





# Leveraging semantic technologies for digital interoperability in the European Railway domain

Julián Andrés Rojas<sup>(r)</sup> , Marina Aguado, Polymnia Vasilopoulou, Ivo Velitchkov, Dylan Van Assche<sup>(r)</sup> , Pieter Colpaert<sup>(r)</sup> , and Ruben Verborgh<sup>(r)</sup> 

IDLab, Department of Electronics and Information Systems, Ghent University – imec  
Technologiepark-Zwijnaarde 122, 9052 Ghent, Belgium  
`{firstname.lastname}@ugent.be`

**Abstract.** The European Union Agency for Railways is an European authority, tasked with the provision of a legal and technical framework to support harmonized and safe cross-border railway operations throughout the EU. So far, the agency relied on traditional application-centric approaches to support the data exchange among multiple actors interacting within the railway domain. This lead however, to isolated digital environments that consequently added barriers to digital interoperability while increasing the cost of maintenance and innovation. In this work, we show how Semantic Web technologies are leveraged to create a semantic layer for data integration across the base registries maintained by the agency. We validate the usefulness of this approach by supporting route compatibility checks, a highly demanded use case in this domain, which was not available over the agency’s registries before. Our contributions include (i) an official ontology for the railway infrastructure and authorized vehicle types, including 28 reference datasets; (ii) a reusable Knowledge Graph describing the European railway infrastructure; (iii) a cost-efficient system architecture that enables high-flexibility for use case development; and (iv) an open source and RDF native Web application to support route compatibility checks. This work demonstrates how data-centric system design, powered by Semantic Web technologies and Linked Data principles, provides a framework to achieve data interoperability and unlock new and innovative use cases and applications. Based on the results obtained during this work, ERA officially decided to make Semantic Web and Linked Data-based approaches, the default setting for any future development of the data, registers and specifications under the agency’s remit for data exchange mandated by the EU legal framework. The next steps, which are already underway, include further developing and bringing these solutions to a production-ready state.

## 1 Introduction

The establishment of an interoperable European railway area without frontiers, while guaranteeing railway operation safety, is the prime objective of the European Union Agency for Railways (ERA) [7]. Since 2019 ERA became the

European authority<sup>1</sup> for cross-border rail traffic in Europe, mandated under the European Union (EU) law, to devise the technical and legal framework for supporting harmonised and safe cross-border railway operations.

The European railway ecosystem presents a particularly challenging scenario for interoperability, not only regarding physical aspects (e.g., infrastructure, energy systems, etc.) but also digital ones (e.g., information). Multiple organisations, such as Infrastructure Managers (IMs)<sup>2</sup> and Railway Undertakings (RUs)<sup>3</sup> [6], need to interact and exchange information to ensure safe cross-border railway operations. These organisations rely on different information management systems from multiple vendors, that are often incompatible with each other. To increase digital interoperability among heterogeneous data and information systems, ERA supports and maintains a set of base registries<sup>4</sup>, in the form of relational databases, where organisations input and access the different aspects of the information they manage and require.

However, following such traditional approach lead to isolated digital environments that consequently added barriers to digital interoperability. Tightly coupling base registries to the applications that operate over them, triggered the proliferation of overlapping and difficult to manage data models hidden inside application code, which also increased maintenance and innovation costs. Moreover, stakeholder organisations such as IMs, have to report the same information multiple times for different registries, increasing the probability of data inconsistency issues, while adding more costs to IMs due to duplicated efforts.

To address these issues, we propose a digital interoperability strategy for ERA, that adheres to the Linked Data principles<sup>5</sup> [19] and relies on standard Semantic Web [1] technologies. We built the foundations to establish a *semantic layer* for data integration within the agency, initially spanning three different base registries<sup>6</sup>: Register of Infrastructure (RINF), Register of Authorized Types of Vehicles (ERATV) and the Centralized Virtual Vehicle Register (ECVVR). We validate the usefulness of the approach by reusing the produced semantic data to support route compatibility checks (RCC), a highly-demanded use case in the railway domain. The RCC use case is stipulated and specified in EU regulations 2016/797 and 2019/773 [8, 10] and was so far, unsupported by ERA

<sup>1</sup> ERA is the European authority for cross-border rail traffic in Europe: [https://www.era.europa.eu/content/era-becomes-european-authority-cross-border-rail-traffic-europe\\_en](https://www.era.europa.eu/content/era-becomes-european-authority-cross-border-rail-traffic-europe_en)

<sup>2</sup> An Infrastructure Manager is defined as any body or firm responsible in particular for establishing, managing and maintaining railway infrastructure, including traffic management, control-command and signalling.

<sup>3</sup> A Railway Undertaking is defined as any public or private licensed undertaking, the principal business of which is to provide services for the transport of goods and/or passengers by rail with a requirement that the undertaking ensure traction.

<sup>4</sup> “A base registry is a trusted and authoritative source of information which can and should be digitally reused by others, where one organisation is responsible and accountable for the collection, use, updating and preservation of information.” [24]

<sup>5</sup> Principles of Linked Data: <https://www.w3.org/DesignIssues/LinkedData.html>

<sup>6</sup> Base registries of ERA: [https://www.era.europa.eu/registers\\_en](https://www.era.europa.eu/registers_en)

due to interoperability issues among base registries. Additionally, we show the flexibility of graph-based data models, by integrating an additional external data source that complements the resulting Knowledge Graph.

The contributions of this paper include (i) an ontology<sup>7</sup>, modelling railway infrastructure aspects, rolling stock and authorized vehicle types, and 28 independently managed reference datasets; (ii) a public and reusable RDF Knowledge Graph<sup>8</sup> with 13.8 million triples about the European railway infrastructure and more than 800 thousand rolling stocks; (iii) a cost-efficient system architecture that enables high-flexibility for use case support; and (iv) an open source and RDF native Web application<sup>9</sup> to support and process RCC queries.

