Applications of Linked Data in the Rail Domain

Christopher Morris*, Dr John Easton, Prof Clive Roberts
Birmingham Centre for Railway Research and Education
University of Birmingham
Birmingham, United Kingdom
Cxm210 @ bham.ac.uk
*Corresponding author

This paper presents early findings from a larger study, into the use of linked data in the rail domain. The study and other literature has shown there to be benefits from improved integration of data in this domain and proposes that linked data in general and ontology in particular will address this. The paper will set out the current state of data integration in the British rail domain, highlighting issues found there. The manner in which linked data is employed in the broader transport domain will then be examined along with previous work pertaining to the rail domain.

Keywords—Linked Data; Semantic Data; Ontology; Railway data;

I. INTRODUCTION

The British Rail industry works as one large system, with a diverse range of stakeholders collaborating to produce a coherent whole. Ridership on the UK rail network is growing, as evidenced by statistics from the UK office of Rail Regulation [1] the number of UK passenger kilometres travelled trending strongly upwards. However, recent public reviews [2] of the industry have found that there remains a need for improved information sharing throughout the industry.

The Technical Strategy Leadership Group [3] has reported that “Excluding Network Rail’s own information systems, research discovered over 130 information systems maintained by approximately 20 suppliers were in operation in 2011”. The same document then goes on to identify the benefits of integrating that data – primarily centred on the reduction of costs and duplicate effort ([3], part 2.124). Previous work, such as [4] has noted that most existing formats are proprietary, locking people into a single vendor and preventing data interchange. This point is made again in [5] which states that “most data are archived for “future use” and never looked at, unless a specific need occurs”. These isolated information systems are often referred to as “information silos” or “data stovepipes” in the literature. When [6] examined this field it was found that “a system-wide data framework for UK rail, preferably in combination with wider data sharing between stakeholders, does have the potential to both improve asset management and encourage a shift in transport modes towards the railways”.

Verstichel in [7] – an extention of the inteGRail project - outlines how the need for better integration of information systems is made all the more urgent by the increased number of actors in the rail sector post privatisation. The steadily growing amount of information communication technology in use in the rail sector is another key factor. The report [7] adds that an accessible data model is advantageous in that many teams can develop simultaneously with it. As stated by the authors: “It is after all undesirable and nearly impossible to centralise application development in world-wide fragmented and large systems, such as the railways.” Other benefits of integrated information systems are identified in [8] and include a reduction in CO2 emissions brought about by a shift to rail transport. This in turn will be driven by the benefits it will be possible to offer consumers, both in terms of better information and journey planning, as well as a reduction in long term costs.

This project aims to meet the needs of the rail sector by building upon the Rail Domain Ontology currently being developed by the Birmingham Centre for Railway Research and Education. This in turn will allow for better data integration across the rail domain. The key parts of this are to represent existing data in the Rail Domain Ontology and create tools to allow non-ICT expert members of staff to query the ontology.

Other aspects of the research will include the appropriate form for an ontology driven user interfaces to information systems. Search with an auto-complete function or visual methods are both to be considered as solutions at this stage. In order as to accomplish this it will be necessary to select tools for making existing data available semantically and map the existing data onto the rail domain ontology.

II. WHAT IS AN ONTOLOGY?

In [9] an ontology is defined thus: “An ontology defines a set of representational primitives with which to model a domain of knowledge or discourse”. Previously in [10] ontology was defined as “An ontology is a specification of a conceptualization”, which has since been heavily cited. “Knowledge” in itself also requires definition - this differs from data by having not just meaning but also context, though it must be noted that much philosophical debate exists as to the exact definition of each of these terms. This report will use the definition provided in [11] namely: “[knowledge is] information that has been made part of a specific context and is useful in this context”. Other disciplines would define ontology...
differently, [12] suggests: “In philosophy, ontology is the study of the kinds of things that exist”.

Terms such as “Vocabulary” and “Semantic Data” are also used to refer to different types of data with meaning. The exact dividing line between these two terms and an ontology has yet to be defined: “There is no clear division between what is referred to as “vocabularies” and “ontologies””[13], however an ontology will contain a vocabulary. In this context a vocabulary continues to have the same meaning as in day to day English; that is a list of terms, pertinent to a given domain. Semantic Data is defined as data with meaning as discussed in (Shadbolt, et al., 2006). A good explanation of semantic data can be found in Roberts et al in under the heading “Conceptual Data Models and the Preservation of Semantics”.

