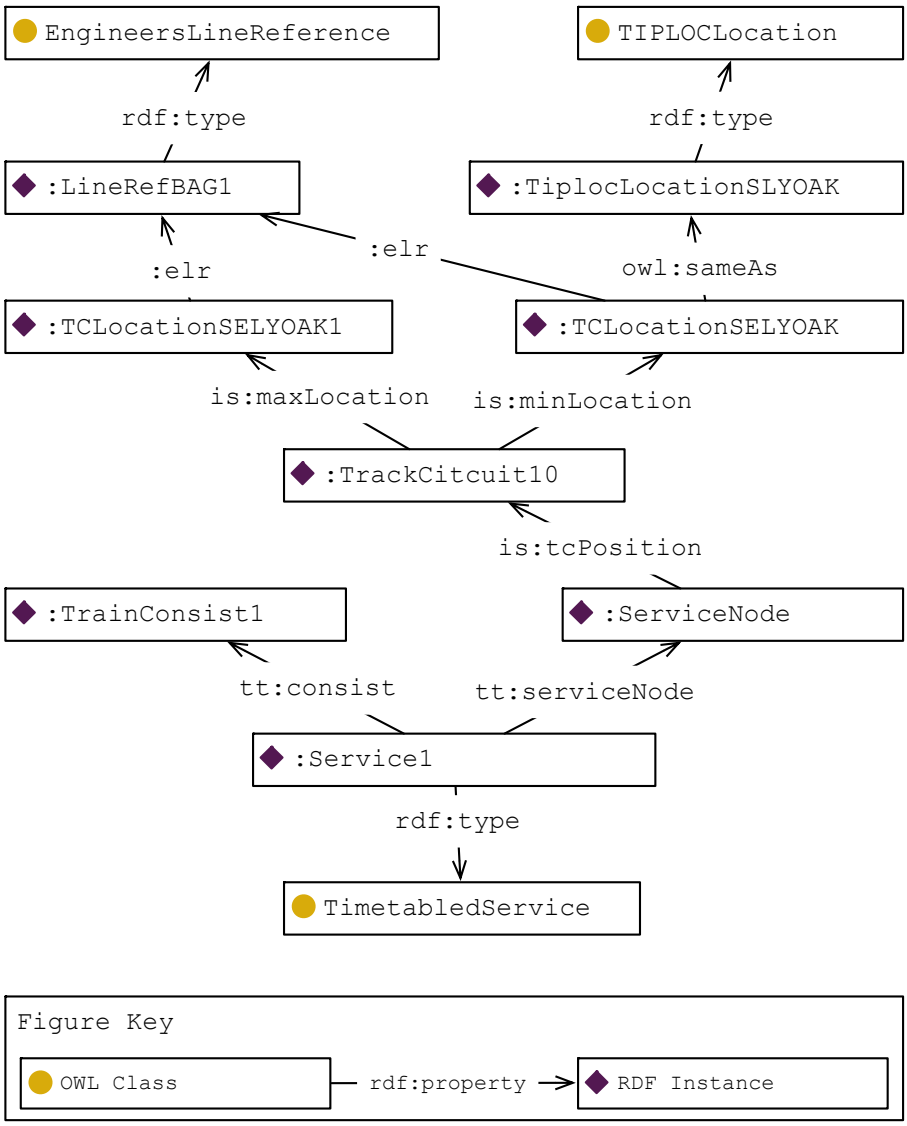


FIGURE 3-6 ONTOLOGY GRAPH SHOWING TRACK CIRCUIT POSITIONING

Reasoning to Allow Legacy System Functionality Given New System Input Data

In order to provide legacy system functionality when a system upgrade occurs, a rule is constructed and added to the triple store. Rules are custom-based reasoning patterns that a triple store applies to matching data at query-time. The aim is to capture the following knowledge:

> *"If a train's current mileage is between the minimum and maximum mileages of a particular track section, and on the same line, the train is defined as being in that track section"*

When encoded as a SPARQL rule, this logic led the reasoner to performed the following actions:

- Check for current node's line reference;
- Filter list of possible track circuits to only those on current line;
- Retrieve minimum and maximum mileages for each candidate match;
- Identify track circuits with mileages within range of current train's mileage;
- Assert that the current node is associated with the matching track circuit.

Consequently, whenever a legacy application now requests a node's track circuit location, this rule is checked and the correct track circuit returned whether it was encoded explicitly by an input system, or calculated based on a train's current mileage position.

Reasoning to allow improved resilience of information systems during degraded service

The strengths of an ontology-driven data store do not only allow the mapping of new data back into other forms for use in legacy systems, but also make it possible to increase data availability during periods of degraded system reliability. Using the capability of the system to interlink data, a hierarchy of “preferred” properties were specified for each system concept, and these hierarchies used with closed world rule reasoning to find the best available data for a particular application. Recall the following scenario from the storyboard:

- A railway line has recently been upgraded to ERTMS operation, and now provides very rich location information for each train on the track, rather than only track circuit occupation details;
- New applications for customer information and service monitoring are built using the new, more accurate ERTMS location information. It is desirable, however, for these systems to continue functioning in times of degraded operations - for instance if ERTMS systems are unavailable and the line reverts to fixed block operation.

In this case, the usual approach would be to include application logic to search for available systems and make a decision specified at system design time as to which data source to choose an approach which is in flexible and unsustainable in a complex system.

To enable the data model to find which data to provide for a train location application, the following pattern encodes knowledge of “preferred systems” (see Figure 3-7). This shows several OWL classes (marked with yellow circles) related to each other through RaCoOn properties (marked on arrows) forming transitive “:preferredOver” relations. Thus, a reasoner can infer that an “is:CrsLocation” instance is preferred to a “vocab:RailwayMileage”, and can choose to prioritise data of this type.

With this knowledge of which system of measurements is preferred given data availability, it is now possible to encode a rule that states:

> *"If entity X has multiple locations associated with it, and one is preferred (location Y) over the other (location Z), then insert a new fact: entity X - > preferredLocation - > location"*

As a result of the inclusion of these rules, systems utilising the ":preferredLocation" property will automatically be presented with the most accurate data for their needs. It is important to note that applications have varying requirements for location data (some rely on GPS co-ordinates, others rely on Computer Reservation System (CRS) codes, and others on data with other constraints). The pattern above does not ignore these constraints; they are represented through other clauses in the query.

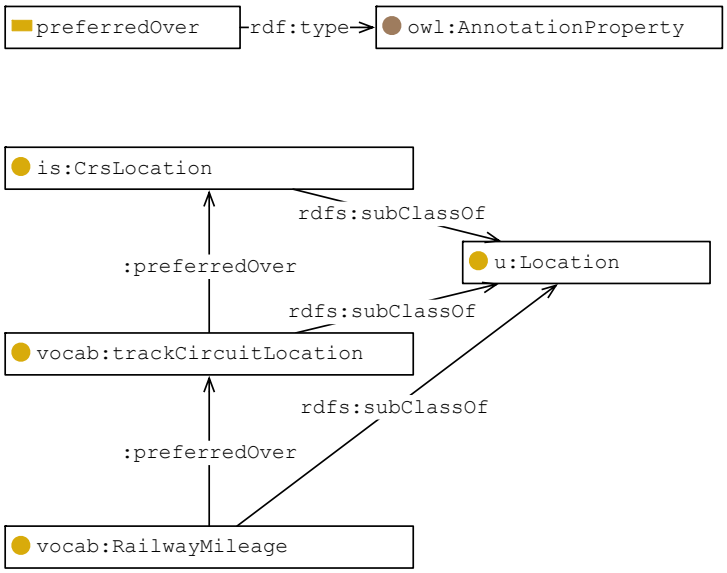


FIGURE 3-7 ONTOLOGY GRAPH SHOWING "PREFERREDOVER" RELATION BETWEEN LOCATION SOURCES

Implementation

The demonstrator web application includes several views which show the effect of reasoning based on location, as discussed above. These views are described in the following sections.