This work demonstrates how data-centric system design, powered by Semantic Web technologies, provides a framework to achieve data interoperability and unlock innovative use cases and applications. The results of the work presented in this paper had a strong impact on ERA<sup>10</sup>, which decided on making Semantic Web technologies the default setting for any future development of data, registers and specifications, under the agency’s remit, for data exchange mandated by the EU legal framework. The next steps, which are already underway, include further extending the ontology with additional aspects, aligned with the requirements of the railway domain and evolving the system architecture towards a production-ready solution, fully integrated with the data management workflows of ERA.

The remainder of this paper is organized as follows: Section 2 presents an overview of related work in the context of modelling approaches and interoperability for the railway domain. Section 3 describes the data sources and the RCC use case requirements. Section 4 gives an overview and description of our proposed solution architecture. Section 5 discusses advantages and limitations of the approach and Section 6 presents our conclusions and perspectives for future work.

## 2 Related Work

In this section, we present different (semantic) data models to describe the railway domain focusing on different aspects of the domain and motivated by different use cases. Also existing related work applying semantic technologies in the railway domain. We studied these models, aiming on reusing as much as possible their embedded domain-specific knowledge (e.g. definitions, categorizations, naming conventions, etc.), during the creation of ERA’s ontology.

Multiple domain data models were proposed (some still under active development), to bridge the interoperability challenges by uniformly describing the

<sup>7</sup> <http://era.ilabt.imec.be/era-vocabulary/index-en.html>

<sup>8</sup> <http://era.ilabt.imec.be/>

<sup>9</sup> <http://era.ilabt.imec.be/test/compatibility-check-demo/>

<sup>10</sup> ERA’s roadmap for Linked Data mainstreaming: [https://www.era.europa.eu/sites/default/files/agency/docs/decision/decision\\_n250\\_annex1\\_linked\\_data\\_en.pdf](https://www.era.europa.eu/sites/default/files/agency/docs/decision/decision_n250_annex1_linked_data_en.pdf)

different technical aspects related to the railway domain. However, most lack semantic definitions that promote/guarantee the use of persistent identifiers across data sources, hindering interoperability when exchanging data across organizations. Available models range from company-specific to industrial consortium-driven standardization efforts. For example, the *Informatie Model Spoor*<sup>11</sup>, developed by the Dutch IM ProRail, provides an XML Schema-based model with integrated functional and geographic information about the railway infrastructure. IMSpoorXML is currently used within ProRail and is under active maintenance and development. The International Railway Standard IRS:30100 Rail-TopoModel<sup>12</sup> [22] was developed under patronage of the International Union of Railways (UIC) and provides a systemic UML-based model for describing the topological aspects of railway infrastructure. It relies on the *connectivity graph* mathematical concept [16] to describe the interconnection of the different railway network elements. Implementations of RailTopoModel include RailML<sup>13</sup> [21] and the EULYNX<sup>14</sup> initiative, both currently developed by industrial consortiums.

The use of semantic technologies, for modelling the railway domain is not new. In 2011, the EU project InteGRail created an ontology integrating the major railway sub-systems, to achieve higher levels of performance in terms of capacity, average speed and punctuality, safety and the optimised usage of resources in railway systems [28]. Smart Rail is another EU project that applied semantic technologies for modelling organizational aspects of the railway domain. It produced an ontology<sup>15</sup>, focused on modelling stakeholders and physical resources of the railway infrastructure. RaCoOn (Rail Core Ontologies) is a set of domain ontologies that model areas of the rail domain commonly used in railway data exchange [26]. A study of how Linked Data was applied in the British railway domain, highlights the reduction of costs as a consequence of more efficient data flows, and hints towards the need of increasing adoption from industry [23]. More recently, Bischof et al. [2] outlined the requirements and challenges to define an open standard ontology for railway topologies based on existing standards. None of these approaches evolved beyond academic exercises and the produced ontologies are currently unmaintained or no longer available. In contrast, one of the main goals of ERA, as an European authority for the domain, is to provide a fully supported and open reference ontology not only for internal data management operations but targeting also its adoption and extension by the stakeholders of the railway domain, as an asset that supports their own use cases.

<sup>11</sup> <https://confluence.rigd-loxia.nl/display/IMSP/IMSpoor+Publicatie+Home>

<sup>12</sup> [http://www.railtopomodel.org/en/download/irs30100-apr16-7594BCA1524E14224D0.html?file=files/download/RailTopoModel/180416\\_uic.irs30100.pdf](http://www.railtopomodel.org/en/download/irs30100-apr16-7594BCA1524E14224D0.html?file=files/download/RailTopoModel/180416_uic.irs30100.pdf)

<sup>13</sup> [https://wiki3.railml.org/wiki/Main\\_Page](https://wiki3.railml.org/wiki/Main_Page)

<sup>14</sup> <https://dataprep.eulynx.eu/2020-10/index.htm>

<sup>15</sup> <https://ontology.tno.nl/smart-rail/>

### 3 Data Sources and Use Case

In this section, we outline the different data sources reused by our proposed solution and describe the RCC use case as main motivator for this work.

#### 3.1 ERA's base registries

Our approach considers, so far, 3 of the base registries<sup>16</sup> maintained by ERA, namely the Register of Infrastructure (RINF), the Register of Authorized Types of Vehicles (ERATV) and the Centralized Virtual Vehicle Register (ECVVR). These registries contain overlapping conceptual definitions, represented as properties of different type of entities, which are locked within their respective data silos. Next we give a brief description for each of these registries.

