Rail Research UK Feasibility Account (FA):

Factor 20 – reducing CO_2 emissions from inland transport by a major modal shift to rail (EP/H024743/1)

The Specification of a System-wide Data Framework for the Railway Industry – Final Report

Lead Investigator

Professor Clive Roberts (University of Birmingham)

Feasibility Study Team / Partners

Professor Clive Roberts, Dr John Easton, Rhys Davies (University of Birmingham)

Dr Sarah Sharples, Dr David Golightly (University of Nottingham)

The feasibility study team wish to thank all those who kindly volunteered their time to attend the workshop sessions.

Executive Summary

This report summarises the findings of the RRUK-A, Factor 20 project feasibility study into the "specification of a system-wide data framework for the railway industry". The study was designed to determine whether a cross-industry, semantic data exchange format would be of value to the UK railway industry with respect to better asset management and patronage, leading to reduced CO_2 emissions from the transport sector. In its investigation of this question, the study team held two workshops in which members of the industry were invited to, amongst other activities, propose use cases where they felt a system-wide data framework would have value. This resulted in over 150 separate applications for such a framework being identified. The workshop participants were also asked about perceived problems with the adoption of a common data framework within the industry; amongst the responses issues relating to the ownership and security of data, levels of IT competence within the industry, and the difficulties involved with the generation of a business case were felt to be the most significant. Priority areas for framework development were seen to be the provision of geographical data at varying levels of granularity, and real-time operational data. The provision of historical information, while important, was not seen as a critical component of an initial data framework.

Following the workshops, a number of example conceptual data models were developed to support some of the use cases proposed. This exercise highlighted a number of potential pitfalls to any wider data modelling exercise within the industry, including the difficulties associated with obtaining the required domain knowledge, and the need for greater than expected industry involvement. As an output from this work, a revised roadmap for future model development activities within the industry was proposed.

In conclusion, the study 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, thereby leading to reduced CO_2 emissions from the transport sector.

Table of Contents

Introduction

As increasing numbers of studies point towards carbon emissions being a key factor in global warming, the need for businesses and individuals alike to carefully monitor their carbon footprint is clear. Transportation, and in particular the air travel industry, is a major producer of carbon dioxide (CO_2) and therefore one way in which carbon emissions may be reduced is through a shift in travel mode, where possible, from cars and aeroplanes to lower carbon alternatives such as rail. However, before such a change can take place, railway stakeholders at both national and international levels must address the incompatibilities and performance issues within railway networks; producing a more integrated, efficient and harmonised service both from the point of view of railway operations and the quality of service experienced by passengers.

The RRUK-A Factor 20 Project

Factor 20 is the umbrella term for a group of 6-month feasibility studies performed by the Rail Research UK Association (RRUK-A), a partnership between UK universities and rail industry stakeholders. The project, which was funded by the Engineering and Physical Sciences Research Council (EPSRC), was set up with the aim of investigating a range of methods by which a modal shift in inland transportation methods towards rail could be encouraged, thereby leading to an overall reduction in CO_2 emissions across the UK.

The feasibility studies performed as part of the Factor 20 project covered a diverse set of topics, ranging from the reduction of headway between vehicles, through visions of the potential shape of the rail freight market in 2035, to ways in which passenger comfort on overcrowded commuter trains could be improved. Alongside these was a study that was designed to determine whether the specification of a cross-industry, semantic data exchange framework would be of benefit to the industry with respect to better management of assets, leading to improved energy efficiency and increased patronage; it is the results of that work that are reported in this document. The data management study had three main aims; these were to:

1. Investigate the benefits of developing a standardised, cross-industry, system-wide capability of integrating data, which could be used in the UK, and potentially throughout the world;
2. Construct well researched and supported recommendations towards the required standardisation effort the industry will need to undertake and develop a road map for the railway industry;
3. Develop a prototype data model that demonstrates the concepts behind such a model and illustrates how the railway system could become more efficient, and hence more attractive to the travelling public and freight operators, through the better use of existing data sources.

Data Management in the UK Railway Industry

With stored data volumes globally growing at rates estimated to be up to 60% per annum but the capacity to process that data growing at a much slower rate, and the US stock exchange making more money from the provision of stock information to traders than it does from the trading of stocks, it is probably safe to say that data is a greatly undervalued resource in may large businesses. Indeed, the provision of the correct information to the right people at the right time and presented in a way that supports their working practices may give a business a huge competitive edge over its rivals. Unfortunately the reverse of this is also true, and as the recent McNulty report into value for money within the UK rail industry (DfT, 2011) found "…the effectiveness of the industry's IS is inhibited by a suite of legacy systems that are expensive to run, unable to communicate with new technology and encourage users to develop a wide range of bespoke local systems to overcome limitations. Many legacy systems were created and managed in company silos, with only a few systems crossing industry boundaries." This situation is beautifully exemplified by the case of Network Rail, who the Rail Safety and Standards board (RSSB) have reported as managing over 1,500 different IT systems (DfT, 2011), many of which operate either in complete isolation or via manual system interfaces.

In fairness to Network Rail, the somewhat lamentable state of IT within the railway industry isn't solely their fault and there are, in actuality, a number of contributing factors, most of which are rooted far further back that Network Rail's stewardship. To begin with, many of the "core" legacy systems such as TOPS and TRUST that are used to record vehicle information and to route trains around the country are decades old, and were never designed to work in conjunction with other packages. Over the years those systems have been built upon to the extent that their replacement has become very difficult and hugely expensive, largely due to their safety-critical roles and the heightened risk of unforeseen dependencies that other software may have on them; it is the logistical issues associated with updating legacy systems, particularly when coupled with the more general reluctance at a national / governmental level in the wake of the recent failure of the NHS national IT system, which makes the decision to replace a "working" IT system a hard one to justify.

The privatisation of the rail industry in the mid-90s is also unlikely to have helped the situation, encouraging siloed thinking during a period when other industries were beginning to integrate their IT systems in readiness for the new (at the time) Web technologies. The fractured nature of the industry post-privatisation now means that it is hard to build a business case for wider data sharing, since while many members of the industry admit that there are benefits that could be realised, a whole-system benefit cannot easily be accounted for in individual company balance sheets.

Aside from technical and financial issues, one key question when considering any future data sharing system is of course whether such a system is even needed. The industry seems to be split on this issue; some industry members, when asked if data sharing is needed, state that those stakeholders who need access to a particular information source already have access to that data and that wider sharing of the data is just an unnecessary complication, wasteful of time and expensive to implement. Others however take the viewpoint that while railway undertakings generally understand and know how to work with the data they currently have access to, the reason that they don't know they need accesses to other data sources is that they don't know the data sources exist; if they knew of the existence of other relevant sources, then they might use them.

Assuming that a common data framework is needed, there are other important questions that must be asked: can the security of information (particularly information that is commercially sensitive) be ensured? Who would be using the information and for what purposes? Should the end-users of commonly available data be made to contribute towards the cost of gathering and maintaining that data? Does the data need to be archived for future use, or can it be assigned a shelf life? If data needs to be retained, how should it be stored and at what cost? Then there are questions of traceability, consistency, validity and liability: can the quality and provenance of the data be guaranteed? Is the information in the system internally consistent and can this even be tested in a general case? Is the information being refreshed at a rate that is appropriate for the task proposed? If information taken from the system is incorrect and that information is then used as the basis for a business decision, who picks up the bill for the mistake - the data owner, the individual or company allowing it to be accessed via the common framework, or the end user?

Despite these issues, there are groups attempting to either share data or create data models for the rail sector, both in the UK and overseas. In terms of data sharing initiatives, the most significant development in the UK to date must surely be Transport for London's Developer Area, a site providing access to around 20 different information feeds with content such as the location of stops, timetables, passenger counts and live traffic camera data; such was the demand for this data when the service first launched, that within weeks it had be temporarily suspended after the high volume of requests caused servers to crash. Other good examples of shared datasets include the DfT's National Public Transport Access Node (NaPTAN) data, listing transport stops across the country, and the National Public Transport Data Repository (NPTDR), which lists all the journeys made by public transport for a week in October each year. Both of these can be accessed via data.gov.uk. Amongst the railway data models, the XML-based RailML has seen growing support in Europe, although has not yet found wide support in the UK. In terms of more general transportation data models for cross-modal concepts such as timetables, NetEx and JourneyWeb are both good examples. Models also exist for condition monitoring data, with one commonly used example being the MIMOSA association's Open Systems Architecture for Condition-base Maintenance (OSA-CBM).

Another set of data models that are beginning to attract interest within the industry worldwide are conceptual data models. In particular, the FP6 InteGRail project in the EU has suggested that conceptual models may be the way forward for the wider railway industry as it starts to move towards offering ever more cross-border services, but what is a conceptual data model and what advantages does it have over traditional data modelling techniques?

Conceptual Data Models and the Preservation of Semantics

Conceptual data models differ from more traditional data model types as they allow not only the transfer of data but also a machine-interpretable set of metadata describing where the data came from and, more importantly, what the data means (its semantic). By placing data in an unambiguous context, conceptual data models allow computers to reason with information, generating new facts from those they have been provided with; put simply, conceptual data models have the potential to make computer systems into repositories of inferred knowledge rather than simply collections of stored data as they are today.

In practice, two things are required to develop a conceptual data model; a set of controlled vocabularies of terms used within the domain, and a set of related classes and rules that can be used to describe the domain from a particular viewpoints. When using a conceptual data model, developers state the relationship between an item of data and the model of the world, allowing that data to be seen in context by the computer. As an illustration, consider the US English and British English use of the word jelly. Jelly is a perfectly acceptable term in both US and British English, it even has the same syntax (can be used in the same places in sentences), but the meaning of the term, its semantic, is different in the two languages; in British English jelly is a gelatine-based dessert, whereas in US English jelly is a fruit preserve (a jam in British English). In a conventional data model such as a standard XML document, the tag “jelly” can be used ambiguously, because XML schema only enforce the positioning of the tag in the document and the values it can take, not its meaning in that context. Participants in a solely XML-based data exchange could legitimately use the tag “jelly” for either meaning, even if the designers of the XML schema had a particular usage in mind. By representing the data in a conceptual model this situation can be avoided, because the term “jelly” is defined as being a type of dessert that is composed of gelatine, and which may have a particular colour, flavour, shape and wobbliness.

As a key component of the forthcoming Semantic Web, conceptual data models have attracted a great deal of interest within the software industry in recent years and many large companies such as Amazon use them to help provide search functions and present adverts for related products to their customers. The web’s governing body, the World Wide Web Consortium (W3C), has produced recommendations for the implementation of one form of conceptual data model, ontology, using a new XML-compliant representation language called the Web Ontology Language (OWL). It is ontology that this document will propose as the basis of a common data model for the railway industry.

Methods

Standard Approaches for Ontology Development

While in an ideal world the way in which an ontology represented a domain would be unimportant as long as the model was conceptually correct, the practicalities of algorithmic design mean that the structure of an ontology and the types of information it captures about a domain can have a significant impact on its usefulness for particular tasks. This issue, known as task dependence, is currently one of the major challenges to the field, and is probably one of the best reasons for adopting a structured, engineering approach to ontology development.

The use of formal ontology engineering methodologies can help address the problem of task dependence by guiding engineers through the process of creating an ontology. In particular, a formal methodology can be useful in the early stages of the development process, ensuring that appropriate consideration is given to the requirements of the project, the available sources of domain knowledge, the possibilities for the reuse of existing work, the correct ways of managing the knowledge being gathered, and the appropriate levels of documentation. In the last 15 years a number of different ontology engineering methodologies have been proposed, often as the result of a successful modelling project. These can be split into three broad categories: traditional "monolithic" methodologies such as TOVE (Grűninger & Fox, 1995), Enterprise (Uschold, King, Moralee, & Zorgios, 1998), METHONTOLOGY (Fernández, Gómez-Pérez, & Juristo, 1997) and On-To-Knowledge (Sure, Staab, & Studer, 2009), methodologies designed for argumentation-based modelling by teams with different views of the domain or in diverse geographical locations such as DILIGENT (Pinto, Staab, Sure, & Tempich, 2004; Pinto, Staab, & Tempich, 2004; Tempich, Studer, Simperl, Luczak, & Pinto, 2007), and ontology learning methodologies, including Text-To-Onto (Maedche & Staab, 2000a, 2000b) and Rapid Ontology Development (ROD) (Zhou, 2007).