Admin page: scenario control

The Train Mapper home page, accessed when the user first contacts the system, briefly explains the aims of the demonstration and gives users control of the various scenario configuration options. These options influence the behaviour of both the legacy "Departure

Boards" view, and the "Map View" application. On-screen controls allow users to select the data supplied to the system by the simulator: either track circuit data, mileage-based position data, or both.

Further configuration options turn reasoning on and off within the web application, allowing users to see the effect with or without rules being triggered in the ontology.

Legacy departure boards view

The departure boards view (see Figure 3-8) shows trains soon to arrive and depart from a station. These are determined by querying the triple store for relevant services with an appropriate arrival time, and station information if present. Expected train times are naively obtained through adding a ":trainTime" property to every track circuit, and calculating the difference in this property's at the current train's location and the station being viewed.

If the "Track circuit data" data source is turned on, the departure boards view utilises no ontology reasoning whatsoever. Instead, it is presented as a legacy system using linked data as a data storage and interchange format. There are advantages even to this approach, as can be proven by the success and uptake of the Linked Open Data movement on the World Wide Web. If the "track circuit data" data source is missing, however, ontology reasoning steps in, and resolves live train locations to track circuits for the benefit of this application.

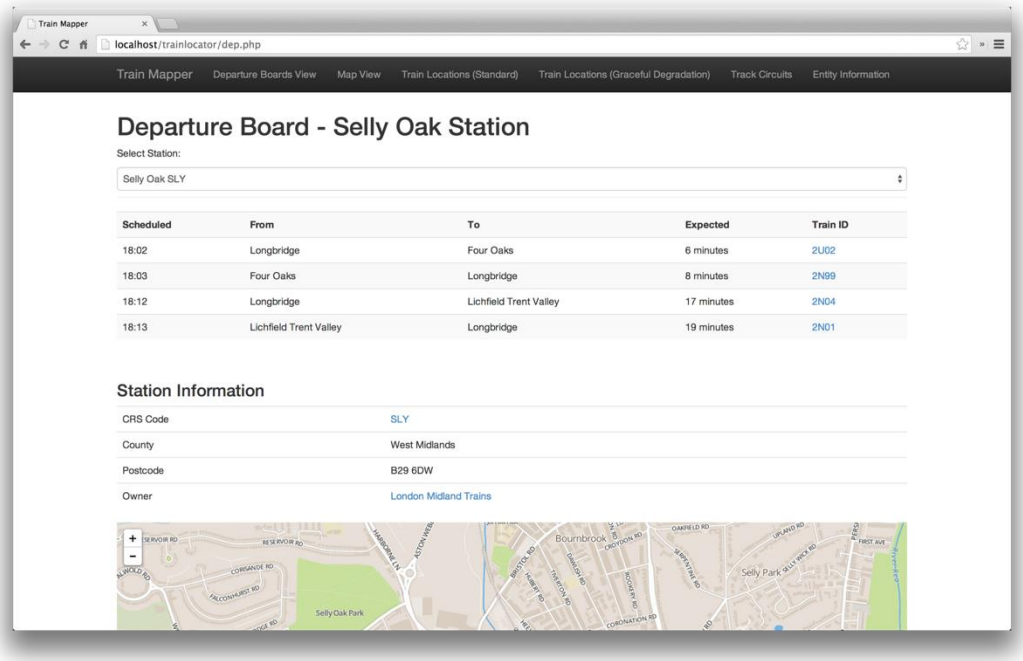


FIGURE 3-8 TRAIN LOCATION DEPARTURE BOARDS VIEW

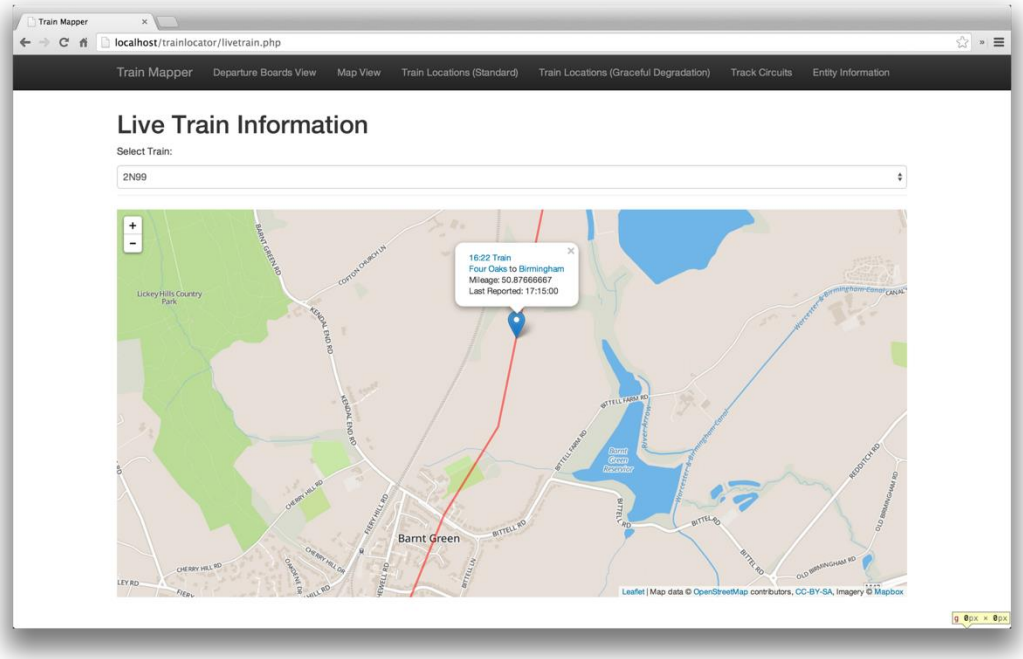


FIGURE 3-9 LIVE TRAIN INFORMATION MAP IN TRAIN LOCATOR

Map view: dynamic train progress

The dynamic train progress page (see Figure 3-9) allows a user to track the progress of a train in real time, using mileage values resolved from a fictitious moving block signalling system.

Users can select the train they want to track, and watch its position change across the map.

- With only the “mileage data” source turned on, this display uses no inference and displays the current mileage of the train selected on a map.
- With both mileage data and track circuit data, this display calls the ontology to ascertain the priority of these location values (as described above), and displays the mileage location, with its track circuit displayed as a secondary information source.

With only track circuit data available, the ontology resolves a less accurate position for the train based on available information. Whilst it would have been possible to build this logic into the application itself, this approach quickly becomes complicated and hard to maintain when deployed as part of a more complex system.