**Register of Infrastructure** The European Register of Infrastructure (RINF) was introduced following Article 35 of the EU regulation 2008/57/EC [4]. RINF contains the main features of fixed installations related to subsystems of infrastructure, energy and parts of control-command and signaling. It publishes performance and technical characteristics mainly related to interfaces with rolling stock and operation. It is maintained as a relational database and its content is provided by different European IMs, by means of a predefined XML Schema<sup>17</sup>.

**Register of Authorized Types of Vehicles** The European Register of Authorized Types of Vehicles (ERATV) is introduced by Article 5 of the EU regulation 2011/665/EU [5]. It aims to publish and keep an up-to-date set of authorized types of vehicles including information that references the technical specifications for each parameter. ERATV is maintained as a relational database populated through a Web application by multiple authorizing entities. It provides also additional information for a certain vehicle type, such as manufacturing country, manufacturer, category and different physical and operational parameters.

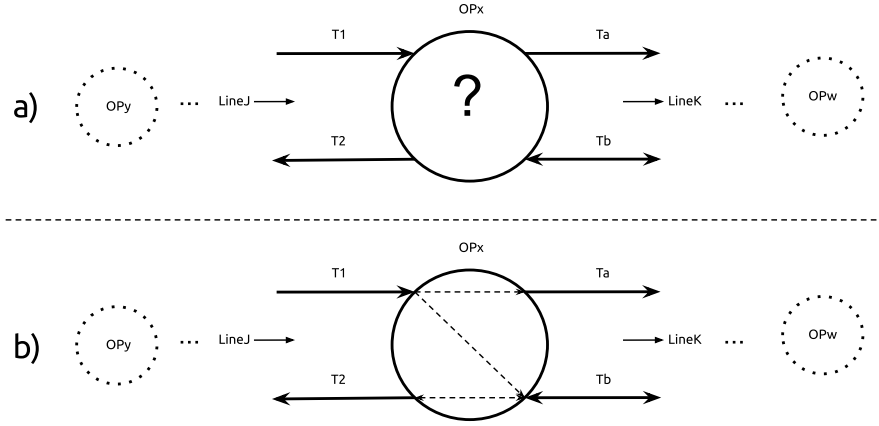
**Centralized Virtual Vehicle Register** The European Centralised Virtual Vehicle Register (ECVVR) is a base registry maintained by ERA, in accordance with the EU regulation 2018/1614 [9]. ECVVR defines a decentralized architecture for information search and retrieval of rolling stock data, where each Member State hosts and publishes their own national vehicle registry(ies), accessible through Web-based interfaces.

#### 3.2 External data source

There are known limitations for ERA's base registries, as is the case of RINF and the limited granularity it gives over the railway topology. RINF provides a

<sup>16</sup> <https://www.era.europa.eu/registers.en>

<sup>17</sup> [https://www.era.europa.eu/sites/default/files/registers/docs/rinf\\_schema\\_en.xsd](https://www.era.europa.eu/sites/default/files/registers/docs/rinf_schema_en.xsd)



**Fig. 1.** a) a schematic diagram of an operational point where its internal connections are unknown; b) how this information can be completed from data provided in Table 1.

view over the railway infrastructure, commonly referred as a meso-level view<sup>18</sup>, where complex topological structures inside stations, junctions, switches, etc., are abstracted into single nodes in the network graph. Route calculations over this limited view, may wrongfully assume certain direction changes, not possible in the real world. Calculating end-to-end routes with high accuracy, requires further data about the connectivity within each network node. This connectivity issue currently stands as one of the main challenges, for an accurate and reliable data source description of the European railway infrastructure topology. For this reason we also consider an external data source, provided by the Dutch IM ProRail, which provides an additional topological description for addressing this issue limited to the region of Utrecht in The Netherlands.

**Connectivity data in the Utrecht area** The Dutch IM ProRail, provided us with an additional data source for exploring an alternative solution for the lack of real information about the internal connectivity inside network nodes (also called operational points). It consists of a table that groups all the different permutations of incoming and outgoing tracks for a set of operational points, and states if they are connected or not.

The operational point  $OPx$  (Figure 1) has two incoming tracks ( $T1$  and  $T2$ ) from  $OPy$  and belonging to the national line  $LineJ$ . We know these are incoming tracks thanks to the logical direction defined for  $LineJ$ , despite  $T1$  being a bidirectional track.  $OPx$  also has two outgoing tracks ( $Ta$ ,  $Tb$ ), going towards  $OPw$  and belonging to another national line  $LineK$ . Based on this information we establish the correct connectivity that reflects real-world behavior.

<sup>18</sup> See section 1.6 of [22] for a description of railway vie levels.

IN_Line	IN_OP	IN_Track	OP	OUT_Track	OUT_OP	OUT_Line	Connected
LineJ	OPy	T1	OPx	Ta	OPw	LineK	<b>true</b>
LineJ	OPy	T1	OPx	Tb	OPw	LineK	<b>true</b>
LineJ	OPy	T2	OPx	Ta	OPw	LineK	<b>false</b>
LineJ	OPy	T2	OPx	Tb	OPw	LineK	<b>true</b>

**Table 1.** All the possible permutations between incoming and outgoing tracks of OPx, plus a column that states if there is a possible connection between two pairs of tracks.

### 3.3 Use Case: Route Compatibility Check

Article 23 (point b) of the European regulation 2016/797 stipulates [8] that: *“Before a railway undertaking uses a vehicle in the area of use specified in its authorisation for placing on the market, it shall check: ... (b) that the vehicle is compatible with the route on the basis of the infrastructure register, the relevant TSIs or any relevant information to be provided by the infrastructure manager free of charge and within a reasonable period of time, where such a register does not exist or is incomplete”*.