Monolithic Ontology Engineering Methodologies

In this report we will refer to the first category of ontology engineering methodologies as "monolithic" methodologies, since they are probably best suited to use by ontology developers working either alone, or as part of a team in a single location. The group of monolithic ontology engineering methodologies includes many of the oldest methods (e.g. TOVE and Enterprise) but is also the most popular, with a survey of over 600 Semantic Web practitioners and researchers by Cardoso (2007) finding that 13.9% had used On-To-Knowledge, and 7.4% had used METHONTOLOGY.

The Toronto Virtual Enterprise (TOVE) Enterprise Modelling project (Grűninger & Fox, 1995) was set up to create the first in a new generation of business models, one that had "the ability to deduce answers to queries that require relatively shallow knowledge of the domain". The TOVE ontology engineering methodology was much less focused on the domain knowledge acquisition and modelling tasks than those that followed it, and instead concentrated on a method for checking the adequacy of the final ontology. This was achieved by the formal specification of the ontology as a series of competency questions that could be expressed in First Order Logic and then used to check the completeness of the final ontology. The use of formal specifications set TOVE apart from other contemporary methodologies, such as that used in the construction of the Enterprise ontology (Uschold *et al*., 1998), which relied on natural language specifications.

METHONTOLOGY (Fernández *et al*., 1997) was developed after the TOVE and Enterprise methodologies, and was one of the first approaches to move ontology away from its rather philosophical roots towards becoming a software engineering discipline, making clear links to standard software engineering ideas such as lifecycles. It was also the first true engineering approach to ontology, defining not only a set of standard activities to be carried out during an ontology development exercise (namely specification, knowledge acquisition, conceptualisation, integration, implementation and evaluation), but also the documentation that should be produced to support each phase of the process and facilitate reuse by other ontology development activities in the future. Despite its age, METHONTOLOGY is still a widely used approach to ontology development.

Recently shown to be the most popular choice of ontology engineering methodology (Cardoso, 2007), On-To-Knowledge was developed as part of a larger EU-funded project that drew a distinction between the two aspects of a knowledge management process; the development, introduction and maintenance of a new ontology into a knowledge management application, known as the "Knowledge Meta Process", and the use of that ontology in knowledge management activities, the "Knowledge Process" (Sure, Staab, & Studer, 2009). The On-To-Knowledge methodology draws on the results of other European research projects, using formal methods to perform steps that many other methodologies treat in a less rigorous way, such as the use of CommonKADS during the project feasibility phase. It also has a strong focus on use-cases, and encourages ontology designers to think about not only how the ontology might be treated by an application, but also the needs of both users who are maintaining a knowledge base that uses the ontology, and of users mining that knowledge base for information.

Argumentation-based Ontology Engineering Methodologies

Whether it is as part of a political system, an ongoing debate between research groups, or a discussion over who gets to use the remote control, as humans we love to argue. Through argument we identify common ground and devise solutions that are acceptable to more than one party. The building of consensus between parties in ontology engineering is an important concept, as an ontology that has been developed from a number of different viewpoints is much more likely to be suitable for reuse in a range of different applications than an ontology developed in isolation.

The DILIGENT ontology engineering methodology (Pinto, Staab, Sure, & Tempich, 2004; Pinto, Staab, & Tempich, 2004; Tempich *et al*., 2007) is based on the principle that in virtual organisations there are several domain experts, with different and complementary skills, who may belong to competing organisations or be geographically dispersed; in these situations, the traditional, monolithic ontology engineering methodologies will be less effective. DILIGENT tries to address this issue by adopting a more community-driven approach to ontology development. Initially, a core ontology containing the main concepts for the domain is created to serve as a basis for discussions. The core ontology, which is maintained by a central board, is then distributed to all participants in the development process, who can apply their own local modifications and provide reasons for any changes that they have made. The changes made by each participant, along with their reasons, are made available to the central board for review, which may choose to add them to the next release of the core ontology. Participants are informed of new releases of the core ontology, and can choose to accept or reject the changes, with that information being passed to the central board to guide their future decisions; this cycle is then repeated for as long as modifications to the ontology are deemed necessary.

Argumentation-based approaches such as DILIGENT are particularly important in standardisation projects; here, different stakeholders may have very different views on how a system should be modelled, and negotiation between those stakeholders, along with more neutral viewpoints offered by parties with fewer interests in the specific area of contention, and the mediation provided by the central board, may be vital to reaching a consensus.

Ontology Learning

Even with a good methodology, the construction of a moderately sized ontology is a time-consuming task. One of the main bottlenecks in the ontology engineering process lies in the acquisition of domain-specific terminology and relationships from which the conceptual model is constructed. Ontology learning, which seeks to discover ontological knowledge from various forms of data automatically or semi-automatically (Zhou, 2007), is starting to emerge as a sub-area of ontology engineering due to the rapid increase in numbers of web-accessible documents and the advanced techniques shared by the information retrieval, machine learning, natural language processing and artificial intelligence communities (Ding & Foo, 2002a).

Technically more of a tool than a design methodology, but very much an early example of ontology learning, Text-To-Onto (Maedche & Staab, 2000a, 2000b) was a semi-automatic system for the learning of ontologies from domain texts. As one of the first systems to attempt to perform the extraction of domain terminology and concepts, and then structure that information, Text-To-Onto did not extract whole ontologies from domain sources; rather, it guided an ontology engineer in the selection of concepts and relationships. The Rapid Ontology Development (ROD) model (Zhou, 2007) by comparison, is a complete ontology design methodology based on ontology learning techniques, and is aimed towards complete autonomy in the ontology learning process as machine learning approaches mature. ROD consists of three main phases: design, learning and validation. The design phase differs very little from that of a more traditional ontology engineering approach, and requires input both from ontology engineers and domain experts. It includes an analysis of the domain, the establishment of the requirements for the ontology, the identification of resources that can be used as sources of domain knowledge, and possibly the development of a very general, top-level ontology to guide later work. The learning phase, which will ideally be autonomous in the future, can itself be subdivided into information extraction, ontology discovery and ontology organisation, and each of these in turn relies on a range of different techniques to extract information from the domain sources and subsequently build an ontology. Depending on the domain, and on the knowledge sources available, these might include the analysis hierarchies taken from web site navigation pages, thesauri, glossaries and indexes to establish key terms, natural language processing of textual documents, and the use of cluster analysis to identify the relationships between terms. The results of the ontology learning exercise are checked during the validation phase, with particular regards to redundancy, conflicts and missing information.

As a word of caution, it should be noted here that ontology learning is still very much in its infancy. Until machine learning approaches have matured further, or ontology editors come with out-of-the-box support for ontology learning, methodologies such as ROD are really more of a subject for academic investigation than serious ontology development work, at least by anyone other than experts in machine learning.

Commonalities between Ontology Engineering Methodologies

There are of course many other ontology engineering methodologies in use than those described here, and while the general type of methodology (monolithic, argumentation-based, or ontology learning) chosen for an ontology project may be important, the specific methodology that an engineer chooses is almost certainly of far less importance than the choice to follow a formal, well-defined process. Indeed, in many cases engineers may select a methodology simply because it fits in well with their own/current organisational working practices. Large numbers of ontology developers also follow other methodologies, including methodologies of their own devising (over 13% in Cardoso's (2007) survey), and so it seems appropriate at this point to mention that work by Pinto & Martins (2004) has shown that many of the formal ontology engineering methodologies share common stages that may be a good starting point for engineers who wish to develop their own approach. These are:

- Specification: Establishing the requirements and scope of the ontology.
- Conceptualisation: The creation of a conceptual model of the domain.
- Formalisation: The creation of a formal model of the concepts identified in the previous step through the assignment of relationships etc.
- Implementation: The coding of the ontology in a representation language such as OWL.
- Maintenance: The process of updating and correcting the ontology throughout its lifetime.

In addition to these, Pinto & Martins (2004) also defined tasks that should take place throughout the lifecycle of the ontology:

- Knowledge acquisition: The gathering of appropriate domain knowledge.
- Evaluation: Judging the quality of the ontology.
- Documentation: Recording what was done, how it was done, and why it was done.

Reusing Existing Ontologies

As with object oriented programming, the ability to reuse model components is a very important part of ontology engineering. Indeed, many ontology engineering methodologies include it as formal stage that takes place after the identification of concepts and relationships, just prior to the implementation of the ontology. The reuse of ontologies can take place in two main ways, either by the direct combination of material from one ontology into another, or through the identification of patterns in the structure of existing ontologies and the subsequent use of those patterns as templates in a new ontology.

Ontology Reuse by Fusion and Composition

The ability to directly reuse parts of an ontology is one of the major advantages of ontologies over XML schema, since the semantic information contained in an ontology model makes it much more likely that a new ontology, formed by the combination of elements from existing ontologies, will retain the original meaning. Two main sub types of the direct reuse of ontologies exist, reuse by fusion and reuse by composition (Pinto & Martins, 2004).

Ontology reuse by fusion is essentially a merging of two or more ontologies; the relevant concepts and relationships are identified in each of the source ontologies, along with the criteria used to classify them. Areas of consensus are retained and the remaining information fitted into the model as appropriate, forming an entirely new ontology. Ontology reuse by composition on the other hand, describes the identification and combination of discrete modules from the source ontologies in such a way that they are still recognisable; for example, a complete description of a vehicle taken from one ontology, and the description of a track section from another, being combined into a new ontology describing a railway. Generally speaking, ontology reuse by fusion is the preferred method when the source ontologies contain a great deal of overlapping knowledge, whereas reuse by composition is more common where there is little overlap.

One very important example of reuse by composition exists in the form of upper level ontologies. Upper level ontologies model general concepts that will not alter across domains, aiming to provide a common base that domain-specific ontologies can extend. While there is still a great deal of debate within the ontology community over whether task dependency makes a true upper level ontology an impossibility, several attempts have been made. These include Cyc (Matuszek, Cabral, Witbrock, & DeOliveira, 2006), SUMO (Niles & Pease, 2001; Pease, Niles, & Li, 2002), COSMO (Cassidy, 2008) and the OWL implementation of ISO 15926 (Batres *et al*., 2007).

Ontology Design Patterns

In software engineering, design patterns are proven, conceptual solutions to common design problems. It is important to note that software design patterns are not reusable blocks of code; if they were then the details of their implementation, such as choice of programming language, might conflict with the requirements of the system being developed. Instead, a software design pattern is a description of a problem, a common solution to that problem that has been successfully used in the past, details of where the solution can be applied, and a discussion of the consequences and trade-offs that result from its use (Sommerville, 2007). Software design patterns are not invented; rather, they are discovered from experience in building practical systems (Devedzic, 1999) and exist on a range of different scales, from algorithmic skeletons, an example of the template method design pattern (Gamma, Helm, Johnson, & Vlissides, 1995), to architectural patterns, such as the model-view-controller (MVC) pattern for software with a graphical user interface.

By contrast to software design patterns, a software antipattern is a commonly occurring solution to a problem that generates decidedly negative consequences (Brown, Malveau, III, & Mowbray, 1998). Examples of software antipatterns include the golden hammer, which describes the inappropriate use of a particular technique because the programming team is familiar with it, cut-and-paste programming, where several similar instances of a piece of code occur throughout a program due to the use of cut-and-paste, lava flow and spaghetti code.