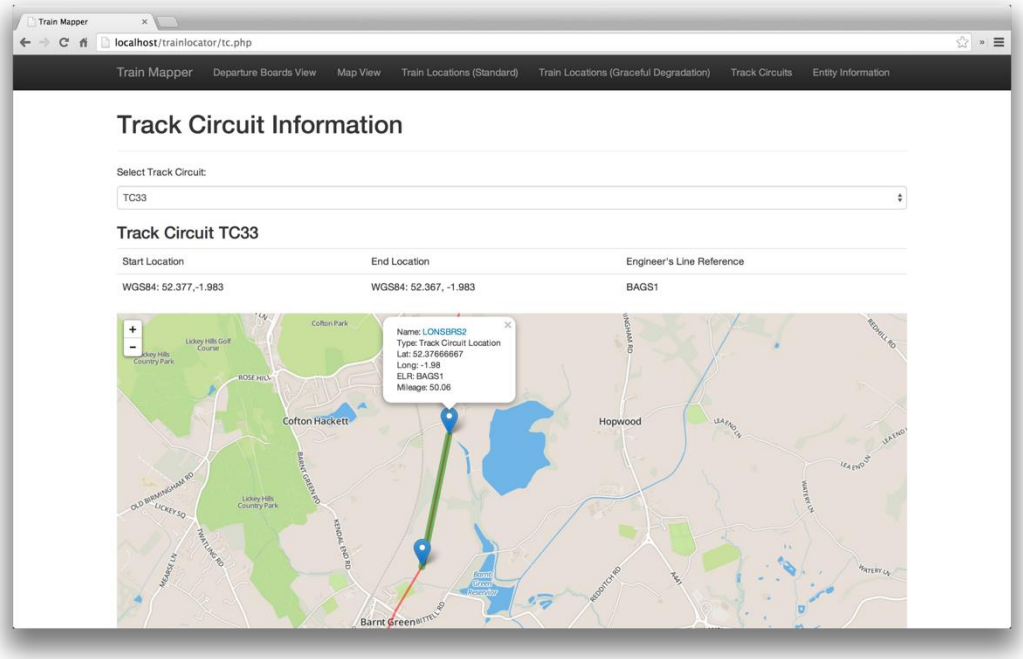


FIGURE 3-10 TRACK CIRCUIT DETAIL AND BOUNDARY OVERVIEW SCREENSHOT

Map view: track circuit information

Finally, the track circuit and entity views (see Figure 3-10) allow users to view more detailed information about each track circuit, or other entity. With reasoning disabled, queries used to populate this view bring back only explicit information held in the infrastructure database about track circuit information. However, with reasoning enabled, links between track circuit locations and other infrastructure items become apparent, and users are able to browse information about train stations, maintainers, and nearby trains. This view is included to further illustrate the use of ontology reasoning to enrich knowledge and convey useful inferred information.

3.3 COSTS AND POTENTIAL BENEFITS OF THE ONTOLOGY APPROACH IN THE CONTEXT OF A EUROPEAN RAILWAY

Industry-wide ontology models for rail have been the subject of significant discussion in the UK rail sector in recent years. The 2013 Network Rail Technical Strategy (Network Rail, 2013), which outlines the UK Infrastructure Manager's priorities for investment in new technology over the period 2014 - 2019 and beyond, suggests that developing the research into ontologies for rail to an “implementation ready” level (i.e. Technology Readiness Levels 5 - 7) would cost the industry up to £1 million; however, as is often the case in this area the document presents no indication of the value of the potential

benefits. An estimate for the financial benefits resulting from the implementation of ontology in the UK rail industry can be found by reference to other domains. As previously mentioned in this paper, the US National Institute of Science and Technology's cost analysis of inadequate interoperability in the capital facilities industry (Gallaher et al., 2004) found that $15.8 billion could have been saved in 2002 through improved information interoperability, a figure that represents between 1% and 2% of revenue for that year. The capital facilities industry, which deals with the construction and management of large commercial and industrial facilities, is similar to the UK rail industry in many respects; it consists of a large number of stakeholder organisations, each with their own ICT provision, which specialise in delivering infrastructure with a long lifecycle - as a result of this, the industry is an appropriate analogue to the railways. On this basis, taking the 1% to 2% revenue figure for the capital services industry and translating it into the UK rail industry, where the Train Operating Companies received fare revenues of £8.2 billion from passengers in the year 2013/2014 (ORR, 2015), results in between £82 million and £164 million of potential savings annually. If only a very small proportion of this figure were to be realised in practice by the industry through the use of a common data model such as the RaCoOn ontology, then the financial benefits would be very significant.

The design patterns and processes demonstrated in this paper have diverse applications across the railway; in particular, the demonstrator highlights a fundamental technique (the ability to utilise the most appropriate available resource of a given type) that can be implemented wherever multiple real-world systems provide the same type of information into a data store. Across the industry the ability to automatically select the most appropriate information resources for the selection available means that legacy software packages can still function in environments using upgraded information stores, maximising the useful lifetime and return on investment from these software packages. Furthermore, by moving the data dependency to the models and data repositories, rather than the applications, adopting the proposed design patterns will enable data-centric, rather than application-centric, management of the selection of appropriate information; reducing the complexity and cost of implementing business logic changes in software.

3.4 LIMITATIONS OF THE SEMANTIC APPROACH

The adaptation of common semantic models, such as RaCoOn, have many potential benefits to offer the railway industry. Care must be taken however, to avoid thinking of the technology as a `silver bullet' that will fit perfectly into every possible data exchange scenario. In enterprise contexts, OWL/RDF systems offer a pragmatic solution to the representation of domain semantics, however, there are limitations to the current implementation technologies as outlined in the following sections.

Scalability, Reasoning Performance, and Expressivity

Many of the benefits of using ontological models in information systems arise from their ability to infer new knowledge from existing data. Whilst some of this inference can be done in an efficient manner, much of the OWL DL language requires reasoning algorithms that do not scale to large volumes of data. A trade-off between reasoning performance and scalability is required, which currently prohibit many useful axioms being used in large applications. Ongoing research in `web-scale' reasoning techniques combined with state-of-the-art RDF graph storage technology is likely to bring increased performance in the future, but applying reasoning techniques over large datasets, represented using highly expressive OWL models, is currently a significant technical challenge.

Architecture and Distribution

Cross-enterprise data exchange is necessarily decentralised, and requires transmission and consumption of information between many systems and parties. While ontological models make it easy to refer to the same concepts universally, they do not address the practicalities of actually publishing and consuming information. Data sharing on the wider semantic web shares this issue: to make use of another dataset, one must either download it in its entirety to interact with locally, or rely on the data provider's processing power and availability using a SPARQL endpoint. Possible alternative architectural approaches to the data provisioning issue include the use of bespoke Service Oriented Architectures, and Linked Data Fragments, which use small, targeted data dumps to facilitate local querying of federated data. However, work remains to be done in this area before a `gold standard' architectural template can emerge.

Versioning and Change Control

Although ontological models afford a great deal of flexibility and backwards compatibility in their design and modification, it is still possible for changes to, or removal of, existing axioms from published ontologies to result in incompatibilities between systems. Web ontologies often present version-specific Internationalized Resource Identifiers (IRIs), so-called `Version IRIs' in addition to canonical IRIs to allow for users that wish to fix their application on one version of an ontology, but the problem of change management through a network of ontologies has yet to be formally addressed.