The specific procedures for assessing if a certain vehicle is compatible with a certain route, are further specified by the Annex D1 of the EU regulation 2019/773 [10]. These specifications directly refer to specific data properties within RINF and ERATV, of 22 different technical aspects that need to be compared to determine if there is technical compatibility. This specification already highlights a clear need for interoperability at least between RINF and ERATV, which we address with the proposed ontology and derived Knowledge Graph.

To determine if a certain vehicle type is compatible with a certain route, is necessary to first find possible routes through the railway infrastructure, which involves a very particular type of queries, namely graph path finding queries. The standard query language for RDF graphs (SPARQL) does not support finding complex relation paths between RDF entities [17]. The Property Paths querying syntax, introduced in SPARQL 1.1, only allows for testing path existence but falls short on counting and retrieving the actual paths between two nodes [25], which is crucial for the RCC use case. Currently there exist non-standard extensions to SPARQL (e.g. Stardog path queries<sup>19</sup>) that address this limitation they are not widely supported across RDF graph databases. We consider this limitation in our proposed architecture and propose an alternative solution (see Section 4.2) to non-standard SPARQL extensions and according to the current Web standards to prevent vendor lock in issues.

## 4 Proposed Solution

Considering the interoperability obstacles that exist among the base registries maintained by ERA, we propose and design a solution architecture, capable of

<sup>19</sup> <https://docs.stardog.com/archive/7.5.0/query-stardog/path-queries>

creating a semantic interoperability layer for data integration over them. Moreover, we exploit the inherent flexibility of graph-based data models to also include an external data source, that enriches the resulting Knowledge Graph (KG) and addresses intrinsic limitations of the original base registries. The proposed architecture relies on an ontology, defined to cover, but not limited to, the explicit interoperability requirements brought forth by the RCC use case. The architecture implements an ETL (Extract Transform Load)-based pipeline that relies on a fully declarative approach for the KG generation process, and leverages fundamental Web principles such as caching, to reduce computational infrastructure costs while maintaining a high querying flexibility.

In this section we present a description of the main architectural components of our proposed solution. We describe the proposed ontology and give an full overview of the solution architecture, which includes a fully functional application to support route compatibility (checks available online<sup>20</sup>).

#### 4.1 The ERA Vocabulary

Our proposed ontology, the ERA Vocabulary<sup>21</sup>, was created in a collaborative effort with domain experts from ERA, ProRail, SNCF and Semantic Web experts from DG DIGIT and IDLab-imec. The ERA Vocabulary provides unique identifiers and semantic definitions for concepts and properties, common to the railway domain. We make available online its documentation, using Widoco [15] as a template generator, and the source files in a public GitHub repository<sup>22</sup>.

Following Semantic Web best practices, the ontology reuses external ontologies such as OGC GeoSPARQL, Schema.org and the EU publications office authority table<sup>23</sup> for country definitions. It defines a layered model (see Figure 2), inspired from RINF’s relational model, where the topological and functional aspects of the railway infrastructure are defined by independent entity types. The *abstraction* layer defines logical entities form the network topology graph, with *era:NodePorts* acting as nodes and both *era:MicroLinks* and *era:InternalNodeLinks* acting as edges. The *implementation* layer, represents concrete and functional objects in the real world, such as tracks, operational points (stations, switches, etc.) and vehicles (types). The link between these two layers is given by the *era:MicroNode* - *era:OperationalPoint* and *era:MicroLink* - *era:Track* relationships. Additionally, 28 reference datasets<sup>24</sup> were extracted from the base registries and defined as SKOS controlled vocabularies. They contain definitions for different domain-related technical aspects, which are envisioned to be independently managed by relevant authorities.

<sup>20</sup> <http://era.ilabt.imec.be/test/compatibility-check-demo/>

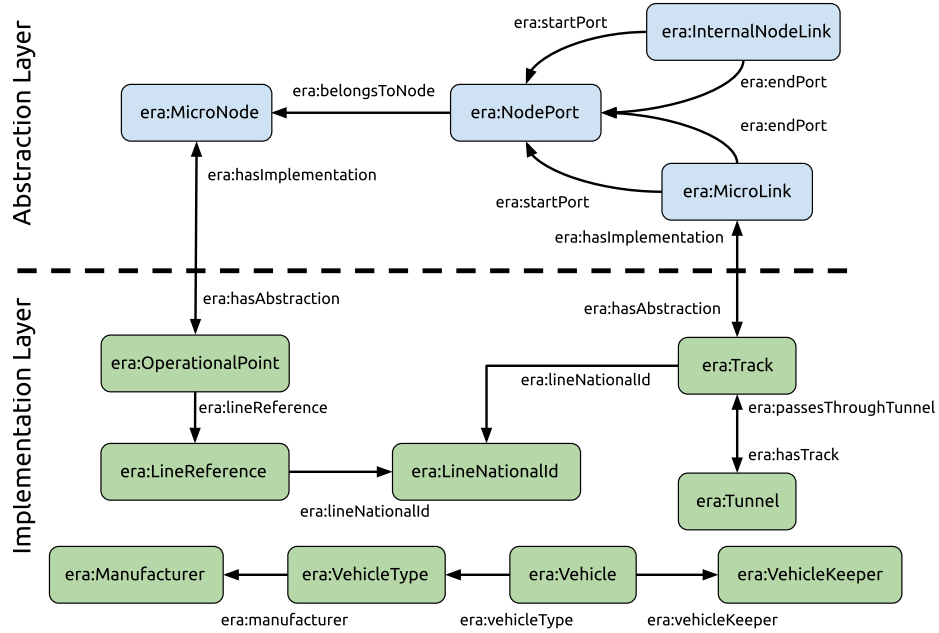
<sup>21</sup> <http://era.ilabt.imec.be/era-vocabulary/index-en.html>