A number of authors have noted that software design patterns and ontologies have much in common (Devedzic, 1999; Blomqvist & Sandkuhl, 2005); on one hand, software design patterns have sets of standard features and properties that can be represented by an ontology, on the other, the core ideas of software design patterns, those being the reuse of components and that particular general solutions can be of use in a range of applications, can also be applied to the design of ontologies. Both techniques also involved knowledge transfer and lead to the creation of common vocabularies, although the vocabularies produced for use with ontologies are far more formal than those used in software design patterns (Devedzic, 1999). Given these similarities, it is hardly surprising that ontologies have been used to formally describe software design patterns (Dietrich & Elgar, 2005), or that design patterns have begun to be used in ontology engineering. Examples include “knowledge patterns” (Clark, Thompson, & Porter, 2000), “semantic patterns” (Staab, Erdmann, & Maedche, 2001), and the work by Gangemi (2005) on the practical use of conceptual ontology design patterns (CODePs).

One of the most interesting possibilities raised by ontology design patterns lies in their potential uses in ontology learning (Blomqvist, 2005), where concepts and relationships that have been extracted by linguistic analysis could then be matched to libraries of known ontology patterns to assist in model development.

Ontology Engineering Methods Adopted for use in the Project

Although the huge scope of the railway domain suggested that a collaborative approach, such as DILIGENT, where groups of domain experts directly contribute to the areas of the data model relevant to their expertise would be the most appropriate methodology for the project, the very short timeframe available, small team of developers and lack of the necessary collaborative IT framework for the activity led to the METHONTOLOGY approach being adopted for use during the prototype model development work within the project. METHONTOLOGY is a six-step methodology comprised of specification, knowledge acquisition, conceptualisation, integration, implementation and evaluation tasks; alongside each of these is the ongoing process of documentation, with a particular document being produced as a required output of each step (requirements specification, knowledge acquisition document, etc.). A detailed description on the use of METHONTOLOGY within the project can be seen in the prototype model development section later in this document.

Standard Approaches to Stakeholder Knowledge and Requirements Elicitation

Many of the ontology development methods described earlier involve a stage where domain knowledge is elicited with the aid of domain experts (the 'knowledge acquisition' phase (Pinto and Martin, 2004)), and / or experts are used in the validation process (the 'evaluation' phase). Methods of knowledge elicitation and knowledge acquisition have emerged since the late 1980s; many are based on the use of prior psychological accounts to understand the nature of expert knowledge and expert reasoning (e.g. Chi, Feltovich and Glaser, 1981), and are now a recognized component of ontology design.

Much of this work in expertise has highlighted the distinction between explicit and tacit knowledge. Explicit knowledge is easily accessed either through documentation or through discussion with experts. Tacit knowledge, however, is more challenging; this knowledge is hidden from immediate access, often because it is personal and has value to the expert (it is what comprises their expertise). Also, tacit knowledge is heavily embedded in the processes that experts engage in, and often only comes to light when that process is explored or simulated. Therefore, a range of knowledge elicitation techniques have been developed that are often used in combination to facilitate the most rapid, accurate and complete access to expert knowledge.

Methods of Elicitation

Broadly, methods of knowledge elicitation are split into two categories – natural and artificial (or contrived) techniques. Natural techniques are those where the conditions in which the knowledge is elicited are as close as possible to those that the expert will be familiar with. They comprise interviews (either as individuals or sometimes in the form of focus groups), observation and ethnography. The advantage of these methods is that they are often quick to administer (though ethnography can be very time consuming) and allow the expert knowledge to emerge in a natural manner. Some tacit knowledge, and particularly expert behaviours, may only emerge serendipitously, such as behaviours in control room environments. The difficulty, particularly when there is very little structure to the inquiry, is that expertise is often covered in a non-systematic manner and it is difficult for the analyst to understand the completeness of the data. Also, the analysis process can be very time consuming and open to subjectivity on the part of the analyst.

An alternative set of techniques, contrived techniques, use structures or probes specifically designed to elicit expert knowledge, particularly tacit knowledge and relationships in the domain that may not be apparent even to the expert. Concept mapping and laddering are two related techniques where the expert is asked, with the assistance of the analyst, to develop visualizations (akin to mind maps) that represent the domain in question. Also, while a task analysis – sequential representations of tasks and processes – can be the outcome of an analysis, it can be useful to develop these in tandem with the expert. Critical Decision Method uses scenarios to help elicit key decision points and criteria for decision making used in processes, particularly those where decision-making is rapid and naturalistic, and therefore knowledge is again tacit. Another form of contrived technique is repertory grids; based on techniques derived from personality psychology, 'rep grids' present the expert with groupings of domain elements and ask them to quantify or rank items across varying criteria. This allows certain characteristics of the domain to emerge, e.g. key material characteristics for distinguishing between minerals. One final technique is card sorting, where experts are asked to develop groups of cards corresponding to groups of previously identified concepts, and then naming and describing the different characteristics of those groupings. The advantage of such contrived techniques is that they offer fast and targeted access to tacit knowledge. A secondary benefit is that some techniques, concept mapping especially, generate outputs that can be turned into ontology-like representations with little additional effort. The disadvantage of such techniques is that they may involve some explanation, and require skill and practice on the part of the analyst.

In practice, methods of knowledge elicitation are often combined. For example, while focus groups may be difficult to manage without clear direction, collaboration towards defining a concept map or task analysis may help give the structure to focus group activity. Also, methods can be combined over time – initial exploration may involve more open-ended methods to elicit key concepts or decision points. Contrived methods are then used to highlight the key characteristics of these concepts.

Requirements Analysis

In parallel with the need to capture expert knowledge for ontology development, any technology, ontology included, needs a concrete definition of the requirements of the product and the intended context of use. This is essential to ensure the finished product meets a clear and necessary aim. Several methods of requirements elicitation, such as interviews, focus groups and task analysis, overlap with those of knowledge elicitation. Also, theoretically, at least some of the stakeholders who contribute to requirements can be the same people who act as experts for knowledge elicitation.

Within the feasibility study, the primary areas of interest lay in the perceived uses of, value in, and barriers to, the adoption of ontology in the rail sector. In other words, the study would focus on the methods for the establishment of the business case behind the use of ontology in the rail domain, as much as on their specification and subsequent content. This required expert knowledge, and opinion, from stakeholders. However, it also indicated that more exploratory techniques, such as 'brainstorming', might be of value. These activities often benefit from very open-ended approaches, or approaches where there is much scope for expressing and generating ideas, before any firm decision is made on selecting those ideas.

Knowledge and Requirements Elicitation Approaches Adopted for use in the Project

The methods used in this project therefore were selected to meet a number of aims:

- To elicit knowledge from experts
- To elicit business requirements, as well as some of the likely barriers and motivators to implementation, from rail stakeholders
- To see if both could be carried out concurrently to maximize use of stakeholder time

To this end, three stages have been used (see Figure 1).

- A stakeholder workshop – initially participants were asked to generate cards relating to potential uses of an ontology. They were asked to do this in groups, with as little filtering or critiquing of ideas as possible. They were then asked to sort these cards according to themes and, using one case scenario, generate a number of factors relating to that scenario. As such, this phase used a focus group approach, combining card sort and semi-structured interview questions to structure the session.
- A knowledge elicitation workshop – this session aimed to explore one scenario (franchise handover) in depth. This was based around a task analysis / elicitation, with participants asked to highlight the key steps in the franchise handover process, and discuss barriers relevant to each of those steps. A second activity involved a structured survey (essentially, a paper-based structured interview), where the participant was asked to give their input regarding themes highlighted at the knowledge elicitation workshop.
- Validation – this activity involved presenting concept maps based on the ontology developed out of all of the prior activities, and presenting this back to key stakeholders (yet to take place).

More details about each of these sections are presented in the relevant findings sections.

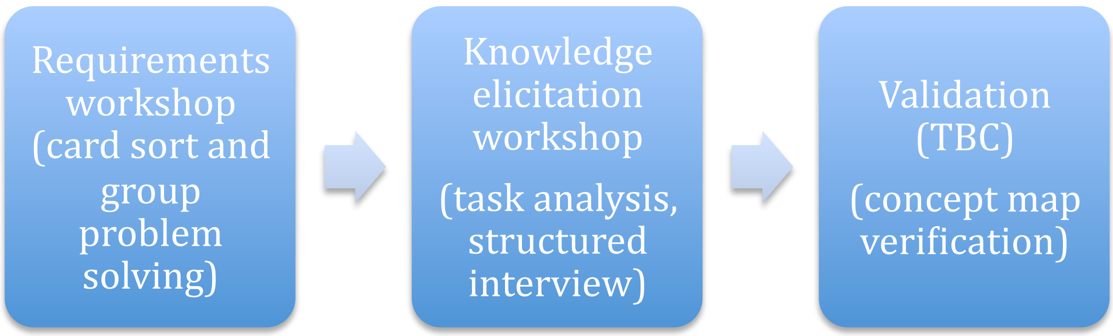

Figure 1: Stages of the knowledge and requirements elicitation process used in the study.

Findings of the Project Workshops

Requirements Workshop

As part of the feasibility study, an industry stakeholder workshop was held on the 23rd March 2011 at the Railway Safety and Standards Board (RSSB) offices in London. The workshop brought together representatives from across the rail sector to capture some of the major scenarios where a common data framework would be of value, some of the components of the framework that would be required to support those scenarios, and of the existing data sources that could be brought to bear.

In total 25 representatives of 15 industrial stakeholders (plus the project team from the Universities of Birmingham and Nottingham) took part in the workshop, which began with a series of presentations providing the technical background to the project alongside a presentation from Jeff Brewer, RSSB, on the range of IT systems in use within the industry. The participants were split across 5 tables in such a way that there was a cross-section of industrial expertise on each table (see Table 1). The participants were then asked on an individual basis to list as many scenarios as possible where they thought there may be some value to be gained from a common data framework for the industry. Across the five tables over 150 scenarios were identified (see "Appendix A – Suggested Applications for a Railway Data Framework" for the complete list). The scenarios were then subdivided into categories made up of Network Rail's "4 Cs" (cost, customer, carbon, capacity) and an additional group "commission" initially intended to represent interactions with Europe, either through regulation or cross-border services, but rapidly adapted during the meeting to mean any co-operative action.

Group	Stakeholders
Table 1	Porterbrook, Network Rail, Thales, Alstom, ATOC, University of Birmingham
Table 2	Information Junction, Network Rail, Virgin Trains, Invensys Rail, University of Birmingham
Table 3	Information Junction, Network Rail, RSSB, Atkins, First Group, University of Nottingham
Table 4	Network Rail, Data and Process Advantage Limited, TATA Steel, Trapeze, DB Schenker Rail, University of Nottingham
Table 5	Network Rail, Alstom, TATA Steel, University of Nottingham

Table 1: Industrial stakeholders represented on each of the workshop tables.

Analysis of Generated Scenarios

Of the 153 different scenarios identified, the majority were assigned to either the cost or commission groups (44 scenarios each). In addition to these, 28 were assigned to customer, 23 to capacity, and 6 to carbon. A further 6 were assigned as "miscellaneous", usually because participants had placed them in the centre of the classification diagram suggesting that they were relevant to all of the categories (note that there was no order suggested by the structure of the classification, or the layout of the classification diagram). Within these categories there were many recurrent and similar examples. A sample of these is presented in Table 2.