4. IMPROVING SITUATIONAL AWARENESS USING OPEN DATA RESOURCES

Throughout the WP3.4 activities the C4R SP3 team have maintained a strong focus on the importance of open data in delivering the reliable, high-capacity railway envisaged for 2050. Although open data resources of potentially unknown heritage will never entirely replace dedicated data harvested by railway undertakings as the basis for railway operations, there is an obvious and growing use case for the adoption of open data resources, used with appropriate safeguards, in order to supplement internal industry data and give operators and infrastructure managers improved situational awareness of the state of the network, particularly during disrupted operations or in the context of other connecting services that together deliver end-to-end journeys for the customer.

4.1 OPEN SITUATIONAL DATA

4.1.1 DRIVERS FOR GROWTH

A key driver in the growth in provision of timely data on the state of an individual's local environment has been the development of the smartphone. Since Apple's iPhone first launched in 2007 the growth in connected devices has been staggering, with the technology becoming indispensable for a wide demographics of users, and for a variety of uses from business to recreation. As an indicator of the rate of growth in the smartphone market in 2013 the number of connected devices in use was estimated to be approximately 3 billion (daCosta, 2013), a figure that had risen to just over 7 billion around 1 year later (GSMA, 2015)(Boren, 2014).

Typically, a consumer-grade connected device will have geolocation capabilities, accelerometers, and of course data connections providing the user with access to the web and a huge range of mobile applications in real-time, making it into an ideal local sensing platform for a wide range of applications, including giving transport providers a better understanding of how a customer moves through the network. These capabilities are also beginning to be used as the basis for other, more specialist applications within the industry, such as the monitoring of ride quality (Azzoug, 2017), although at the time of writing there is some debate as to the usefulness of such data at the resolutions that can be provided by the sensors in the phone, relative to specialist bogie-mounted or in-car solutions.

Smartphones have been instrumental in one of the most significant growth areas in near-real-time data contribution online by individuals – the growth of social media platforms. Social media represents a rich vein of information for transport providers, as for the first time passengers can comment on their journey / experience in real-time (see Figure 4-1 for an example of a chain of social media messages posted during a passenger journey), and do so in a pseudo-anonymised forum where they are likely to be less restrained in their commentary than might be the case on

conventional reporting forms. Aside from commenting (or more likely complaining) about their journeys, research has shown that many passengers look to social media channels for information on their journeys (Morris et al., 2010). Li et al. (2011) claim that about 11% of messages on one social media platform are questions, and 6% have actual information needs (i.e. are not rhetorical). Data collected by C4R team members corroborates this finding, with common queries asking about compensation / refunds, or directions between connections and final locations. Other topics discussed range from jokes to serious information or news including road, railway and traffic conditions, and computer scientists are already beginning to leverage such data to improve situational information (Rahman et al., 2012), including event/incident detection and monitoring or management (Bodnar et al., 2014). Active research is also taking place to improve predicting road traffic information (He et al., 2013).

The analysis of social media data is a classic Big Data problem, with the IT provider Oracle estimating that 15 million posts were being made a day as far back as June 2013 (Oracle, 2013). For the industry to leverage the content effectively, there will be a need in the coming years for significant further technical developments in large-scale interpretation of feeds, particularly in terms of sentiment analysis (the automated interpretation of the “feelings” expressed by the user posting the content), and high throughput identification of unique entities in near-real-time (for example railway services, specific locations, and differing perspectives on the same event, be those as a result of location, interpretation, or politics of the end user). Fortunately, the inclusion of social media-based intelligence in business operations is already nearing maturity in a number of other sectors, e.g. the utilities, and therefore the large ICT providers can be expected to continue to develop these capabilities without

significant specific investment needing to be made by the rail industry at this time (beyond the identification of key use cases and metrics to be used within the platform).

FIGURE 4-1 EXAMPLE OF A STRING OF SOCIAL MEDIA MESSAGES, POSTED BY A SINGLE USER OVER THE COURSE OF A JOURNEY

4.1.2 CHECKS AND LIMITATIONS ON USAGE

The growth of social media analysis has been held in check by a series of new challenges that the technology presents to industry, in particularly issues around anonymization, privacy, and trust. While most social media analysis work is restricted to content left open / public by users, data protection legislation means that, without the specific permission of the user to store their details, it is necessary to anonymise the data before use. Even if this is done with one-way transforms that allow groups of posts from a unique user to still be identified without identifying the user themselves, data anonymization often requires the removal of a lot of valuable contextual

information about the age, social grouping, and other demographic details of the user, and can significantly reduce the benefits achieved from the analysis activity.

Trust in the veracity of data derived from social sources is also a major concern, particularly if derived business intelligence is to be used in the planning of operations. The lack of proof of the identity of an individual on the web (a situation which is not helped by the need for anonymization) means that it would be comparably easy for false information to be delivered via social media channels, and comparably difficult for standard Big Data analysis tools to detect those posts. In the worst case scenario, it is possible to imagine situations where an orchestrated attack involving a botnet or similar could post large numbers of predefined, untrue messages from multiple accounts, all of which would appear to corroborate each other, and which could then be used for malicious purposes (e.g. to distract the attention of staff from another incident).

4.1.3 Industry provision of open data

As an industry, the railways have been surprisingly slow to open up their data. Despite this, good progress is now being made on the publication of key data, such as timetables and vehicle movements, that enable application developers to create tools for journey planning etc. This type of data will be vital for the successful soft-integration of modes as components of the multimodal public transport system in the next 5 years, and this will be critical to achieving the levels of modal shift from private cars envisaged in the European Commission's transport white paper.

A good example of the rail industry's engagement with open data can be found in Network Rail's public data feeds (Network Rail, 2016). The feeds are made freely available to the community via a license system, which enables management of load on the service, and the easy removal of users

who do not comply with the conditions of fair usage. Public feeds that users can then subscribe to include:

- SCHEDULE - daily extracts and updates of train schedules from the Integrated Train Planning System, in CIF and JSON formats
- MOVEMENT - train positioning and movement event data
- TD - train positioning data at signalling berth level
- TSR (Temporary Speed Restrictions) – details of temporary reductions in permissible speed across the rail network
- VSTP (Very Short Term Plan) – train schedules created via the VSTP process which are not available via the SCHEDULE feed
- RTPPM (Real-Time Public Performance Measure) - performance of trains against the timetable, measured as the percentage of trains arriving at their destination on-time
- SMART - train describer berth offset data used for train reporting
- Corpus - location reference data
- BPLAN - train planning data, including locations and sectional running times
- Train Planning Network Model - contains very detailed information on the network model used by ITPS, the Integrated Train Planning System.

4.2 Social media for improved situational awareness

In order to demonstrate the potential utility of social media data as a supplementary data source for operational staff, the SP3 team have developed a simple system for the gathering, analysis, and presentation of social media posts relevant to the railways. In section 4.2.1 we will present the data resources leveraged by the system, in sections 4.2.2 and 4.2.3 we will describe how social media data and route data were managed in the system, in section 4.2.4 and 4.2.5 we will show how messages have been associated with specific services (thus providing a filter of sorts that reduces the likelihood of non-rail messages being put forward for analysis), before finally discussing the presentation of the data in section 4.2.6. A flowchart illustrating the key steps in the process can be found in Figure 4-5.