<sup>22</sup> <https://github.com/julianrojas87/era-vocabulary/tree/master>

<sup>23</sup> <http://publications.europa.eu/resource/authority/country>

<sup>24</sup> <http://era.ilabt.imec.be/era-vocabulary/era-skos#>





**Fig. 2.** Layered data model of the ERA Vocabulary.

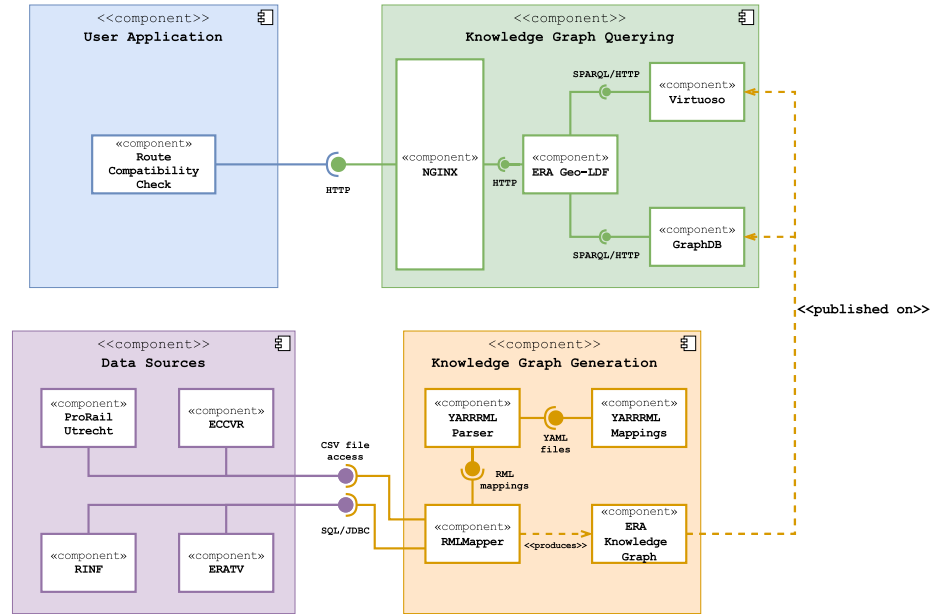
## 4.2 Architecture Overview

Our proposed solution architecture is composed by 4 main modules (see Figure 3), namely the *Data Sources*, *KG Generation*, *KG Querying* and *User Application* modules. The *Data Sources* module represents the considered data sources (previously described in Section 3). The components from the *KG Generation* module, access the data sources to produce the RDF triples that compose the ERA KG. The ERA KG is published and made available for querying by the *KG Querying* module, which provides the necessary interfaces for the *User Application* module to support specific use cases. Next, we provide a description and the rationale behind these modules.

**KG Generation** The KG generation process in our solution follows an ETL-based approach and uses the RML [13] technology stack for declaratively generating the RDF triples of the ERA Knowledge Graph. RML was selected for handling heterogeneous data sources, which in our case are relational DBs and CSV files, but XML Schema-based data sources (e.g., RailML) are also envisioned as a next step. The steps followed in this process are:

1. Definition of RML rules<sup>25</sup> in YARRRML [20] syntax.

<sup>25</sup> <https://github.com/julianrojas87/era-data-mappings>



**Fig. 3.** Overview of the proposed solution architecture for semantic data interoperability across ERA's base registries.

2. Translation of YARRRML rules to RML using the yarrml-parser<sup>26</sup>.
3. Production of RDF data via the RMLMapper<sup>27</sup>, according to the set of given RML rules.
4. Publishing of the resulting KG in a triple store. At the time of writing the ERA KG, had a total of 13.8 million triples, which we also make available as a raw data dump<sup>28</sup>.

**KG Querying** We published the ERA KG in two different triple stores (GraphDB<sup>29</sup> and Virtuoso<sup>30</sup>) to prove that our proposed solution is vendor-independent. This module includes one of the core components of the architecture: the ERA Geo-LDF, which is implemented as a Node.js application<sup>31</sup>. The main purpose of this component is exposing a Linked Data and Hypermedia-based API over the ERA KG. It builds on the Linked Data Fragments [27] approach to provide metadata

<sup>26</sup> <https://github.com/RMLio/yarrml-parser>

<sup>27</sup> <https://github.com/RMLio/rmlmapper-java>

<sup>28</sup> <https://drive.google.com/file/d/1KofPzYx2ovgAz85rLuO5J98SEs2BjWbO/view?usp=sharing>

<sup>29</sup> <http://era.ilabt.imec.be/sparql>

<sup>30</sup> <https://linked.ec-dataplatform.eu/sparql?default-graph-uri=https%3A%2F%2Flinked.ec-dataplatform.eu%2Fera>

<sup>31</sup> <https://github.com/julianrojas87/era-ldf>

annotated fragments (tiles) of the ERA KG, based on a predefined geospatial pattern. It follows the slippy maps specification<sup>32</sup>, where the grid-based partition of the world is specified based on a zoom level  $z$  and the  $x$  and  $y$  cartesian coordinates. A live example of a tile for the area of Brussels can be accessed on <http://era.ilabt.imec.be/ldf/sparql-tiles/implementation/10/524/343>.

The tiles are built by the ERA Geo-LDF component via template SPARQL queries that select and filter the entities based on their geospatial properties. In this way, client applications can request relevant data for their purposes and since the API returns unmodified triples from the KG, further querying and processing becomes possible on the client-side. Following this approach, we address the limitation of performing graph path finding queries directly on the SPARQL endpoints. Our client application implements a shortest-path algorithm and proceeds to download the relevant tiles based on the geospatial information given by origin-destination queries. Furthermore, tiles can be cached both on client- and server-side, which reduces the overall computational load on the server and improves query performance for client applications.