‘C’ (n responses)	Examples
Cost (44)	Maintenance planning, Maintenance life cycles, Understand real asset usage for renewal planning, Extending vehicle life, Data sharing to optimise performance (capacity availability), Data for whole system models for maintenance and renewal decision tools, Increase productivity by enabling easy discovery of data
Commission (44)	Standard time table between bus and train and EU, Knowledge management, Interoperability reference, One industry geographical data model, Reference for stakeholders, Inter organisation incident management e.g. Olympic coordination, Freight logistical flows + pan-European tracking of goods
Customer (28)	Delay causation, End to end tickets for cross modal journeys, Consistency of information to TOCs / pubic w.r.t. service disruption due, Signalling infrastructure problems -> need for re-scheduling, re-routing etc., How we mange delays, performance management and improvement – including KPI, customer information, seat reservation free seat location available – bookable from platform, train movements - accurate movement / capacity of trains
Capacity (23)	Train consist information (which vehicle numbers in which order), Sharing of train service characteristics between TOCs / FOCs and operations control (formulations, coach class etc.), Journeys taken -> future planning, How good are our assets, Where is my rolling stock, How to optimise track and rolling stock, Adhesion information to help with SPADs, Wheel flats, Delays, Problems (lowering speed across the network due to small LA)
Carbon (6)	Real time optimal driving control (adaptive on current state), Information display based on end user roles, Driving optimisation i.e. train regulation and advisory speeds to avoid wasted energy, Energy usage – how to improve it, How to pay for it, Train class, Sharing of passenger congestion information to support passenger information and behavioural change applications
Misc (6)	What is our inventory of assets, Better records of asset status saves effort in need for asset survey, Asset records capture essentials of functions, Interfaces – key for considering impact of change

Table 2: Analysis of potential applications by 5 Cs

Categorisation	Numbers
Asset management	40
Operations	30
Business and strategy	26
Interoperability	15
Customer	12
Data	12
Vehicle monitoring and management	8
Standardisation	4
Testing / modelling	4

Table 3: Alternative categorisation for potential applications

Looking through the potential applications, there was some duplication, and many recurrent themes cutting across the analysis by 5 Cs. In order to interpret this, a second analysis was carried out which categorised applications by some of these recurrent themes. This analysis is presented as Table 3 (NB in many cases the notes provided by participants were somewhat ambiguous and reflected the viewpoint of that particular participant. For example, 'asset management' could mean track, other infrastructure, vehicle etc. While this means only a surface level analysis is possible, the main aim of this analysis is to show the breadth and general coverage and, to this end, the analysis presented in Table 3 is sufficient.).

Analysis of Specific Scenarios - Barriers and Benefits

Following on from the first phase of the workshop, each table was invited to select one of their scenarios for more detailed development; identifying potential sources of any data needed for the scenario and listing the benefits of, barriers to, and stakeholders involved in the process (see "Appendix B – In-workshop Elaborations of Key Scenarios" for the details of this exercise).

The scenarios selected by each table covered, along with benefits and barriers, are presented in Table 4.

Analysis of Specific Scenarios - Stakeholders

In addition, participants were asked to identify major stakeholders for each of the scenarios. Many of these stakeholders were identified across scenarios. The analysis of the applications generated in phase one highlighted a number of relevant stakeholders, several of which were referred to at the scenario stage. The lists of identified stakeholders were merged to give the following:

- Infrastructure operators (e.g. Network Rail)
- Government
    - EU
    - National (DfT)
    - Regional
- Regulators and standard bodies (e.g. RSSB)
- Train operating companies (TOCs), Freight operating companies (FOCs)
- Customers
    - Passengers
    - Freight
- ROSCOs
- Suppliers
    - Original Equipment Manufacturers (OEMs)
    - Materials suppliers
    - Component
    - Energy suppliers
    - Software suppliers
- Vehicle maintainers
- Emergency services
- Ordnance Survey
- Weather (MET office)
- Other transport mode operators
    - Cross-modal operators (e.g. TfL)

Scenario	Benefits	Barriers
Understanding vehicle status and history for franchise transfer	Improved ppm Safety Warranty Tracking fitment and notifications to reduce potential for dip in performance as stock is transferred	IT competence Poor documentation Lack of open / standard interface High level collaboration (directors) Legacy Incentives for sharing Costs benefits Who owns what?
Track to train condition monitoring	More reactive maintenance Reduced delays minutes costs Reduced failures better customer experience Improved safety Increase asset life and availability	Need to agree on the data type / flow transmitted Business case - why should NR spend money to help TOCs Commercial arrangements cost / benefits in the right organisation
Operations planning (long term and real time)	Optimise performance Utilise network effectively Asset productivity Customer experience (delay) Operational costs	Structure of the industry Govt direction Inflexibility Contracts No feedback cycle
Increasing route utilisation (increasing usage of existing track; optimising track to support increased usage)	Increased utilisation More money and passengers Better passenger info More predictable and reliable	Data standards (lack of) Fragmentation Supplier led Uncertain future Complexity
Supply chain data.	Better equipment for purpose Better knowledge of assets Better product design	Data not available or in enough detail Confidentiality A “shared vision” Volumes and sources of data Getting data to the right place at the right time Interpreting a lot of data

Table 4: Detailed scenario analysis - benefits and barriers

Franchise Handover Workshop

Due to depth of information provided at the initial workshop, and the support of a relevant organisation (Porterbrook) we elected, as a second knowledge elicitation activity, to explore franchise handover in further detail.

A day workshop was arranged at Porterbrook, Derby, attended by relevant members of the organisation with an operational interest in how a data framework might be used to help support their business. The five attendees included the Operations Director, Engineering Applications Manager, Engineering Development Manager, Director of IT and Commercial Manager.

The workshop was recorded and transcribed. This transcription supported a number of different major areas of analysis. These were examined using NVivo - a content analysis tool - to gather comments regarding themes such as current gaps in activities, uses of a data framework, and barriers to adoption. While the intention was to use the workshop to discuss the specifics of the franchise handover scenario, a certain amount of early discussion aimed to elicit the more general business aims of a ROSCO and to introduce the data framework concept. In the end, this phase of discussion generated much useful data about the context in which a rail data framework could be used, along with the challenges to, and specific applications of a framework.

The Business Context

The first section of the meeting focused on understanding the **general characteristics and aims of the organisation** in order to ground how a data framework might be used and add value, and the business aims it would need to support. These general business characteristics included:

- A core business emphasis on managing and maintaining vehicle as the asset, while providing value to TOCs
- Emphasis on two major functions key for the ROSCO - operations and asset management
- The need to receiving a vehicle back at the end of the lease in a manner that is commensurate with the maintenance plan (and negotiation around this if this is not the case)
- The central importance of the maintenance plan - critical to strategic planning, costing and understanding the life expectancy of assets
- The importance of the maintenance plan meant that historical views were of limited importance. Once maintenance had been performed and new components fitted, the history of the component or vehicle was effectively 'overwritten' and the lifespan of the asset could be projected forward (and extended) through the maintenance plan
- Different levels of lease agreement requiring different levels of involvement from the ROSCO - sometimes involvement is relatively infrequent, sometimes continuous
- Lease agreement size is also a factor in how the lease is managed, as it can influence how often or easily the ROSCO can review its assets (can be more difficult with larger fleets)

- Type of lessee (freight or passenger) also influences the type of commercial agreement - generally there is less visibility with freight leasing
- Assets themselves are not always under inspection - inspection is more concerned with how well the vehicle is being maintained with regards to the maintenance plan, rather than MOTing vehicles on a regular basis
- Much data does not have to be analysed in real-time

Later, discussions moved on to **specific issues** relating to limited data or lack of information, in order to understand the potential for data frameworks. These included:

- Generally, the industry is good at producing data but cannot use it because it cannot be interpreted efficiently
- One reason for this is increasing diversity in the systems being used by the various stakeholders in the railways. Often systems were being chosen on the grounds of procurement cost, rather than on factors such as data interoperability. As a result there was a view that getting industry-wide views, or consolidating views across different stakeholders, was now very difficult, and that extra resources were being used to interpret data
- The lack of clarity or emphasis on data exchange in commercial agreements did not help the cause of data exchange
- A repeated theme was that the knowledge of current status of the world (e.g. recurrent issues, or performance differences) was already there, but that the proof or detail was either not available, or that it was too difficult or time consuming to access, because of issues with system incompatibility
- Data quality issues, especially around mileage, which is the core of the heavy maintenance plan
- Issues with data entry, where minor slips in data, once entered become 'sacrosanct'
- Reverting to the lowest common denominator (usually excel) in order to access data from the wide variety of data sources and systems available
- Each TOC could have different configurations requirements or other varying characteristics such as different duty cycles - understanding this was not easy
- Accessing and assessing vehicles against route characteristics was an area for improvement – this would serve a number of uses including assessing route acceptance certification, calculating mileages, assessing gauge limitations for freight and assessing performance of assets
- Missing location data, also for TOCs, so they can have a real-time picture of where their assets are
- Issues around cataloguing and tracking parts in stores
- In terms of life extension and operations, currently have to treat vehicles the same. With better data on usage and performance you can treat vehicles individually
- Differences between freight and passenger vehicles - generally the passenger can be a source of data (about incidents, about performance). Don't have that with freight. Also, data generally on freight is less complete

In addition the discussion raised a number of **general issues** with operations. While such issues were not specifically related to data, they were indicative of business gaps where improved knowledge and information, based on new data or improved data analysis, could play a role. These issues included:

- Improving on configuration control for all stakeholders in train ownership / operation process (e.g. where gaps or errors in the configuration management process had led to old parts being refitted because of known, fixed issues resurfacing with newer components)
- A limited window of opportunity to assess asset status (vehicle) and a desire to be in a position to see that picture more dynamically
- Operational competence levels are not consistent - one TOC may be maintaining well in one respect but not in another.
- Managing stores and the stores float

A number of **specific uses** for a data framework were highlighted. Not surprisingly, several of these reflect some of the gaps in data management described above. These included:

- Being able to have a dynamic view of maintenance status of a vehicle
- Automating data entry to overcome data entry errors
- Assessing route clearance through using accurate route information
- Assessing gauges etc for freight
- Vehicle location - for understanding where a train is through GPS, during incidents (e.g. bad weather)
- Performance and energy monitoring to understand duty cycles and modify or propose operational changes accordingly
- Data entry for stores and component management
- Understanding repeat defects - minimise risk of persistent faults
- Abstracting value from condition monitoring reports
- Being able to see performance drop off in a class or fleet so it can be compared with other operators
- Being able to diagnose vehicle faults (i.e. those case where the fault was known, but the causes were yet to be understood)

The next theme highlighted the ways in which improved data usage could have **value** for a ROSCO, for other stakeholders, and for the industry generally:

- OEMs would prefer a standardisation because they are having to build for multiple interfaces
- A 'level' playing field i.e. common data and knowledge within the rail industry
- Value of being seen as a contributing stakeholder within the rail industry
- Having better knowledge and prediction to enable more effective life extension for a vehicle, ultimately leading to cost savings
- Ultimately, reduced costs for individual organisations, and rail as a whole, in keeping with McNulty
- The ability to prove with data what is already known - e.g. different performance characteristics of vehicles in different areas, value of decisions regarding life extension
- Better data is better responsiveness is better reputation
- While information might be known or available, better system integration would make this process faster and less resource intensive

- The rail industry as a whole being able to predict and understand performance everywhere across the network

Potential **barriers** to the adoption of a rail data framework (and some potential **solutions**) included:

- Commercial sensitivity and value - There were concerns that organisations would be reluctant to share, as it would reveal commercially sensitive and valuable information about their processes and products. This is a particular concern for franchise holders, as it is likely they do not own the rolling stock etc. that they operate, and therefore the only advantage they have is in the data they hold. One perspective stated during the discussions was that the only way OEMs really make their money is by maintaining trains at a premium, and divulging to others how to maintain these trains more effectively would undermine their business model. However, there was also a view that a certain amount of IPR was perceived in data when it didn't really exist, or given that the value was not understood, organisations would prefer to assume that it could have value and protect it. In the absence of knowing exactly what their data was worth or whether it really was sensitive, an organisation is likely to adopt the most conservative case and restrict access. The view was that, in terms of data exchange, the philosophy (and potentially the mandate from a body like RSSB or DfT) was that data should be exchanged unless there was a specific case for not doing so, rather than the opposite, which appears to be the current case. Even for that data that cannot be shared, there may be a form that can.
- Data ownership - One issue would be clarifying who exactly owned the data in certain circumstances. For example, in the case of a vehicle, would this be the TOC, the ROSCO or the OEM? The view (not surprisingly for this context) was that this would firmly be with the ROSCO as owners of the vehicle. However, the view was also that data exchange should still be as free as possible as it gives all stakeholders a level playing field, particularly in the area of franchise negotiation and tendering.
- Data volume - one potential barrier is that additional information and knowledge generated from data would highlight factors for better operations, for example finding repeated faults to fix. In practice, there was already too much information in the system to do the work that was already required, so it was questionable whether there would be capacity to act on this additional data.
- Cost - the cost of implementing the changes required for a data framework were perceived as being a barrier to the adoption of such an approach. The view was that if such changes were integrated at the change of a franchise, however, it would be a marginal issue as systems were often refreshed at that point anyway. Also, demonstrating the commercial value of data exchange could offset some of the perceived costs, and providing clear scenarios of how this could be achieved would be advantageous.
- IT Competence - One issue is that information systems are not perceived as the core of rail activity ("the level of IT sophistication does not match the level of engineering sophistication"), and that the industry did not have a good appreciation of data modelling. Real examples would again be of benefit. Also, there was a view that it was about getting data to the right people at the right

time. Whether people could understand that data and embrace it was another matter.