4.2.1 Data resources leveraged

The demonstration system drew on a number of publicly-available data resources. For brevity, these can be seen summarised in Table 4-1, however, a more detailed breakdown of the resources is as follows:

- Publicly accessible, geotagged social media posts – these were harvested from a well-known social media platform and were filtered to extract only those posts that had a direct relationship to the railways or railway operations
- Map of the UK's Railway infrastructure from the Ordinance Survey (Meridian 2 dataset) (OS, 2016)
- Train positioning and movement event data (so-called “TD feed” messages) from Network Rail's Open Data Feed service. The messages are batched and updated positions are reported in near real-time (NR, 2016)
- Railway Observation Points and location codes (e.g. STANOXs, TIPLOCs, etc.), derived from Network Rail's Corpus database, and accessed via the Railway Codes website (Railway Codes, 2016)
- The National Public Transport Access Nodes (NaPTAN) database. There is a NaPTAN record for every bus stop, railway station, airport, ferry terminal etc. in England, Scotland and Wales (DfT, 2016a). We use the NaPTAN database to acquire the geographical coordinates of locations corresponding to TIPLOCs as the Railway Codes database does not include the geocodes. NaPTAN database is maintained by Landmark under contract to the Department of Transport (DfT, 2016b). TIPLOCs (Timing Point Locations) are used by the train planners to identify what time trains should arrive at, depart or pass a particular point (Railway Codes, 2016).

Table 4-1 Data resources leveraged in the situational awareness demonstration

Data resource	Examples

Social media data	Content, geolocation, time of creation, links to external content
IM public data	Live vehicle movements, train describers, notifications of TSRs etc.
Ordnance survey	Infrastructure layouts
ATOC data	Timetables, fares and supporting information
NaPTAN	Information on access points / interchanges

4.2.2 Initial route data

The Ordnance Survey offers a wide range of digital map products, including the digital map of the UK railway network as an ERSI shapefile. The geographic coordinates of the railway tracks were extracted from the shapefiles by transforming the data into the geoJSON format using the QGIS package. The geoJSON was then parsed using a custom-written application, and the coordinate pairs saved to a local database instance. Around 253 thousand coordinate pairs from the UK railway network were gathered, and these were used for the identification of social media messages posted at or near the railway tracks. Later assignment of harvested messages to specific segments of the railway was performed using the shapefiles themselves (the coordinate pairs were necessary in the initial stages of the process because of limitations to the social media platform's API).

4.2.3 Harvesting social media data

Social media messages were gathered via APIs provided by the social media site. A radius of 100m around each coordinate pair known to lie on the rail network was used for the initial filtering of content – while this meant that a large number of posts were gathered initially that were not necessarily associate with rail traffic, it did mean that no messages were missed due to the low accuracy of mobile device geotagging, a known problem in cuttings or built up areas. The API returned the messages in a

simple JSON format, which was parsed and added to the local database. Associated metadata (date and time of posting etc.) was also stored.

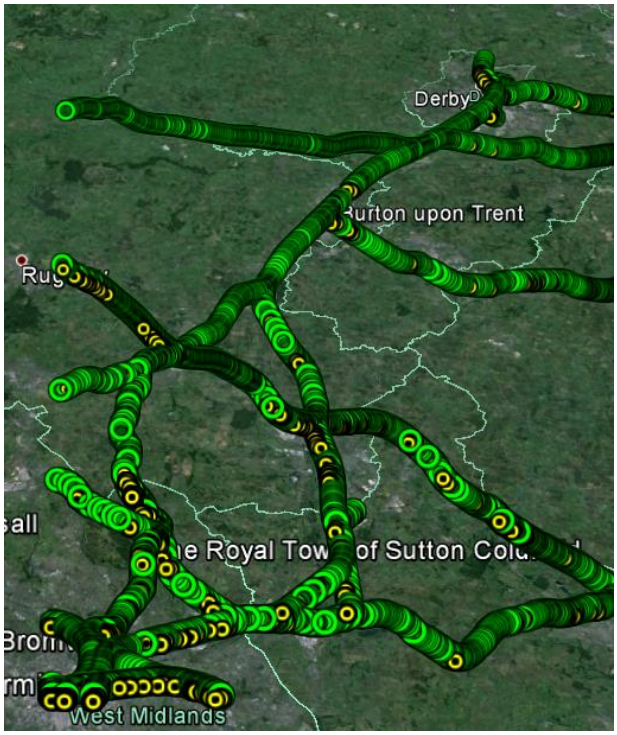


FIGURE 4-2 LOCATIONS OF SOCIAL MEDIA POSTS (YELLOW) OVERLAID ON KNOWN TRACK POSITIONS (GREEN)

Once candidate social media posts had been identified, PostGIS, a GIS data plugin for the PostgreSQL database system was used to assign messages to the closest rail track segment in the OS shapefiles. Figure 4-2 shows the points corresponding to the geographical coordinates of a small number of messages that are collected from trains in the Birmingham / Derby area. Green circles indicate the initial coordinates drawn from the shapefiles, while yellow circles represent the geographical coordinates of the messages. Although the messages shown had not yet been filtered (to remove those within the 100m radius not posted from trains) it can be seen that most are actually located on the rail tracks.

4.2.4 RELATING OPERATIONAL CONTROLS AND PHYSICAL INFRASTRUCTURE

A key element of the association of social media messages with the services running at the time revolves around the operational control points on the network. Each of these control points has a unique (with a few exceptions) STANOX code associated with it, and further location codes, for example TIPLOCs (Timing Point Locations), may also be used by various stakeholders to describe when vehicles should arrive at, depart, or pass through the same area.

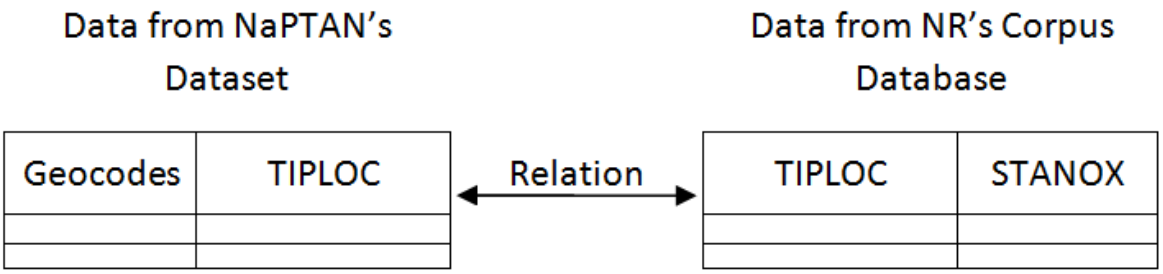


FIGURE 4-3 RELATIONSHIP BETWEEN NAPTAN AND CORPUS DATA

Network Rail's Corpus database provides a list of STANOX, TIPLOC and NLC codes. The National Public Transport Access Nodes (NaPTAN) database provides the geographical coordinates of locations corresponding to their TIPLOC codes. Since both data sources contain TIPLOC codes, the team established a relationship between the two tables using TIPLOCs as the primary and foreign keys, resulting in a table where the geographical coordinates of locations could be identified from their corresponding to their STANOX codes. Locations with STANOX codes were then assigned to segments in the OS maps based on the geographical coordinates corresponding to the STANOX codes, providing a relationship between the operational locations used by the railways, and the track sections that the social media posts had been assigned to.