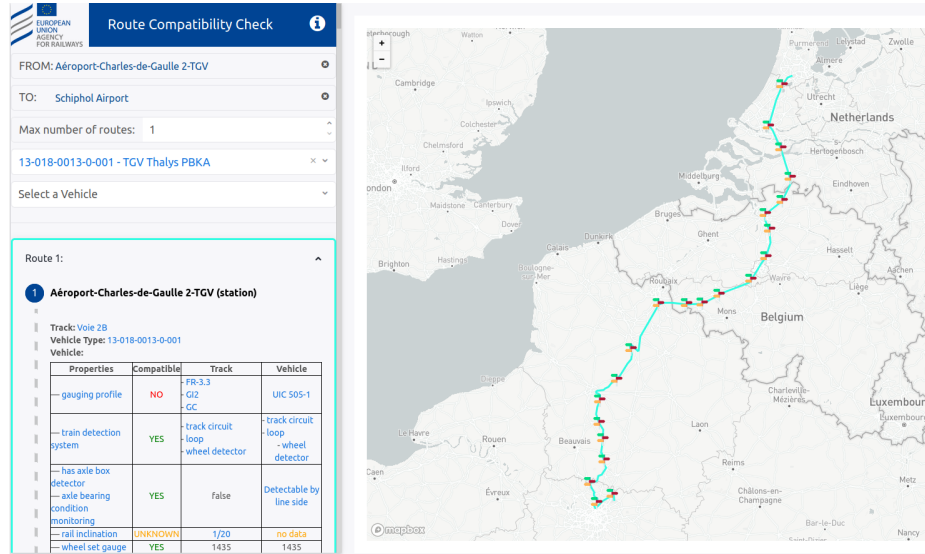
**User Application** This module represents any user-oriented applications that would perform querying tasks over the ERA KG to support a given use case. We developed a React-based Web application<sup>33</sup> for supporting the RCC use case and demonstrating data interoperability via the ERA KG. The application allows users to select an origin-destination pairs of operational points (visible in map-based UI) to calculate one or more routes between them. Once selected, it proceeds to download the relevant KG tile fragments and perform the path finding process. It handles RDF triples natively and implements the A\* [18] and Yen’s [29] algorithms for graph shortest path and top-k shortest path calculations respectively. Once a route is found, users may select a vehicle (type) to assess technical compatibility. Currently the application evaluates compatibility for 15 different parameters of both track sections and vehicle (types). Users can also visualize the internal connectivity of operational points that form part of a calculated route, by means of a schematic diagram that shows the possible internal connections defined in the ERA KG. This feature is particularly interesting for operational points around the city of Utrecht in The Netherlands, considering the additional data source from ProRail (Section 3), that was integrated into the KG.

## 5 Discussion

The implementation of our proposed solution allowed us to achieve semantic interoperability over the considered data sources, which stood as independent and disconnected data silos before. Our architecture relies entirely on semantic web technologies and tools, starting from the KG generation and ending with a RDF native Web application that supports the addressed RCC use case.

<sup>32</sup> [https://wiki.openstreetmap.org/wiki/Slippy\\_map\\_tilenames](https://wiki.openstreetmap.org/wiki/Slippy_map_tilenames)

<sup>33</sup> <https://github.com/julianrojas87/era-compatibility-check>



**Fig. 4.** The RCC Web application: a route calculated from the Charles de Gaulle airport in Paris to the Schipol airport in Amsterdam. On the lower left panel, the results of the compatibility check process for the TGV Thalys PBKA vehicle type.

## 5.1 Solution Features

Next we outline the main features of our proposed solution:

**Fully declarative KG generation** One key feature of our proposed solution relates to the ERA KG generation process, which is accomplished following a fully declarative approach. In other words, no pre-processing steps nor dedicated software/scripts are required to generate the RDF triples of the ERA KG. The KG generation rules are defined as RML mapping rules, which are executed by an existing and general purpose engine, that follows the given rules to produce the desired RDF triples. This feature has an important value from a data governance perspective, considering that no additional ad hoc software needs to be maintained. The RML mapping rules become the central resource for the ERA KG generation process, which can be adjusted or extended to include additional data sources, with significantly less effort compared to developing and maintaining additional software for every new data source to be included in the ERA KG. Furthermore, the mappings can be reused and adapted by IMs to produce their own internal KGs.

**KG enrichment flexibility** We were also able to explore an alternative solution to integrate additional data originated directly from an IM, to address the missing connectivity issue of the railway infrastructure. This approach demonstrated the flexibility that graph-based data models hold, considering that adding

additional data sources requires significantly less effort, than for example, altering a relational data model, potentially introducing breaking changes for the applications that depend on it.

**Cost-efficient KG publishing and querying** Our architecture design was made, with data publishing and querying cost-efficiency as a guiding principle. As described in Section 4.2, the ERA KG is published on triple stores with support for SPARQL querying. However the user application that supports the RCC use case does not perform direct SPARQL queries over these triple stores. Instead, it downloads specific parts of the KG via an API, over which it applies its business logic. Such an approach is no different to traditional REST-based application design over relational databases, where applications are given access to data via APIs only, and do not have unbounded querying access to the database(s) [3, 14]. In contrast to most API implementations, the APIs implemented in this architecture, follow the hypermedia constraints defined by REST, providing self-describing data responses via hypermedia metadata controls. In other words, the API data responses include additional metadata that describe how it can be used by client applications to retrieve more relevant data for a particular query. Such descriptions enable the creation of smarter and more autonomous client applications, avoiding the need of hard-coding the application according to specific API interfaces.

More importantly the API design in this architecture has been done to maximize the cacheability of API responses. By following a geospatial fragmentation approach, which suits the RCC use case, the API publishes fragments of the ERA KG that can be cached both on client- and server-side. This further reduces the computational cost on the triple stores, which only need to process once the query for a given fragment. A client application that requested a certain data fragment does not need to request it again (client cache) and has full flexibility to perform any type of further processing on the data it contains. When another client application needs access to the same type of data, it can rely on server-side cached API responses which also improve overall application performance.