- Business case - One further barrier was the view that generally, the key data and information that was required to run the business was available, and that it was difficult to build the business case for unforeseen, marginal or infrequent events and activities. Also, industry fragmentation meant it was difficult to build an industry wide business case. Again, concrete examples of the value of improved data exchange would help address this barrier.

Finally, one theme of questioning and analysis covered the **mechanisms** for managing and regulating such a framework. One view was the RSSB should be the driving force behind cross-industry initiatives. The other view was the DfT would drive this initiative - partly because they ultimately were in charge of financing as well as the impetus to cut cost at an industry-wide level, and also because they had the ability to specify adherence to data frameworks as part of their procurement and franchising processes. Adhering to data exchange formats would be part of the 'rules of the game'. Such a body would also be able to drive an industry wide business case.

The Franchise Handover Scenario

The second half of the meeting focussed specifically on the elaboration of the franchise handover scenario identified in the March Stakeholder workshop.

The major stages of the franchise handover **process** included:

- Ongoing through franchise - annual review of 10% of vehicles
- 6 months before handover - inspection
- 3 weeks before handover - inspection
- Outputs include
    - Defect list
    - Commercial cost of dilapidation
    - Costs for any used, non-maintained or lost spares
    - Updates to the maintenance plan

Aims of the franchise process included:

- Checking the state of the vehicle is commensurate with the maintenance plan
- Proposing a plan of maintenance so that the vehicle would be in an appropriate state at handover, or placing a commercial value on incomplete maintenance
- Ensuring the status and value of used, unmaintained or lost spares
- Establishing what the follow on TOC is getting
- Ensuring the bidding and handover process is fair and thorough
- An opportunity to check the fleet - generally, and to ensure vehicle condition is commensurate with the data plan
- Ensuring the process has to be conducted in a structured manner
- Understanding the state of the asset so as to be able to plan any changes into the business

Current perceived **problems** within the franchise process included:

- Configuration control
- Mismatches between paper and electronic records (different views as to which was the correct representation)
- Lack of visibility as to how running maintenance flows from one franchisee to another
- Understanding how changes in duty cycles or other performance demands (including potentially different routes) might impact performance between one lessee and the next

Other relevant characteristics of the franchise process include:

- Most key information was already held electronically
    - Heavy maintenance records, deferred 1-4 maintenance activities
- A lot of information that would help the handover process was historically held elsewhere (RAVERS), or the safety management system
- Data format (and sometimes quality) was determined by systems in place at the time of handover
- The change between one lessee and the next was not as substantial as first might appear - often the operational staff remained unchanged between on franchise holder and the next

A number of **benefits** for improved data and information management (and from that, the case for a data framework) within the franchise process were highlighted. These included:

- Information would be available to potential bidders so that they would all have an equal understanding of what they were bidding for
- Better data exchange would allow the ROSCO better knowledge coming up to a franchise exchange of the state of their asset
- Better data exchange would give the ROSCO better knowledge of how the vehicle was being operated and predict and manage any differences based on potential changes in the new franchise environment
- To be able to offer the new franchise holder advanced information about vehicle performance in order for them to operate a better service - i.e. to add value for the franchise holder
- Much of this process was currently carried out on paper. While not necessarily a major problem, there were occasional mismatches between paper and electronic information. Also, while this was not necessarily a problem for the ROSCO, it could be for the follow on TOC and being able to improve this process could be a way of adding value
- Aiding data collection during the inspection process - facilitating 'quiet enjoyment'
- Checking concerns about configuration
- Being able to reconfigure a fleet to suit what the new operator envisages they will require and matching that to what is possible technically and financially. ("Knowing vehicles, knowing assets and suggesting ideas")

Key **actors** in the handover process included:

- Outgoing franchise holder
- Bidders
- Incoming franchise holder
- OEMs
- Suppliers
- Maintainers
- 3rd party spares managers
- 3rd party inspectors
- Network Rail
- RSSB
- DfT
- Technology suppliers

Discussion of Workshop Outcomes

A number of conclusions can be drawn from the stakeholder exercises. To begin with, there are a number of concrete applications for data exchange within the industry. Many of these are system-wide (e.g. improving operations at a national or EU-level) while others are very specific in nature (e.g. analysing data to determine the best sites for trials). The areas of application are spread across real-time operational issues, short term planning, strategic planning, and across all sectors of the rail industry. Also, it is worth noting that although all of the 5 Cs were covered, the majority of the scenarios were categorised under either cooperation or cost.

While we might expect to see this breadth of applications given the activities and attendees at the first workshop, this was also re-iterated in the second workshop where, even though the focus of the workshop was on franchise handover, there were many suggestions for applications for a number of functions and activities of the ROSCO, and for other rail industry partners.

The scenarios for data exchange captured during the workshops differed in the type of business process they supported. This came out most clearly in the second workshop. Not only were there clear links between the applications expressed, and a number of general business aims of the organisation, they also varied in terms of whether they facilitated bottom-up processing, or top-down processing. Put another way, while some of the uses were related to data discovery or data exchange in order to generate knowledge (e.g. better mileage and route data to help understand what would be required for route acceptance), there was another class of application where the knowledge was already partially known, but the analysis of data would help ground the business case, prove a certain operational fact (e.g. that the same class of vehicle really was performing differently under different TOCs), or diagnose a known issues (e.g. why were the vehicles performing differently under different TOCs). This kind of result is critical as the algorithms, and user interfaces, which will sit on top of a data framework, must be tailored to fit the kind of discovery and analysis processes the operator wishes to carry out. A "one-size fits all" approach is not likely to be sufficient.

There was a consistent pattern of barriers expressed by participants in both workshops – ownership of data and commercial sensitivity, IT competence, lack of standardisation, business case, and the need for clear leadership (and even mandate) from government or regulatory bodies was expressed in both sessions. The feeling was that either RSSB or DfT would be required to push such an initiative through, and while the business benefits were clear to those involved in our workshops, the need to demonstrate clear cases and value was essential for the wider acceptance of any data framework approach.

As a comment on method, the initial workshop played an important role in generating the breadth of applications, and therefore demonstrating the scope of an industry wide framework. Also, an initial analysis of these applications and the detail of the scenario highlighted obvious concepts for a data framework (assets, actors, processes etc), but also some less obvious or more abstract constructs for the data framework that need to be considered. For example, the role of time, and the distinction between real-time operational uses, tactical uses, and strategic planning is important in the exchange of data. For example, data that is currently being used for real-time purposes (e.g. point usage) could be used in a modified, or analysed form, to feed into more long-term analysis (e.g. trends to assess the need for maintenance and renewal). Again, this requires both appropriate analysis and presentation of the data so it can be used effectively. Information at the granularity to support real-time operations may be to detailed to support effective strategic decision-making. Similar issues of granularity are likely to be found in terms of geography (where data captured in terms of meters or even millimetres, might need to be expressed at a regional or even national level). Such concepts will need to be modelled sufficiently to enable flexible data exchange in an industry wide framework.

The second workshop provided invaluable information and followed the expected flow of generating further detail to flesh out information raised in the first workshop. What was so surprising was the amount of information regarding business cases, and business context, in order to ground the discussion on the content of the date framework itself. This serves to highlight the importance of a full understanding of business requirements for any application of a data framework.

Prototype Domain Model Development

Following on from the workshop sessions, the prototype model development portion of the feasibility study aimed to produce a set of simple ontology models that could be used to illustrate how a common, semantic data model could improve the efficiency of data exchange in the railway industry. As mentioned previously, the METHONTOLOGY approach to ontology development was selected for use within the project; this section of the report, which is laid out according to the six steps of the METHONTOLOGY process, will describe our initial attempt to apply METHONTOLOGY in the railway domain.

Specification

A key step in any software development process is to establish what the question to be answered is; under METHONTOLOGY this is formally required at the beginning of the specification process.

In the case of the prototype model development, the questions that the ontologies would be designed to answer / use-cases that they would be designed to support were selected based on the outputs of the brainstorming session of the stakeholder workshop. The initial set of 153 use-cases generated by the group was filtered down to leave only those ideas that had been arrived at by a number of individuals around the room. Obvious relationships between the ideas (for example incidents and major events having an impact on the timetable) were then added to form a graph (see Figure 2). The graph was compared with each of the five detailed use-case scenarios produced later in the workshop session and clusters of topics that might be involved in each were identified. In the end three strong clusters of terms became apparent, which supported the cross-interface condition monitoring and rolling stock configuration management scenarios from the workshop, along with a third grouping centred round a new use-case of end-to-end ticketing of a multi-modal journey; the scenarios are indicated in Figure 2 by red, green and blue shading respectively. The natural language descriptions of the questions that the prototype ontologies would attempt to address were:

Cross-interface condition monitoring scenario – A railway vehicle has been fitted with some ultrasonic inspection equipment. As it traverses the network it gathers data about the condition of the track it passes over. Data from the inspection, along with information on the infrastructure, are used to allow the set of possible fault types for a reading to be inferred.

Rolling-stock configuration management scenario – Railway vehicles have a comparatively long working life during which they will undergo several substantial maintenance processes. Components such as wheelsets may be removed from one vehicle, dismantled, overhauled, reassembled (potentially in different configurations) and then returned to use on a different vehicle, providing a substantial challenge in terms of the tracing of the assets, spares management and record keeping. Could a conceptual data model capture this information effectively?

Figure 2: Diagram of common themes from the brainstorming session of the data management workshop.

End-to-end ticketing of a multi-modal journey scenario – End-to-end journey tickets that allow travel on a variety of travel modes such as PLUSBUS are becoming more common. How could a conceptual data model for the railway industry interact with data from over transport modes to allow this type of information to be captured?

With the scenarios to be handled by the prototype ontologies identified, it became possible to proceed with the rest of the specification phase of the METHONTOLOGY process; the definition of the purpose, scope and sources of information available for the ontologies needed by each use-case. It is worth noting at this point that the method also requires a "level of formality" to be assigned to each ontology, this is largely a point of historic interest since all the ontologies here, were they to be implemented, would be done so in the Web Ontology Language (OWL) and therefore are defined as "semi-formal"; METHONTOLOGY dates from a point before the creation of the semantic web & OWL, at which point it was common for ontologies to be defined in a wider variety of ways including "formal" approaches such as first-order logics. It is for this reason that the level of formality field exists in the method, even though it is now largely redundant.