4.2.5 CANDIDATE VEHICLE IDENTIFICATION

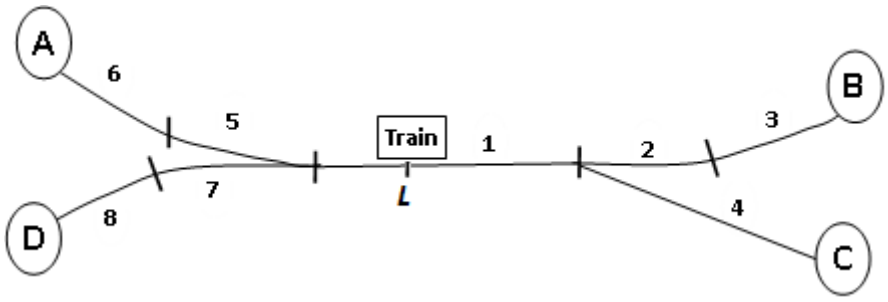


FIGURE 4-4 IDENTIFIYING CANDIDATE PATHS FOR A VEHICLE

Once the control points have been associated with tracks (and hence coordinates where messages could be posted), the vehicles passing through the area at the time of posting must be identified. This requires two stages: firstly, the identification of all the control points surrounding a given location, and secondly, the identification of all the vehicles within that envelope at the time the message was posted.

Identification of associated control points

Figure 4-4 shows a small portion of the railway network, where A, B, C, and D represent four observation points, each of which has a STANOX code. These tracks consist of eight track segments (in the mapping data) and each segment is denoted with a number (from 1 to 8). Now assume that there

is a social media post that has appeared at segment 1; we assume this message has been posted from a train, but at present we do not know the identity of the vehicle, or even the route from which it has arrived in the segment. The train in the figure could come to its current position from any of the four surrounding directions, and may have been observed at any of the four observation points, namely A, B, C, and D.

The following steps are taken to find out all candidate paths and the surrounding observation points (with STANOX codes) for a train at a specific location

1. First we identify the line segment on which the train was situated, by measuring the straight line distance between the message location and the railway line segments. The closest line segment from the posting location is where the train was assumed to be situated.
2. Next we identify all line segments on the map that are connected to the line-segment on which the target train (the train from which the tweet was made) was situated at the time of posting. We call each of these connected line segments a neighbouring segment.
3. We then recursively find the neighbouring segments for each of those segments connected to the segment from which the tweet was made, this process continues until a railway observation point is reached on every branch of the connected line segments.

It is noteworthy that due to the construction procedure, each candidate path must have at least one and maximum two (one in each direction) observation points on it. Dead-end candidate paths will feature only a single observation point, and thus may be easily identified if needed.

Identification of candidate vehicles

In order to illustrate the process by which messages are assigned to vehicles, suppose a post was made at time ***t*** from a train at location ***L***, in Figure 4-4. The train from which the message originated must have passed through at least one of the four surrounding observation points (A, B, C, and D) on the candidate paths. In the public train movement feeds a train is logged each time it moves through one of these points; thus, those trains that have entered but not yet left that region at time ***t*** are valid candidates. It is usual in quiet or rural areas of the network, for only one vehicle to be in the area of interest at any one time. However in busy regions of the network, such as in urban areas, more than one candidate train may exist. In these cases the candidate vehicles can be further narrowed down based on entry/exit point combinations, and the use of assumed line speeds to place vehicles in different segments within the area of interest. The completed process for locating social media posts on the network, and then associating them with particular services is show in Figure 4-5.

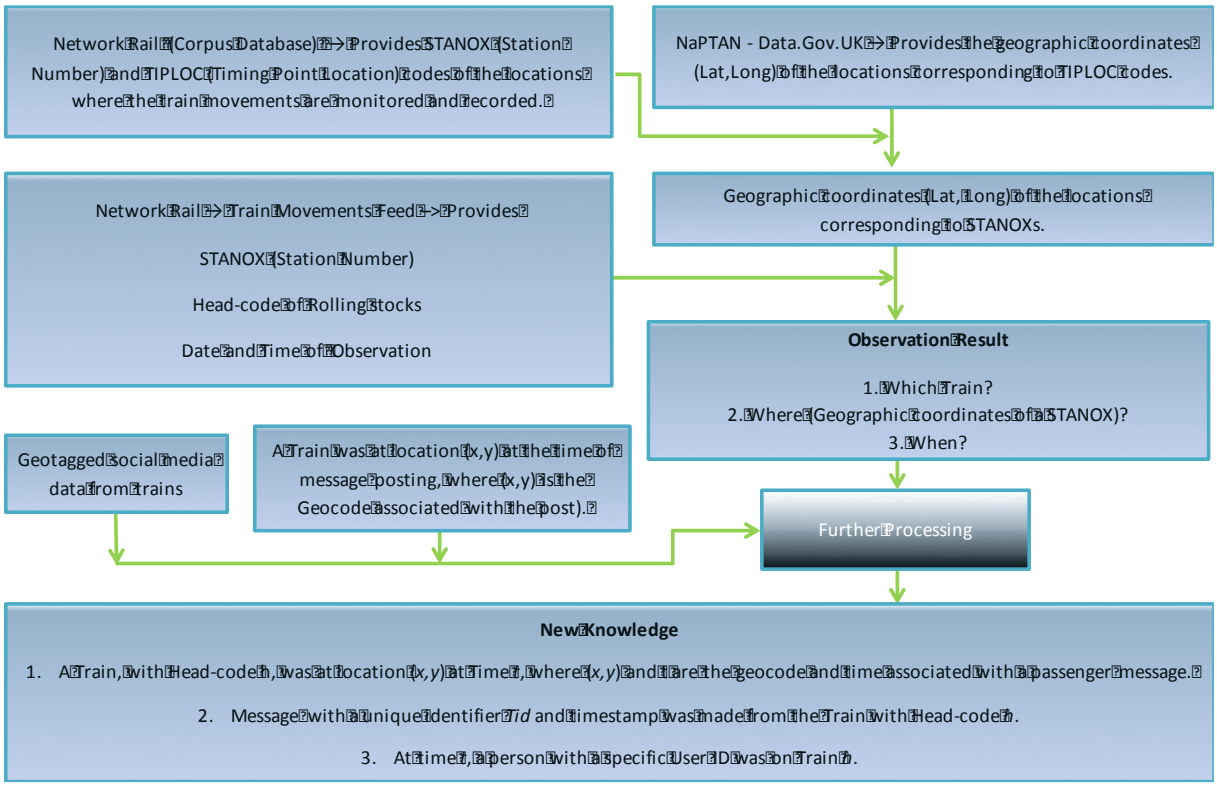


FIGURE 4-5 PROCESS FOLLOWED TO ASSOCIATE GEOTAGGED SOCIAL MEDIA POSTS WITH SERVICE HEADCODES IN THE SITUATIONAL AWARENESS DEMONSTRATION

4.2.6 PRESENTATION AND EXPLOITATION OF POSTS

In order to provide improved situational awareness to operational staff, it is necessary to not only identify posts originating from the network, but also to visualise the most important posts in such a way that they can be reviewed and assigned to locations quickly and easily. Figure 4-6 shows how messages were presented in the context of the demonstration application. Here, a combination of tag detection and a simple sentiment analysis was used to filter the social media posts, and select only those that either were marked as relating to service status (normally originating from "official" sources, such as train operators), or where the sentiment analysis suggested that the poster was angry or unhappy.