Linked data is defined by the World Wide Web Consortium as [14] “[the Semantic Web needs] access to data, but relationships among data should be made available, too, to create a Web of Data (as opposed to a sheer collection of datasets).” Data can then be stored as linked data, without necessarily being part of a full ontology.

III. THE CASE FOR ONTOLOGY IN THE UK RAIL DOMAIN

Other industries such as biomedical research [15], the power distribution industry[16], the media [17] and petro-chemical industry [18], have adopted ontologies successfully, to resolve the problems similar to those outlined in the introduction.

Biomedical research has taken the lead in the use of Ontology. The “The National Cancer Institute’s Thésaurus and Ontology”, described in [19], has been operating successfully for a number of years. This has over the years served to bring together the large amount of data required for biomedical research. Over time it has grown to cover an increasing part of the Biomedical domain as concluded by [20] “Cancer research and clinical practice increasingly require tight integration of large amounts of molecular and clinical data. The NCI Thesaurus has been extended to support such integration.” The same report also discusses the trade-off between scalability and expressivity which is a key issue in many industries and will need further investigation as part of the study.

As examined in [16] the power distribution industry is also using linked data for interoperability. That industry also faces the challenges of a large number of stake holders working together on the same system, which has strong parallels to the rail domain.

In the petro-chemical industry, ISO15926 has seen the supply chain and the design process simplified, by removing the need to manually scan large numbers of datasheets and rekey the information they contain. This in turn reduces the scope for error and increases the scope for automation.

In the case of the media, in particular the British Broadcasting Corporation, this is used to integrate their public facing data and program information.

Directly rail related benefits are suggested in [7] namely: improved route planning, better maintenance and better passenger information. As outlined by Umiliacchi in [21] further benefits to this domain could be achieved through the integration of maintenance and condition monitoring data. Rail vehicles, as hierarchical systems lend themselves well to representation as an ontology or as a taxonomy. The report goes on to discuss the benefits available in terms of fault prediction and thus preventative maintenance, which brings financial benefits. These come both in terms of the avoidance of vehicle failure and hence corrective maintenance and associated costs along with the avoidance of unnecessary replacement of working parts. This can be combined with data integration to take data from suppliers (say the correct values for the sensors) and then when the observed values fall outside a (also supplied specified) envelope report that a fault is likely.

A related area which would benefit from better integrated data is rail infrastructure maintenance. Part D5.1 of the AUTOMAIN project [22] attempts to improve maintenance timetabling, using predictive maintenance techniques and estimated durations for maintenance tasks. The methodology shows great improvement on the state of the art, however, in order to apply that methodology to the real world there is a need for timetabling, infrastructure, parts and condition monitoring data to be brought together as one data source. This process can be achieved very effectively with linked data but is challenging without. As made clear by [23] this would in turn offer reduced possession time for maintenance, leading to an improvement in the number of available train paths.

There are additional, more generic, advantages for software re-use; [24] points out that it is possible to design an application once and use it with different ontologies for different use cases. Furthermore the separation between data and the front end, commonly regarded as a key goal of good software design, can be far more complete. As a result much of the business logic that would otherwise have to take place in the application can instead take place in the ontology. Thus when a new feature must be added or the front end redesigned to work on a new platform the data source does not need to change. Similarly when some part of the reasoning changes – for example the conditions under which a fault should be reported - it is only necessary to update the ontology, rather than rolling out an entire new application.

Other sources such as [25] point out that ontologies can be used as an aid to software design. Use of ontologies for the design of rail traffic management software is discussed in [26]. That study is focused primarily on using ontologies as tool for the verification of software requirements, as a part of the software design process. Whilst the primary focus of this study is on data integration, the development of an ontology for this domain would certainly be of great benefit to those working in the area of software verification.