The specifications for the ontologies required by the use cases are presented below as a set of "index cards". Where a scenario reuses an ontology already defined for one of the previous use-cases, the card only contains the title and is put in as a placeholder.

Cross-interface condition monitoring scenario

Vehicle	
Domain	Rail
Level of Formality	Semi-formal
Purpose	Providing a description of any vehicle (including road-rail) that may be using the railway.
Scope	Vehicle types, vehicle components
Sources	RailML, InteGRail, Railway Group Standards

Infrastructure (track)	
Domain	Rail
Level of Formality	Semi-formal
Purpose	To enable the description of a physical section of railway track.
Scope	Rail types, manufacturers, weld & joint types, switches & crossings, ballast
Sources	Network Rail track design handbook NR/L2/TRK/2049, Railway Group Standards

Rail Fault	
Domain	Rail
Level of Formality	Semi-formal
Purpose	To enable the description of faults found in rails.
Scope	Controlled vocabulary of fault types, hierarchy of faults / classification for faults.
Sources	Rail failure handbook RT/CE/S/O57 Iss. 4 Network Rail track maintenance handbook NR/L3/TRK/002

Test equipment	
Domain	Rail
Level of Formality	Semi-formal
Purpose	Describe equipment that can be used to detect a fault in an asset.
Scope	Controlled vocabulary and classifications of test equipment.
Sources	Rail failure handbook RT/CE/S/O57 Iss. 4

Rail Network Topology	
Domain	Rail
Level of Formality	Semi formal
Purpose	Describe a railway network in terms of its fixed physical components
Scope	Relationships between infrastructure elements
Sources	RailML, InteGRail

Location	
Domain	Rail
Level of Formality	Semi-formal
Purpose	Define the location of an asset in the railway network.
Scope	GPS, metres/kilometres from point, miles and chains from point
Sources	N/A (expected reuse)

Time	
Domain	Rail
Level of Formality	Semi-formal
Purpose	Define points in time and time intervals.
Scope	Dates and times to second accuracy.
Sources	N/A (expected reuse)

Rolling-stock configuration management scenario

Vehicle

Supply Chain	
Domain	Rail
Level of Formality	Semi-formal
Purpose	Define the supply chain for vehicle parts.
Scope	Suppliers, components supplied (link to vehicle ontology), logistics and lead times.
Sources	Unknown

Maintenance Schedule	
Domain	Rail
Level of Formality	Semi-formal
Purpose	Capture of the maintenance processes that are performed on a vehicle.
Scope	Times and places of maintenance performed, maintenance processes, schedules for planned maintenance tasks.
Sources	Rolling Stock Library, Railway Group Standards, RAVERS, existing maintenance scheduling systems (SAP, SPEAR etc.)

Train Consist	
Domain	Rail
Level of Formality	Semi-formal
Purpose	Allow the description of a group of vehicles that form a train.
Scope	All vehicle types (link to vehicle ontology), identification of non-valid configurations based on coupling methods, electronic connections etc.
Sources	Rolling Stock Library.

End-to-end ticketing of a multi-modal journey scenario

Timetable	
Domain	Rail
Level of Formality	Semi-formal
Purpose	Describe a timetable.
Scope	Points of origin, destinations, departure and arrival times.
Sources	RailML, InteGRail, NetEx

Ticketing	
Domain	Rail
Level of Formality	Semi-formal
Purpose	Describe the fares payable for a journey.
Scope	Journey departure point, journey destination, route, cost, ticket types, ticket medium, railcards, ticket delivery method, payment method, refunds.
Sources	Unknown

Non-rail Transport Mode	
Domain	Transportation
Level of Formality	Semi-formal
Purpose	Describe connections to external transportation modes.
Scope	Connection points to external modes, external mode types.
Sources	NetEx, REWERSE

Rail Network Topology

With the specification phase of METHONTOLOGY complete, it became obvious that the amount of time available in the feasibility study was not going to permit even a prototype implementation of the number of ontologies required for all three scenarios of use. As a result of this, it was decided that only the condition monitoring scenario would be followed through to the level of prototype ontologies. The ability to reuse the Vehicle ontology developed for the condition monitoring scenario would mean that a simple example of a change in vehicle configuration could also be given, while the majority of the configuration management scenario, largely owing to its complexity and the time required to gather the appropriate domain knowledge, would be used as the basis for a future stage of the model development work. The end-to-end ticketing scenario would also be left at the specification stage.

Knowledge Acquisition

The knowledge acquisition phase of the METHONTOLOGY process involves the gathering, and to a certain extent the structuring, of the necessary domain knowledge for an ontology from the sources identified during the specification stage. Depending on whether the ontology engineer already has specific knowledge of the domain, knowledge acquisition can be the most time-consuming stage of the ontology development process, potentially taking many weeks to complete; fortunately, previous work at the University of Birmingham has resulted in a wiki-based repository of railway domain knowledge drawn from the Railway Group Standards. At the time of writing of this document, the wiki contained information on over 2,500 topics and covered subjects including vehicle design and signalling systems. For an example section of a wiki page, see Figure 3. Although, based on discussions with industry representatives during the feasibility study, the wiki's current content (drawn solely from the Railway Group Standards) may not provide sufficient information on which to base a complete set of domain ontologies for the railway industry, it did provide a convenient glossary of domain terms for use during the prototype model development work. Further development of the wiki will be an important branch of future work in this area.

With a glossary of appropriate terms already completed, work in the knowledge acquisition phase of the project could focus on the creation of lists of domain vocabulary for each ontology, as identified for the relevant sections of the wiki. Example vocabularies for the Vehicle and Rail Fault ontologies can be seen in "Appendix C – Example Vocabulary Lists from the Development of the Vehicle and Rail Fault Ontologies".

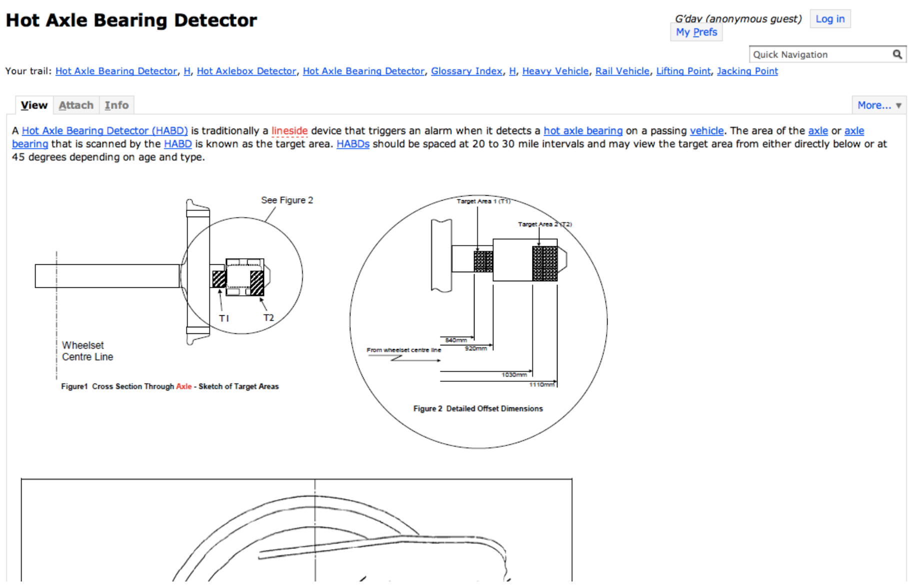

Figure 3: Example section of a page from the University of Birmingham railway domain wiki.

Conceptualisation

Conceptualisation involves the separation of the knowledge into concepts (the "ideas" of things, such as "vehicles" or "inspection methods"), properties (such as number of axles) and instances of objects (377 501) within the domain. This process can often be guided by simply identifying nouns and adjectives within the vocabularies of domain terms. For a relatively informal conceptualisation of the terms from the Vehicle ontology, see "Appendix D – Conceptualisation of Vehicle Vocabulary Terms".

Integration

As mentioned previously in this report ontology reuse is an important component of the ontology development process, both reducing development time and increasing consistency across models. Within the prototype modelling work, two topics for which ontologies were likely to already exist were identified, those of time and location. Upon further research, the W3C OWL-time ontology was found to be a good match for the requirements laid out in the specification phase and would be appropriate for reuse within a railway domain ontology. Location proved harder, with no existing ontology containing miles & chains distances from a point being found. The DARPA DAML program has developed ontologies capable of capturing GPS positions, and it is possible that this could be extended to the railway domain to meet the required functionality.

It should be also noted that this type of concept is also widely covered by upper-level ontologies such as the IEEE-supported Suggested Upper-level Ontology (SUMO) or the OWL implementation of ISO 15926, which also captures changes to the composition of objects over time. Inclusion of either of these upper-level ontologies would likely meet the requirements for location and time at least as well as the solution identified above.

Implementation

In the implementation step of the METHONTOLOGY process concepts, properties and instances are combined to form the final ontology, which is usually expressed in an ontology representation language. At the point METHONTOLOGY was created a number of ontology representation languages were in use, but recent research by Cardoso (2007) has shown that most practitioners now use the W3C recommended language, OWL. Since OWL is an XML-based format, it is possible to code ontologies by hand, however the process is made far easier by the use of ontology editing software; of the available ontology editors Protégé is currently the most popular (and was the package used in the feasibility study), but alternatives include the NeOn Toolkit, TopBraid Composer and OntoStudio.

Evaluation

The methods available for the evaluation of ontologies are still very much a topic of active research, and it is likely that there may well not be a one-size-fits-all solution for this stage of the ontology engineering process. Broadly speaking, ontology evaluation methods can be split into two groups. The first, validation against the original specification, is particularly popular in cases where the ontology has been modelled from logical expressions of some kind and therefore a mathematical analysis of the result is possible. The second method of validation involves the ontology model being evaluated by a domain expert, this method is of course more subjective than a mathematical validation, but it could be argued to be a better test of whether the ontology produced meets the needs of practitioners within the domain, and more importantly can be applied to a much wider cross-section of the models developed than a mathematical evaluation.

For the feasibility study the relatively loose specifications and incomplete domain coverage of the prototype models produced meant that neither of these methods was particularly appropriate and so, instead, it was decided that the evaluation of the ontologies would be performed by illustrating how they could be used in the scenarios outlined during the specification phase.

Cross-interface condition monitoring scenario

The cross-interface condition monitoring scenario involved a vehicle inspecting the track and then inferring a list of possible defects at a site based on its position in the network and the data it had gathered. Since one of the most flexible ways of modelling objects in ontology is to build them up from their component parts, the prototype vehicle ontology defines a “Vehicle” as an object made up of instances of “VehicleComponent”. Figure 4 shows the make-up of the inspection vehicle instance to be used in the demonstration of the scenario, which contains a traction package, cab, bogies and a track inspection system. Orange arrows indicate a “isComposedOf” relationship between the vehicle instance and one of the sub-components it is made up of, this is used in reasoning operations to identify all the sub-components of a vehicle quickly and without the need for recursive algorithms. The grey arrows, which link the wheelsets and axles/wheels is a “hasVehicleComponent” relation that shows a more direct composition. The green arrow between the track inspection system and the ultrasound equipment is different to the others because the ultrasound equipment is not part of the vehicle ontology (it isn’t part of a vehicle and could be used elsewhere, by a track worker on a walking stick for example) even though the track inspection system itself is part of the vehicle. Instead, the ultrasound equipment is imported into the model from the “TestEquipment” ontology, and hence has a special “hasTestEquipment” relationship.