Posts in a geographic area were then grouped together, and a multi-segment marker was assigned; coloured segments were added, given the operator a quick visual prompt as to the mixture of message categories contained within the marker. Clicking on a marker caused the messages it represented to popup in the surrounding space. In order to ensure content remained relevant, messages were assumed to have "expired" after a short timeframe, and when no active messages remained in an area, the marker was removed.

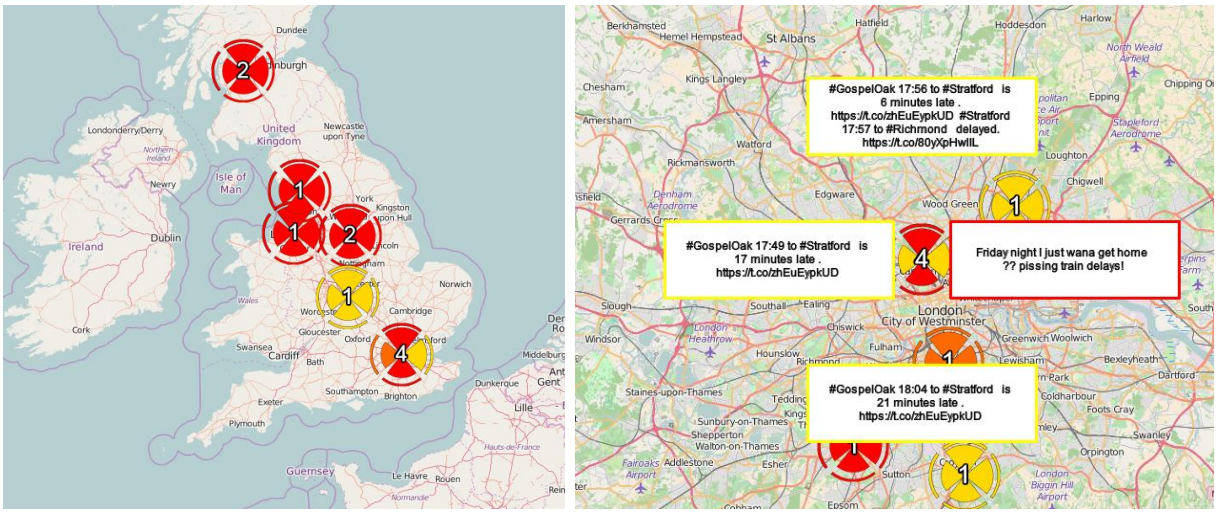


FIGURE 4-6 SCREEN CAPTURES FROM THE SITUATIONAL AWARENESS DEMONSTRATION SHOWING HIGH-LEVEL EVENT MARKERS (LEFT) AND THE EXPANSION OF A SPECIFIC MARKER (RIGHT).

5. CONCLUSIONS

Deliverable D3.4.1 of the C4R project attempted to catalogue the data models in use in the rail industry, and identify gaps that would need to be filled in order to support 3 “visions” of the railway of 2050. This document, Deliverable D3.4.2, has complemented that work by looking at the ICT architectures and data resources the industry may choose to draw on in the short to medium term, taking a particular interest in modular, scalable approaches to architecture, and diversity of data resources.

In section 2 we looked at a candidate software architecture for future TMS platforms, the Enterprise Service Bus. The ESB is becoming the architecture of choice within the software engineering community, and has already been proposed for use in the rail sector by previous projects including InteGRail and ON-TIME, both of which members of SP3 have been heavily involved with. It looked at how the industry was already moving towards the ESB model, and discussed activities in Shift2Rail and the UK’s Digital Railway programme, where architectures of this type are being planned.

Section 3 of the document focussed on semantic data models and supporting architectures. It reintroduced the concept of ontologies (first touched on in D3.4.1), data models that allow the explicit meaning of an item of data to be maintained outside the scope of the originating system in a machine-interpretable form. It gave a detailed explanation of the Railway Core Ontology (RaCoOn) model, and showed through a case study how it could add value to railway operations by decoupling software applications from the data they operate on. It also described how diverse of data resources could be leveraged by the industry, and discussed ontology models that could be used in conjunction with RaCoOn to describe data from sensors on the network. It presented a candidate architecture for handling this type of high-velocity data within a reasoning framework; the architecture includes a key-value store alongside the standard triplestore system, thus preventing the need for computationally expensive updates to inferred values unless derived information is requested by an application. It looked at the potential costs and benefits to the industry of the semantic approach to data integration, and discussed some of the limitations including the scalability of the models, the challenge of distributing the reasoning architecture, and version management – all of which are outstanding issues well known in the community.

Section 4 wrapped up by showing how data external to the railways could be used to support railway operations, particularly through the provision of enhanced situational awareness during periods of disrupted operations. It demonstrated the harvesting of geotagged data from a well-known social media platform, the assignment of that data to a particular service known to be running in the area at the time, and the presentation of the information to control room staff based on an analysis of the sentiment contained in the message. It also discussed some of the issues of the use of social data in a railway context, including the need to ensure user privacy, and issues of the trustworthiness of the data, particularly if that data is being used as the basis for making operational decisions.

In closing, the SP3 team would like to make the following recommendations:

- The Enterprise Service Bus architecture is an appropriate foundation for the development of Traffic Management Systems, and that the outcomes from existing research activities using this technology, including the FP7 ON-TIME project, should be adopted by the industry;
- That continuing with efforts to improve the management and integration of data resources within the industry must be treated as a priority over the next 5 years. The appropriate usage of new technologies, including ontology, has a key role to play in these improvements;
- The harvesting and utilisation of open data from public sources has great potential to provide the industry with enhanced situational awareness during disruptions, but care is needed to ensure user privacy is respected, and that (as far as is practical) the authenticity and provenance of material used for decision making.

References

Allen, J. F. (1984). "Towards a general theory of action and time." Artif. Intell., 23(2), 123–154.

AWT Consortium. (2013-2015). All Ways Travelling. Retrieved May 19, 2015 from http://www.allwaystravelling.eu/

AWT Consortium. (2014). All Ways Travelling - To develop and validate a European passenger transport information and booking system across transport modes (phase 1). Final Report, Contract MOVE/C2/SER/2012 489/SI2.646722.

Azzoug, A., and Kaewunruen, S. (2017). RideComfort: A Development of Crowdsourcing Smartphones in Measuring Train Ride Quality. Frontiers in Built Environment, volume 3 p. 3. Available online via http://journal.frontiersin.org/article/10.3389/fbuil.2017.00003, last accessed 28th February 2017.

Bodnar, T., Tucker, C., Hopkinson, K., and Bilen, S. (2014) Increasing the veracity of event detection on social media networks through user trust modelling. In proceedings of Big Data (Big Data 2014) IEEE International Conference on, Oct 2014, pp. 636–643.

Boren, Z.D. (2014). There are officially more mobile devices than people in the world. Available online via http://www.independent.co.uk/life-style/gadgets-and-tech/news/there-are-officially-more-mobile-devices-than-people-in-the-world-9780518.html, last accessed on 13/04/2015.

Botts, M., Percivall, G., Reed, C., & Davidson, J. (October 2006). OGC® sensor web enablement: Overview and high level architecture. In International conference on GeoSensor Networks (pp. 175-190). Springer Berlin Heidelberg.