Other benefits of linked data in the rail domain are identified in [8]:

- Operations Planning (Long Term and real-time);
- Integrating condition monitoring data for track and train and vice versa;
- Increasing route utilisation (increasing usage of existing track; optimising track to support increased usage);
IV. EXISTING WORK IN THIS AREA

Work has already been done in the field of ontology and linked data as applied to the rail domain. Projects of note include InteGRail, outlined in [5], and work derived from the ArkTRANS project, such as MultiRIT outlined in [27]. One of the key benefits of linked data is improved data integration and a key tenet of achieving that is the reuse of terms from existing vocabularies, as set out in [28] “If suitable terms can be found in existing vocabularies, these should be reused to describe data wherever possible, rather than reinvented.” In the case of the rail domain these benefits come from integration with data from other countries and modes of transport, where appropriate. As such it is necessary to examine the coverage of this domain given by existing ontologies.

Another benefit of examining Ontologies that cover this domain (and the projects that created them) is to learn from their methodology - all projects listed in Table 1 other than the SUMO project have parts beyond an ontology. Table 1 below compares ontologies which cover the transport domain in general and sets out how they cover the rail domain. Many of these ontologies make use of broader linked data technologies such as - RDF: Resource Description Framework – defined by W3C [29] as “The Resource Description Framework (RDF) is a framework for representing information in the Web”, in effect a format for storing linked data. Some also use OWL: Web ontology Language - one of several ways of expressing ontologies such that they can be stored and used electronically. OWL can itself be expressed in RDF.
<table>
<thead>
<tr>
<th>Project</th>
<th>Purpose</th>
<th>Relevance</th>
<th>Gaps</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>REWERSE</td>
<td>Very broad project relating to ontology and reasoning. In particular part A1-D4 “Ontology of Transportation Networks” is discussed here.</td>
<td>Rail path information, in particular as relevant to the spatial domain. Rail interchange information as relating to the multi-modal transport domain. Rail stations etc are also covered.</td>
<td>No Timetabling information. No Detailed rail information, such as infrastructure or Rail Vehicle components</td>
<td>(Lorenz, 2005), ISO 14825:2004</td>
</tr>
<tr>
<td>ArkTRANS</td>
<td>Serves the multimodal transport domain.</td>
<td>Rail transport as part of the multimodal transport domain Applies roles so as not to be tied to any given stakeholder Many projects extend this project (not listed separately in this table)</td>
<td>Intent to cover rail as a mode of transport, not in sufficient detail for maintenance etc. Still very much in development – technical detail hasn’t been implemented for some areas, including rail</td>
<td>(SINTEF, 2010)(M. Natvig &amp; Westerheim, 2008)</td>
</tr>
<tr>
<td>OZONE \ DITOPS</td>
<td>Primarily Scheduling, including transportation scheduling</td>
<td>Scheduling – in particular service scheduling</td>
<td>Limited data has been converted to this format No Detailed rail information, such as infrastructure or Rail Vehicle components</td>
<td>(Becker &amp; Smith, 1997)</td>
</tr>
<tr>
<td>EURIDICE</td>
<td>Adding intelligence to cargo</td>
<td>Some intersection with the rail domain. Containerised freight in particular</td>
<td>No coverage of the passenger domain. No coverage of rail infrastructure.</td>
<td>(Schumacher, Gschweidl, &amp; Rieder, 2010),(Paganelli et al., 2009)</td>
</tr>
<tr>
<td>InteGRail</td>
<td>An ontology aimed directly at the rail domain</td>
<td>Aimed directly at the rail domain Has some rail network data mapped Has a vocabulary for the majority of the commonly used rail data</td>
<td>Limited data has been converted to this format</td>
<td>(Köpf, 2010)</td>
</tr>
<tr>
<td>SUMO</td>
<td>A very broad upper ontology, which covers (amongst many others) the transport domain</td>
<td>An Upper level ontology Contains the highest level transport concepts Can help integrate ontologies Other derived projects look at the transport domain in greater detail</td>
<td>No fine coverage of the rail domain</td>
<td>(Niles &amp; Pease, 2001)</td>
</tr>
</tbody>
</table>
A number of the projects outlined in Table 1 have been expanded upon by further projects, or themselves build upon the work of past projects.