A quick inspection of the figure shows that the instance that has been created does not contain all the components in a railway vehicle (there are no brakes for instance). Under many data modelling approaches incomplete descriptions can cause problems when it comes to applying logic to the data; this is because the designer must either allow for all possible states of the model when implementing the logic to be used with the model, or must choose to only implement very "high level" descriptions for which the components involved are always going to be present. In ontology models a special assumption, known as the "open world", is made that states objects we don't know about ***might*** exist in the real world but not be included in the model. Because of the open world assumption, logic in ontology models will not fail just because an instance is incomplete, making it possible to create instances of only those components of a system needed to answer a particular question given the (potentially incomplete) data available.

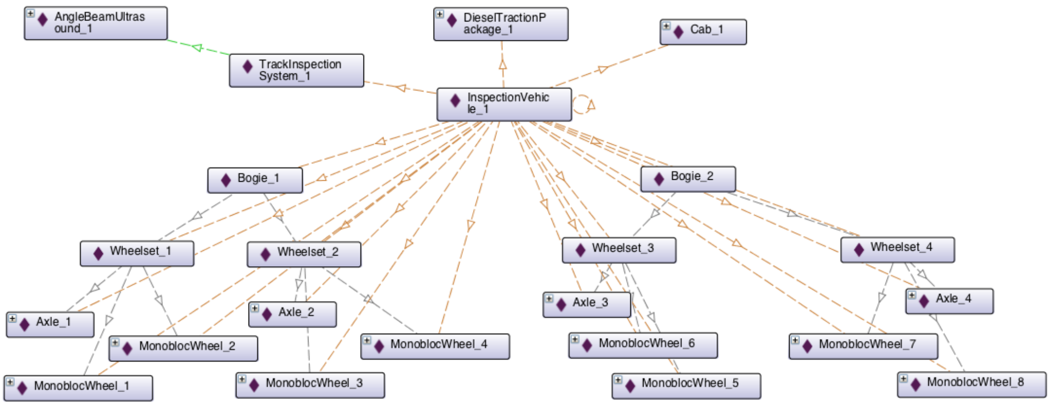

Figure 4: Instances of objects involved in the inspection vehicle.

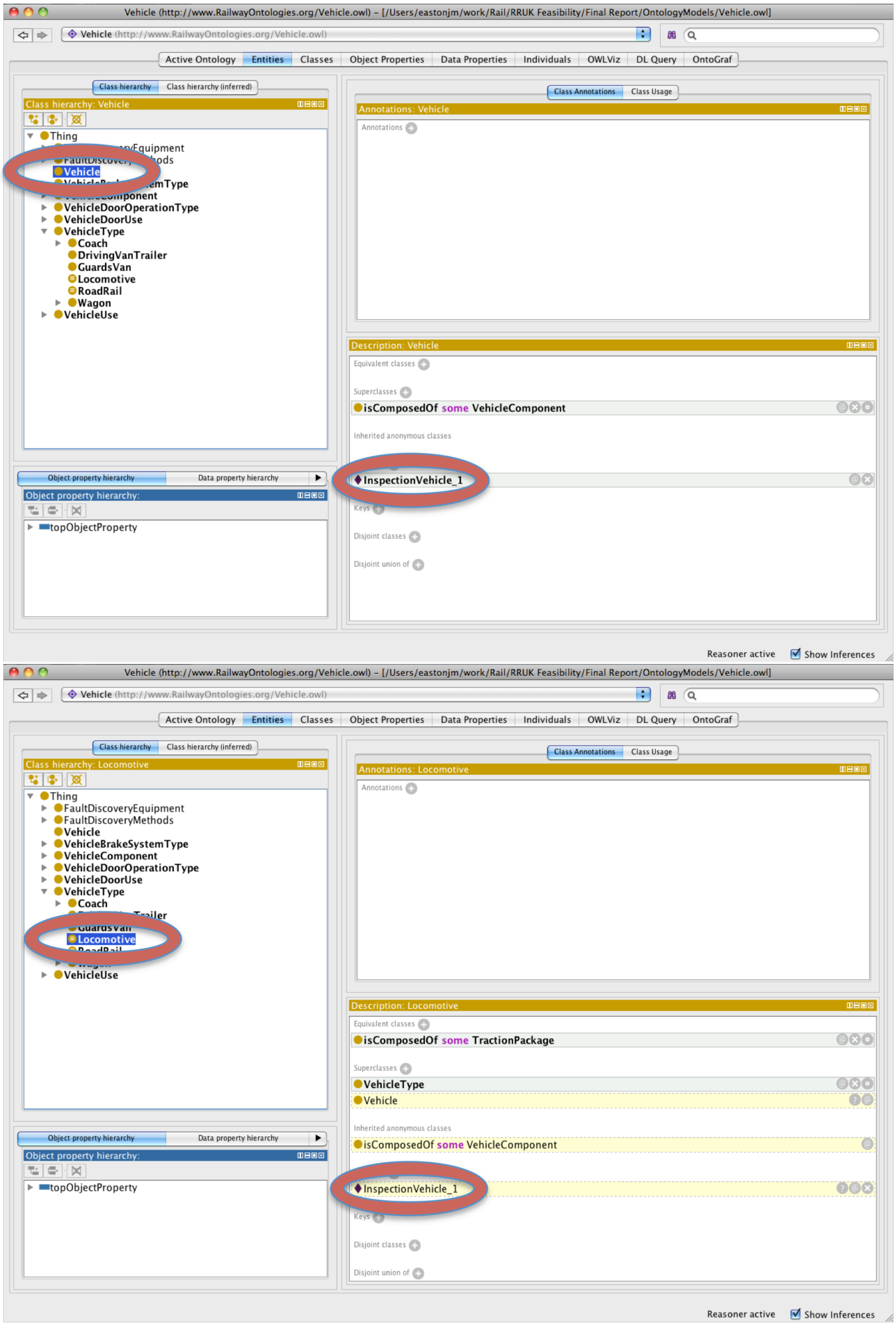

Figure 5: The inspection vehicle is an instance of the Vehicle class, but through inference can also be shown to be a Locomotive.

The upper portion of Figure 5 shows that the inspection vehicle is an instance of the class “Vehicle”. This is an important modelling decision, as it would allow software to request information about all vehicles on the network without the ontology designer having to know about every possible type of vehicle. It is however also quite limiting, as it’s quite possible that a user might only want to query for information on vehicles of a specific type, for example wagons. The lower portion of Figure 5 shows how the combination of an ontology model and inference solves this problem; in the upper section of the figure the inspection vehicle is shown with a grey background, this indicates that the instance was created from the “Vehicle” class, in the lower section it is show with a dotted yellow background, that is because the ontology model has inferred that the instance is also a “Locomotive”. The inference is based on the equivalence rule in the “Locomotive” class that states any “Vehicle” instance that contains a “TractionPackage” is also a “Locomotive”. This is hugely powerful behaviour, as it means vehicles such as the front & rear units of an EMU can be both members of the “Locomotive” and the “Coach” classes depending on the question being asked of the ontology.

In order to relate data recorded by the vehicle in the scenario to possible rail faults, it would be good to know which area of a particular rail the vehicle was passing over at the time; for this we need to know about the route the vehicle is taking and the physical infrastructure that route represents. Figure 6 shows an instance of a route created using the “RailNetworkTopology” and “InfrastructureTrack” prototype ontologies. The “Route” instance consists of a single “Section” (for clarity), and those classes are both defined in the “RailNetworkTopology” ontology. This then imports the “InfrastructureTrack” ontology (see Figure 7), and uses it to create two segments of “Track”, one of which follows the other, and which both consist of two “PlainLine” rails joined with fishplates. In a finished implementation of the ontologies, the “Location” ontology would be used to assign positions to the ends and middle of the rails, and the “Location” and “Time” ontologies could be used to assign recording position and time to the data, allowing them to be related to a part of a particular rail.

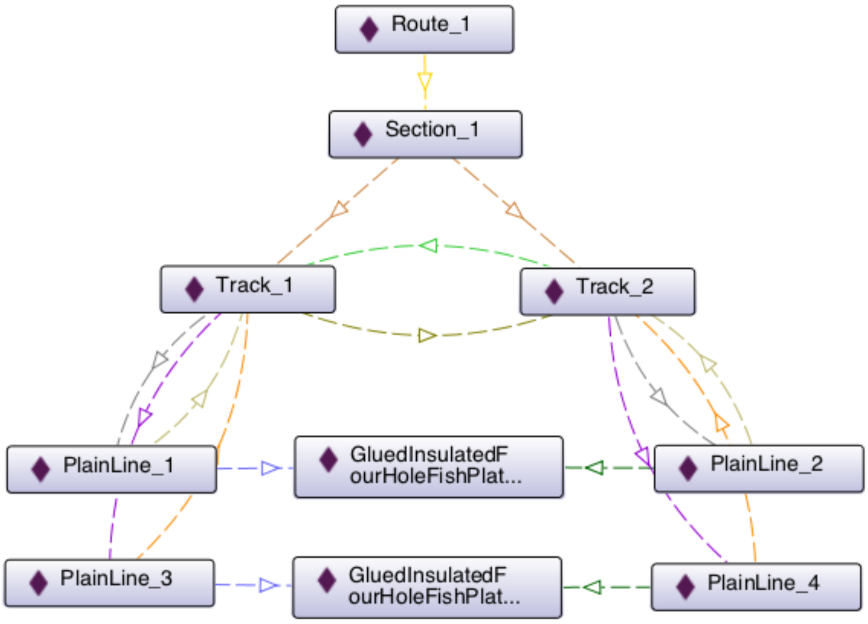

Figure 6: The route instance used in the cross-interface condition monitoring scenario.

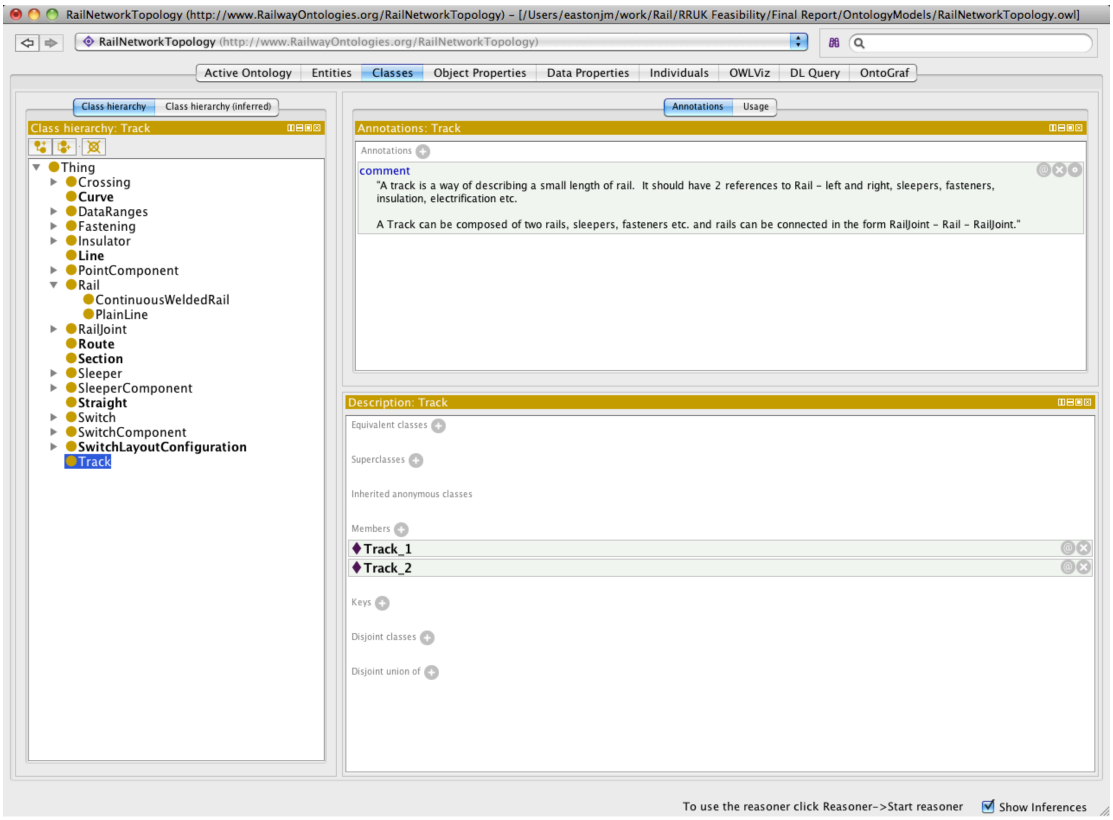

Figure 7: The "RailNetworkTopology" ontology after importing the "InfrastructureTrack" ontology.