Brickley, D. (2003). "W3C basic geo vocabulary." http://www.w3.org/2003/01/geo/

Compton, M., Barnaghi, P., Bermudez, L., GarcíA-Castro, R., Corcho, O., Cox, S., ... & Huang, V. (2012). The SSN ontology of the W3C semantic sensor network incubator group. Web Semantics: Science, Services and Agents on the World Wide Web, 17, 25-32.

daCosta, F. (2013). Rethinking the Internet of Things, A Scalable Approach to Connecting Everything. 1st ed. Apress, Berkely, CA, USA.

Department for Transport (2016a). National Public Transport Access Nodes (NaPTAN). Available online via https://data.gov.uk/dataset/naptan, last accessed on 03/02/2016.

Department for Transport (2016b). National Public Transport Access Nodes. Available online via https://www.gov.uk/government/publications/national-public-transport-access-node-schema, last accessed on 03/02/2016.

ERIM Workgroup. (2014). UIC RailTopoModel Railway Network Description - A conceptual model to describe a railway network (Version RC2). Paris.

European Commission. (2011). WHITE PAPER - Roadmap to a Single European Transport Area - Towards a competitive and resource efficient transport system Com(2011) 144 final. Available online from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0144:FIN:en:PDF, last accessed 26th February 2017.

Gallaher, M., O'Connor, A., Dettbarn, J., and Gilday, L. (2004). "Cost analysis of inadequate interoperability in the U.S. capital facilities industry." U.S. Dept. of Commerce, Technology Administration, National Institute of Standards and Technology, http://iringtoday.com/wordpress/wp-content/uploads/2012/02/NIST-Interoperability-Report.pdf

GSMA (2015). GSMA Intelligence. Available online via https://gsmaintelligence.com/, last accessed on 16/04/2015.

He, J., Shen, W., Divakaruni, P., Wynter, L., and Lawrence, R. "Improving Traffic Prediction with Tweet Semantics," in Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, ser. IJCAI '13. AAAI Press, 2013, pp. 1387–1393. Available online via http://dl.acm.org/citation.cfm?id=2540128.2540328

Hodgson, R., and Keller, P. J. (2011). "QUDT: Quantities, units, dimensions and types ontologies." http://www.qudt.org/

Hurwitz, J., Bloor, R., Kaufman, M., & Halper, F. (2009). Service oriented architecture for dummies. Wiley Publishing, Inc.

InteGRail Consortium. (2009). InteGRail Vision Paper. Available online via http://www.integrail.info/documenti/InteGRail_Vision_Paper.pdf, last accessed 26th February 2017.

InteGRail Consortium (2011) InteGRail - Intelligent Integration of Railway Systems, Final Report. Available from http://www.integrail.info/documenti/InteGRail_Final_Project_Report.pdf, last accessed September 5th, 2016.

IS0 (International Standards Organisation). (2003a). ISO 13374: Condition monitoring and diagnostics of machines — Data processing, communication and presentation. Available online from https://www.iso.org/obp/ui/#iso:std:iso:13374:-1:ed-1:v1:en, last accessed September 5th, 2016.

ISO (International Standards Organisation). (2003b). "Integration of life-cycle data for process plants including oil and gas production facilities." ISO 15926-2:2003.

Li, B., Si, X., Lyu, M.R., King, I., and Chang, E.Y. (2011) Question Identification on Twitter. In proceedings of the 20th ACM International Conference on Information and Knowledge Management, ser. CIKM'11.

New York, NY, USA: ACM, 2011, pp. 2477–2480. Available online via http://doi.acm.org/10.1145/2063576.2063996

Morris, M.R., Teevan, J., and Panovich, K. (2010) What Do People Ask Their Social Networks, and Why?: A Survey Study of Status Message Behavior. In proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 1739–1748. Available online via http://doi.acm.org/10.1145/1753326.1753587

Network Rail Limited. (2013). "Technical strategy: A future driven by innovation." http://www.networkrail.co.uk/publications/technical-strategy.pdf

Network Rail (2016). Open Data Feeds. Available online via http://nrodwiki.rockshore.net/index.php/Open_Data_Releases, last accessed on 16/04/2015.

Office of Rail Regulation. (2015). "GB rail industry financial information 2013–14." http://orr.gov.uk/__data/assets/pdf_file/0005/16997/gb-rail-industry-financials-2013-14.pdf

ON-TIME Consortium (2013a). Library of Data and Communication Models. Available online via http://www.ontime-project.eu/deliverables.aspx, last accessed 26th February 2017.

ON-TIME Consortium (2013b). Architecture specification and integration requirements. Available online via http://www.ontime-project.eu/deliverables.aspx, last accessed 26th February 2017.

ON-TIME Consortium (2014). ON-TIME Project – NTT Data Final Presentation. Rome, Italy. 29th October 2014.

Oracle (2013). Harnessing the value of social media. Available online via http://www.oracle.com/us/industries/utilities/harnessing-social-media-wp-1959234.pdf, last accessed 28th February 2017.

Ordnance Survey. (2014). "Spatial relations ontology." http://data.ordnancesurvey.co.uk/ontology/spatialrelations

Ordnance Survey (2016). Meridian 2 Dataset. Available online via https://www.ordnancesurvey.co.uk/business-and-government/products/meridian2.html, last accessed on 03/02/2016.

Rahman, S.S., Creese, S., and Goldsmith, M. (2012) Accepting Information with a Pinch of Salt: Handling Untrusted Information Sources. In Security and Trust Management, ser. Lecture Notes in Computer Science, C. Meadows and C. Fernandez-Gago, Eds. Springer Berlin Heidelberg, 2012, vol. 7170, pp. 223–238. Available online via http://dx.doi.org/10.1007/978-3-642-29963-6 16

Rail Safety and Standards Board. (2011). National Information Systems catalogue for non-Network Rail systems (project T962). Original catalogue available via www.sparkrail.org, last accessed 26th February

2017. Updated version from 2015 used as basis for material reference in this report – not currently available online.

Railway Codes (2016). Railway Codes Website - CRS, NLC, TIPLOC and STANOX Codes. Available online via http://www.railwaycodes.org.uk/CRS/CRS0.shtm, last accessed on 03/02/2016.

Raskin, R., & Pan, M. (2003). Semantic web for earth and environmental terminology (sweet). In Proc. of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data.

Roberts, C., Easton, J., Davies, R., Sharples, S., and Golightly, D. (2011). "The specification of a system-wide data framework for the railway industry final report." http://p.sparkrail.org/record.asp?q=PB022964

S2R (2015). The Shift2Rail Multi-Annual Action Plan. Available online via http://ec.europa.eu/research/participants/data/ref/h2020/other/wp/jtis/h2020-maap-shift2rail_en.pdf, last accessed 28th February 2017.

Suárez-Figueroa, M. C., Go´mez-Pérez, A., and Fernández-Lo´pez, M. (2012). "The NeON methodology for ontology engineering." Chapter 1, Ontology engineering in a networked world, M. C. Suárez-Figueroa, A. Go´mez-Pérez, E. Motta, and A. Gangemi, eds., Springer, Berlin, 9–34.

World Wide Web Consortium. (2006). "Time ontology in OWL." http://www.w3.org/TR/owl-time/