**Shortest path querying over an RDF KG** The ability to indirectly support calculation of path finding queries, is an important feature of our architectural design. Our approach not only enables solving this particular type of queries, but also opens the door for clients to implement any path finding algorithm, and further customize them to better suit their requirements. Such level of specialization of algorithms is not always possible to be defined through general purpose query languages or it could potentially result in highly inefficient queries.

## 5.2 Limitations and Open Challenges

The identified limitations of our approach include:

**Performance of long distance queries** One of the main limitations of our proposed solution is related to the trade-off between server computational cost and query performance, that is introduced when shifting query processing tasks to the client. This is particularly visible when dealing with long distance route calculations, due to the increasing amount of data fragments that needs to be fetched and processed by the client. Different alternatives could be explored to address this limitation:

*Server-side route planning engine* : This is the most common approach followed by route planning solutions. It requires setting a dedicated engine (e.g., postGIS-based system<sup>34</sup>), which imports the whole topology graph and then is capable of executing a route planning algorithm over it. The drawbacks of this approach include the considerable increase of computational load for the server and less flexibility for client applications to select and tailor the algorithms for their own needs. But more importantly, available solutions do not support RDF data out of the box, which introduces an additional burden for the architecture by having to convert and keep in sync the ERA KG towards the required format of the route planning engine.

*Non-standard Graph Database* : Another alternative is to replace the standard RDF triple store by a graph database that has support for route plan querying (e.g. Stardog<sup>35</sup> or Neo4J<sup>36</sup> both with RDF support). Again the drawbacks of this approach are related to scalability and application flexibility, but they also may lead to vendor-locking issues, since they rely on non-standard solutions.

*Speed-up Techniques* : The application of speed-up techniques for shortest path algorithms, such as Contraction Hierarchies [12] or Multilevel Dijkstra [11], stands as a possible solution. These techniques rely on preprocessing steps that create summarizations of the graph topology, allowing to quickly compute long-distance path queries. They have been applied mostly to road networks graphs, where hierarchies of roads (highway, road, residential street, etc) can be used to create summaries for long distances. In principle they could also be applied to the railway topology graph. The drawbacks of these approaches are related to the introduction of additional complexity for creating the graph summaries that need to be managed and kept in sync with the original KG. However, they could still allow full flexibility for client applications to perform any business logic, since the summaries are only additional data that does not change the original RDF triples of the ERA KG.

**KG based on stale sources** The KG generation process is periodically performed over stale versions of the base registry relational DBs. To accurately

<sup>34</sup> <https://pgrouting.org/>

<sup>35</sup> <https://docs.stardog.com/archive/7.5.0/query-stardog/path-queries>

<sup>36</sup> <https://neo4j.com/docs/graph-data-science/current/algorithms/dijkstra-source-target/>

reflect the real state of the railway network, is necessary to capture in *real-time* the changes introduced into the source DBs, and immediately reflect them in the ERA KG. Other use cases such as signalling and interlocking, require precise and accurate data to guarantee safe vehicle operations. Approaches such as Linked Data Event Streams<sup>37</sup>, remain to be investigated to support this requirements.

**Hardcoded compatibility check rules** The compatibility check rules, were directly implemented into the source code of the RCC client application. This constitutes a limitation, given that it makes more difficult to maintain and evolve the rules. Also, it makes the rules to be indistinguishable from the application, hindering their potential reusability in other use cases. Alternatives to address this issue could explore the use of Notation3 or SHACL Rules to declaratively define the RCC rules, which can be then independently managed and published for applications such as the RCC client to consume and evaluate.

## 6 Conclusion and Future Work

The most important achievement of this work, is the strong impact it had on the decision taken by ERA<sup>38</sup> to make Semantic Web technologies the default setting for any future development of data, registers and specifications, under the agency’s remit. Considering ERA’s position as an European authority this decision could potentially influence the different stakeholders in the railway domain, to take similar paths.

The results obtained from this work, demonstrated with a practical approach, how Semantic Web technologies enable higher data interoperability. Data integration is achieved at the *data level* (data-centric) instead of being locked into application-specific business logic (application-centric), opening the door for new and innovative use cases. We were able to create a semantic interoperability layer over the different considered data sources, which requires significantly less effort to be created and managed, compared to developing ad-hoc applications and 1-to-1 interfaces between different information systems. Furthermore, this work also demonstrated that Semantic Web technologies can be used to create functional Web applications based on modern and developer-friendly frameworks such as React with little additional effort from a development perspective and in a reasonable time frame.

The choice of architecture design made for this prototype leverages HTTP caching mechanisms to achieve higher scalability while providing full querying flexibility to client applications. This is demonstrated by the ability of the RCC client application to perform route planning calculations over the ERA KG, which are not supported by standard RDF triple stores. Yet, this approach establishes a trade-off between scalability and flexibility vs. performance. Further

<sup>37</sup> <https://w3id.org/ldes/specification>

<sup>38</sup> [https://www.era.europa.eu/sites/default/files/agency/docs/decision/decision\\_n250\\_annex1\\_linked\\_data\\_en.pdf](https://www.era.europa.eu/sites/default/files/agency/docs/decision/decision_n250_annex1_linked_data_en.pdf)

optimizations are required to achieve production-level performance without losing the benefits of the proposed solution architecture.

In the future, we aim to explore how more granular descriptions of the railway topology can be integrated, to increase the reliability of the ERA KG. From an architectural perspective, stream-processing and KG virtualization approaches may be studied to support cases with higher requirements on up-to-date data.