A. REWERSE

The REWERSE project, aims “to establish Europe as a leader in reasoning languages for the Web”, as stated by [30] and sets out to establish upper ontologies and rules languages for various domains.

The part of the project of relevance to this study is “A1-D4” which is the transport ontology. As set out in [31] this builds upon past standards, notably that for Geographic Data Files, which is itself a linked data format. This standard defines a standard set of terms, such as “road” and standardises ways in which these terms maybe interconnected. Definitions are split into three levels: topology, route planning and turn by turn driver instructions. Even before adoption by the REWERSE project this has been widely adopted by the mapping community. It is now encoded in an ISO standard, ISO 14825:2004. Of most relevance to this study is the information it encodes relating to the (geographic) path railways take and the location of railway “Features” including junctions. Also modelled are points where one can change transport mode in a multimodal system, for both passengers and freight.

As stated in [31] “Whereas roads are modelled in great detail, the modelling of rails is quite course”. Public transport systems in general (including rail) are modelled separately to the ground they cover (both are modelled, as are the links between them). The model would provide enough detail for the most basic passenger information, (train times, station locations, other public transport to and from those stations) but wouldn’t serve to store, for example, rail maintenance information.

Also part of this project is an ontology of traffic networks based on the ontology from GDF. This is both extended and encoded in a common ontology language – OWL - using an open source tool; protégé. It takes the following as base classes:

- Feature
- Geometric the geometric form of a feature
- Composite_attributes
- Relationship – a non-geographic relationship
- Transfer_point – this includes railway stations, car parks, bus stations etc.

One, minor, limitation of REWERSE is differentiating taxis and private cars – whilst a private car must find somewhere to park (and car park locations are covered by the ontology of transport networks) taxis have no such need. This isn’t yet modelled.

B. ArkTRANS

ArkTrans provides a framework and a reference model for Multimodal transport data. A large number of models based on this project have gone into production, serving both the multimodal transport domain and the freight transport domain. Its aim as stated by [32] is “[to provide] a holistic and mode-independent understanding of the transport sector.”

Arktrans aims to be independent of stake holders and the implementation technology, as such it uses a series “roles”. These are independent even of the mode of transport for example it defines the road domain as having a “Driver” role and defines this as being equivalent to the “Helmsman” roll in the maritime domain and the “Engine Driver” roll in the rail domain. In this vein each domain is described, by [33] to have “Transport Network Management” comprising:

- Transport Network Infrastructure Manager
- Transport Network Utilisation Management
- Emergency Management
- Regulation Enforcement

Of the projects that take ArkTRANS as a starting point most focus on The Freight or Multimodal transport domains, though the very accessible www.nsb.no deals with the rail domain – NSB is the Norwegian state passenger rail operator. MultiRIT, as set out in [34] is the key project for Multimodal passenger transport based on and developed on the principles set out in ArkTRANS. It is aimed at 3rd party travel information providers, of which NSB is but one. MultiRIT aims to work with all modes of transport as such it translates the “Driver” role found in ArkTRANS to be “Road user”, which it then specialises as “Cyclist”, “Pedestrian” and “Driver”. ArkTRANS aims to serve the Multimodal transport domain primarily in a journey planning capacity. This includes information such as accessibility - which goes beyond merely “Disabled accessible” and “not” - information such as audio description of stations and available of step-free access for example. Also supported is intelligent re-planning of journeys as the situation changes.

In order as to support both freight and passenger transport ArkTRANS has a concept called “Transport Item”. This can have needs e.g. goods can need a given, constant, temperature, people can need a power socket for their laptop. This allows the framework to be extended not only for passenger use but also freight. A number of projects make use of the generic nature of the ArkTRANS transport item concept; SMARTFREIGHT as described in [35], Logistics for Life set out in [32] and E-Freight final report: [36] being notable examples.

All these projects have in common taking the ArkTRANS framework as a starting point and extending it into the freight domain. The projects objectives differ; SMARTFREIGHT is concerned with road freight in an urban context, Logistics for Life is aimed at the broader freight domain and E-Freight project is broader still, interested in linked data for the
transport domain. All of these projects, but in particular E-Freight employ linked data to certain extent.