## Acknowledgements

The authors would like to extend their gratitude to ProRail, SNCF, BANE NOR, EIM, UIP, CEDEX, RailML, EULYNX, the Publications Office of the EU and the ELISE action team for providing us with their invaluable data, expertise and feedback to make this work possible.

## References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. *Scientific American* **284**(5), 34–43 (May 2001)
2. Bischof, S., Schenner, G.: Towards a railway topology ontology to integrate and query rail data silos. In: SEMWEB (2020)
3. Chaudhuri, S., Weikum, G.: Rethinking database system architecture: Towards a self-tuning risc-style database system. In: Proceedings of the 26th International Conference on Very Large Data Bases. p. 1–10. VLDB '00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2000)
4. Council of European Union: Council regulation (EU) no 2008/57 (2008), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32008L0057>
5. Council of European Union: Council regulation (EU) no 2011/65 (2011), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32011L0065>
6. Council of European Union: Council regulation (EU) no 2012/34 (2016), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32012L0034>
7. Council of European Union: Council regulation (EU) no 2016/796 (2016), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0796>
8. Council of European Union: Council regulation (EU) no 2016/797 (2016), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016L0797>
9. Council of European Union: Council regulation (EU) no 2018/1614 (2018), [https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L\\_.2018.268.01.0053.01.ENG&toc=OJ:L:2018:268:TOC](https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2018.268.01.0053.01.ENG&toc=OJ:L:2018:268:TOC)
10. Council of European Union: Council regulation (EU) no 2019/773 (2019), [https://eur-lex.europa.eu/eli/reg\\_impl/2019/773/oj](https://eur-lex.europa.eu/eli/reg_impl/2019/773/oj)
11. Delling, D., Goldberg, A., Pajor, T., Werneck, R.F.: Customizable route planning in road networks. *Transp. Sci.* **51**, 566–591 (2017)
12. Dibbelt, J., Strasser, B., Wagner, D.: Customizable contraction hierarchies. *ACM J. Exp. Algorithmics* **21** (Apr 2016)
13. Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: RML: a generic language for integrated RDF mappings of heterogeneous data. In: Proceedings of the 7th Workshop on Linked Data on the Web. CEUR Workshop Proceedings, vol. 1184 (Apr 2014)



14. Fielding, R.T., Taylor, R.N.: Principled design of the modern web architecture. *ACM Trans. Internet Technol.* **2**(2), 115–150 (May 2002)
15. Garijo, D., Geluk, J., Scharm, M., Ruiz-Iniesta, A., McBennett, P., Serafini, A., María, kartgk, Corcho, O., rpietzsch, Schneider, J., Leitschuh, J., Scrocca, M., Lefrançois, M., Garcia, M.A., Zack-83: dgarijo/Widoco: WIDOCO 1.4.15.1 (pre-release): Namespace prefixes fixes and WebVowl update (Dec 2020)
16. Gély, L., Dessagne, G., Pesneau, P., Vanderbeck, F.: A multi scalable model based on a connexity graph representation. *WIT Transactions on the Built Environment* **114**, 193–204 (2010)
17. Gubichev, A., Neumann, T.: Path query processing on very large rdf graphs. In: *WebDB* (2011)
18. Hart, P.E., Nilsson, N.J., Raphael, B.: A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics* **4**(2), 100–107 (1968)
19. Heath, T., Bizer, C.: *Linked Data: evolving the Web into a global data space* (2011), <http://linkeddatabook.com/editions/1.0/>
20. Heyvaert, P., De Meester, B., Dimou, A., Verborgh, R.: Declarative Rules for Linked Data Generation at your Fingertips! In: *Proceedings of the 15<sup>th</sup> ESWC: Posters and Demos* (2018)
21. Hlubuček, A.: Railtopomodel and railml 3 in overall context. *Acta Polytechnica CTU Proceedings* **11**, 16 (08 2017). <https://doi.org/10.14311/APP.2017.11.0016>
22. RailTopoModel - Railway Infrastructure Topological Model. Standard, International Union of Railways, Paris, FR (2016)
23. Morris, C., Easton, J., Roberts, C.: Applications of linked data in the rail domain. In: *2014 IEEE International Conference on Big Data (Big Data)*. pp. 35–41 (2014)
24. Publications Office of the European Union: New european interoperability framework: Promoting seamless services and data flows for european public administrations (2017), [https://ec.europa.eu/isa2/sites/default/files/eif\\_brochure\\_final.pdf](https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf)
25. Savenkov, V., Mehmood, Q., Umbrich, J., Polleres, A.: Counting to k or how sparql1.1 property paths can be extended to top-k path queries. In: *Proceedings of the 13th International Conference on Semantic Systems*. p. 97–103. *Semantics2017*, Association for Computing Machinery, New York, NY, USA (2017)
26. Tutchter, J., Easton, J., Roberts, C.: Enabling data integration in the rail industry using rdf and owl: The racoon ontology. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering* **3** (2017)
27. Verborgh, R., Vander Sande, M., Colpaert, P., Coppens, S., Mannens, E., Van de Walle, R.: Web-scale querying through Linked Data Fragments. In: Bizer, C., Heath, T., Auer, S., Berners-Lee, T. (eds.) *Proceedings of the 7<sup>th</sup> Workshop on Linked Data on the Web*. *CEUR Workshop Proceedings*, vol. 1184 (Apr 2014)
28. Verstichel, S., Ongenaes, F., Loeve, L., Vermeulen, F., Dings, P., Dhoedt, B., Dhaene, T., Turck, F.D.: Efficient data integration in the railway domain through an ontology-based methodology. *Transportation Research Part C: Emerging Technologies* **19**(4), 617–643 (2011)
29. Yen, J.Y.: Finding the k shortest loopless paths in a network. *Management Science* **17**(11), 712–716 (1971)