In conclusion then ArkTRANS is a large and very developed project which does not cover the rail domain in any great detail. As such it may well be wise to integrate with it, in particular for applications in the multimodal transport domain, but it can’t of itself serve as a single source of truth for the rail domain.

C. inteGRail

The aims of inteGRail project come closest to matching those of this study. As set out in [5] the objectives are: “Sharing information between IMs [Infrastructure Managers] and RUs [Railway Undertakings] allows the whole railway to be managed as a single system”. inteGRail worked towards establishing an ontology for this domain and produced a number of demonstrators based upon it. Left outstanding was the question of migrating existing data to the linked format. That question was studied by [24] which took as a starting point the Dutch and Belgian rail networks and produced a “network statement checker” A tool to check whether a given train could take a particular route.

In the case of the Belgian rail network the network description was available as a traditional relational database. Using automated Database to Relational (D2R) tools the data was presented as linked data, in the format suggested by inteGRail. This had the advantage of allowing the existing system to remain in use, for those systems that rely on it. At the same time it makes a real time, constantly updated, copy of the data available for other systems to use as linked data. The only downside of this is performance. The study compares data stored as a native ontology, data stored in a relational database and data accessed via D2R tools. D2R is found to be orders of magnitude slower, a conclusion that is borne out by [37]. Since this market has moved on considerably since the study, it would bear further investigation as to whether this continues to be the case. As stated in the report however; “in the context of non-time-critical applications, such as the Network Statement Checker, this extra overhead introduced by this approach is worth the effort compared with the increased transparency and adaptability of the suggested approach towards future systems”.

In the case of the Dutch system the data was only available as paper books, which had to be manually entered. Since this was done specifically for the project it was manually entered into a spread sheet and then converted to semantic data. In this context, semantic data is simply data with meaning, expressed as a triple. It is worth pointing that, as was the case here, semantic data can use a regular relational database as a triple store, though it needs layers on top of that to present it correctly.

D. SUMO

SUMO is a very high level ontology that exists to allow the integration of other ontologies. As set out in [38] it was formed by combining five upper level ontologies. In itself SUMO covers the transport domain at only the very highest level with very limited coverage of rail as a part of that. It is however worth noting the work presented in [39] in which a detailed ontology of the road network was created, taking SUMO as a starting point. Whilst not the strongest candidate for integration in itself, the principle used here – taking an upper ontology and extending it so as to be compatible with other ontologies is used elsewhere.

E. Summary of Existing Work

The most directly pertinent of the previous work is the inteGRail project. Both the lessons and the ontology itself, from that project will be incorporated into this project. Other projects are however not to be ignored – the possibilities of integration with ArkTRANS or the REWERSE projects for multimodal transport should be borne in mind.

V. CONCLUSION

In conclusion there is a strong case for better data integration within the rail domain. In business terms, this comes from the added value brought by combining different data sources. Condition monitoring, both vehicle and infrastructure, provide many good examples of this, where by conclusions regarding the condition of an asset can better be drawn using the combined sources than any one. This can in turn be used to reduce maintenance costs. Other examples are available in the domain of passenger information and multimodal transport, where by more useful data can be provided by, for example, combining train arrival data with the data on the geographical location of rail and bus stations. This can greatly enhance the passenger experience and help bring about an increase in rail ridership.

Beyond the broader, more strategic, aims there is also the possibility of reducing costs, by the elimination of needless rekeying of data and the prevention of the errors that naturally induces. This saving will, in part, be offset by the upfront cost of software (be it bespoke or commercial off the shelf software) however the software is largely a onetime capital expense whereas personnel are an ongoing cost.

This may be accomplished with the use of Linked Data with further benefits available from using a more complete ontology. Much previous work exists in the general transport domain, in particular the freight and multimodal transport domains with less emphasis having been placed on the rail domain to date. Keys goals of this PhD project will be the conversion of existing data to a linked data format in line with the ontology already designed at the Birmingham Centre for Rail Research and Education along with modifying and extending the ontology as required.

The challenges foreseen include presenting the available data in a linked format and getting buy in from all stakeholders. Another issue will be striking the correct balance between the expressivity of an ontology and the speed of access of less expressive linked data formats.

REFERENCES