With the data gathered and the track position determined, all that remains to be done in order to complete the scenario is the inference of a set of possible rail faults that correspond to the data recorded. The “RailFault” ontology (see Figure 8) allows different rail defects (named in the prototype ontology using the codes from the Rail Failure Handbook but with textual descriptions giving the natural language equivalent such as “wheel burn” or “long-pitch corrugation”) to be categorised based on location in the rail using inference, in the same way as the inspection vehicle was shown to be a locomotive. In the prototype “RailFault” ontology, quite a long list of possible faults for a rail area would be inferred, but with further development, additional ontologies could be used to further categorise the faults based on the types of data that can be used to detect them, the value of the reading etc. leading to much smaller lists or possibly even single defect types being inferred.

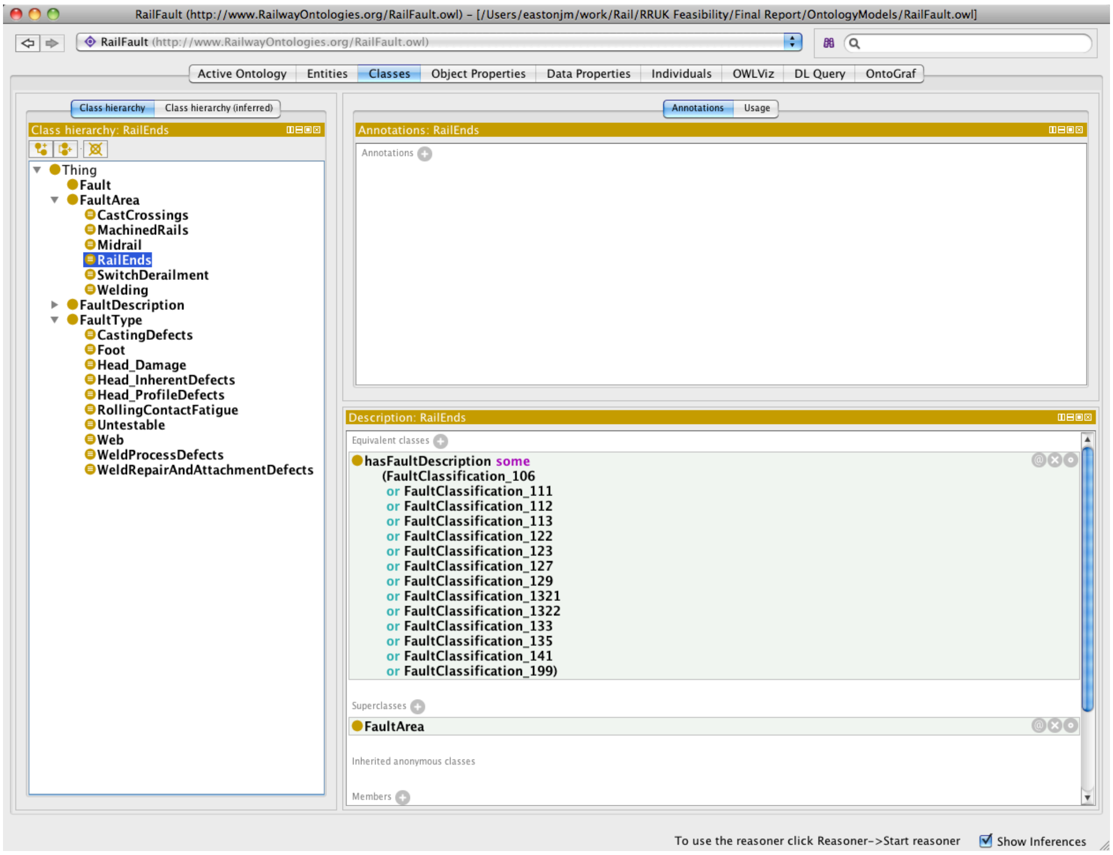

Figure 8: The "RailFault" ontology, which describes possible faults in a rail and allows them to be categorised by type or by the area in which they occur.

Rolling-stock configuration management scenario

While there was not time within the feasibility study to do all the necessary knowledge acquisition work to create a full set of prototype ontologies to support the configuration management scenario, it seemed appropriate since the vehicle ontology had already been created for the condition monitoring scenario, to show how easily ontology models can handle changes in the configuration of instances. Figure 9 shows two configurations of the same wheelset; in the upper portion of the figure, "TyredWheel_2" (and hence "Wheelset_1") is composed of "WheelCentre_2" and "Tyre_2". In the lower portion of the figure, after maintenance, the tyre on "TyredWheel_2" has been replaced with "Tyre_3". While this is a very simple example, it does show how the compositional nature of the ontology model makes both asset configuration and component history ("Tyre_2" still exists as an instance in the model) very straightforward.

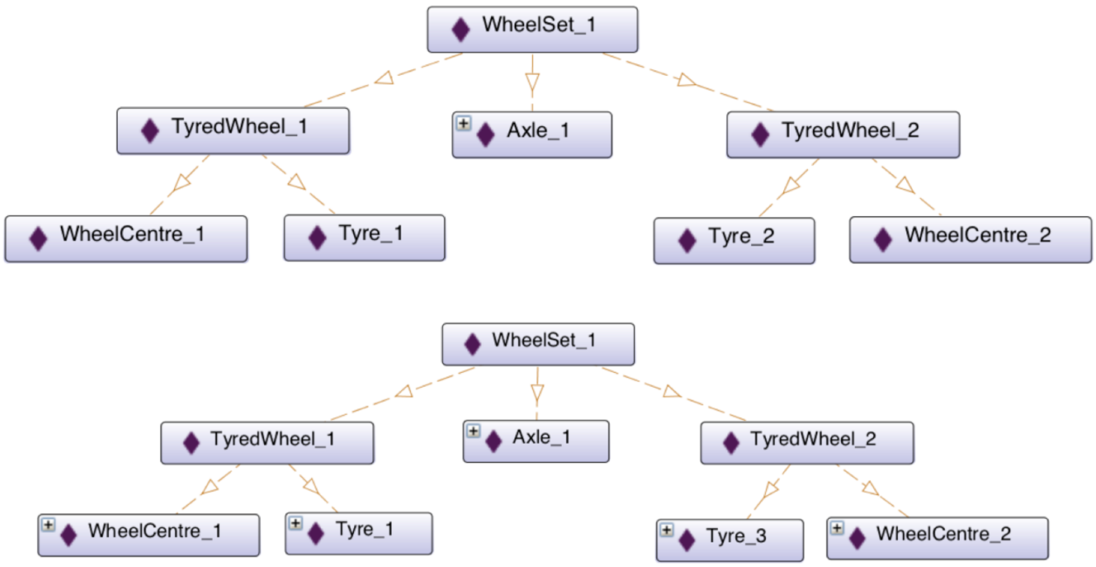

Figure 9: Example of the composition of a wheelset changing after replacement of a tyre.

A Roadmap for Future Conceptual Model Development Activities

Any development of a common, conceptual data model for the railway industry would require careful planning to both ensure that appropriate funding is available for the development activities and to make sure that the coverage of the domain at any point in the process matches the priorities / requirements of the industry. Figure 10 shows one potential roadmap for the development of this work to a useable system and can be broken-down into 3 main phases, architecture development, system-level modelling and complete system modelling.

In the first phase, which is expected to take around 2 years to complete, the core architectural components of the model will be defined leading to a 'lite' version of the data model. The lite model will be capable of describing the "railway system" but without being able to capture detailed information on any one sub-domain. The scope of the lite model is such that the development / further development of a domain knowledge repository similar to the wiki would be vital to this phase of the work, although the feasibility study has shown that the industry does not feel the Railway Groups Standards alone capture the domain sufficiently well to act as a basis for the model. Since it would be nearly impossible to generate a business case for this component of the work, it is envisaged that funding for phase 1 would need to be sought from the EU or similar bodies (see section "Dissemination and Impact of Feasibility Study Results") and strong industry support would be needed for such a bid. The need for the direct involvement of members of the industry in phase 1 of the development processes is expected to be less pronounced than in phases 2 & 3, although it is likely that the project team will need to work closely with both Network Rail and RSSB. An industry steering group should also be set up during this phase to oversee the development of the data model and ensure that it meets current needs.

Phase 2 of the work, "system-level modelling", will see the lite data model extended to include knowledge on technical sub-domains such as condition monitoring and vehicle maintenance. The knowledge in these domains is highly specialised and therefore phase 2 will require far greater involvement of industry members than phase 1, since it unlikely that the information required is all easily available in the public domain. Phase 2 is expected to initially take around 1.5 years, however ongoing maintenance will be needed to ensure it remains up-to-date beyond that point. With the core of the data model completed in phase 1, it is expected that building business cases within a sub-domain should be possible for the phase 2 work; as such it is expected that industry funding will be able to meet the costs associated with this work.

In the final phase of the model development work, the business logic associated with different summaries and reporting methods will be captured. The model will also be further extended to encompass the wider industry supply chain and passenger groups. As with phase 2 of the work, it is expected that the final phase will take approximately 1.5 years to complete and be funded by the industry.

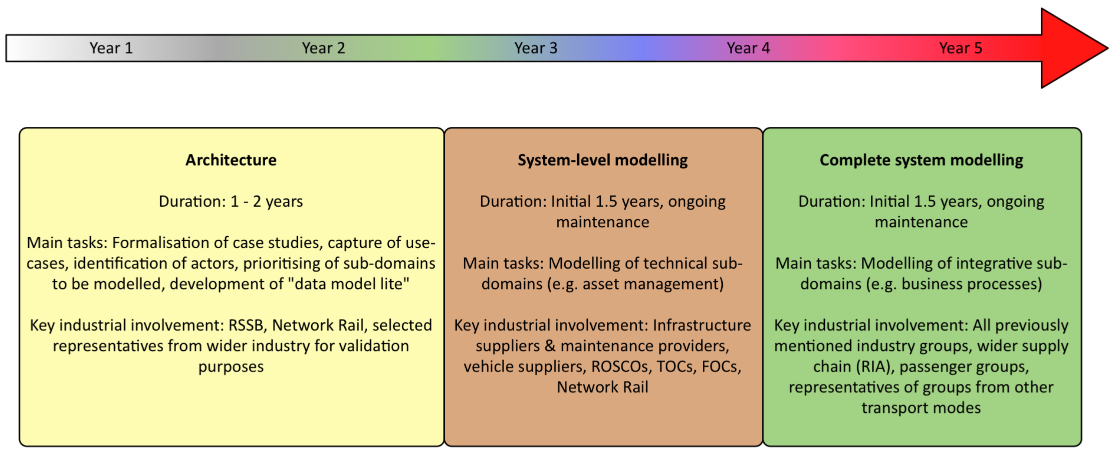

Figure 10: A suggested roadmap for the development of a conceptual data model within the railway industry.

Discussion

In order to evaluate whether or not a common data framework based around conceptual models has a role to play in encouraging a modal shift towards rail, it is important to consider if such a framework can answer several key questions raised at the workshops by industry members regarding the justification for the framework, and the access to, security, and usage of their data. The sections below discuss the outcomes of the feasibility study with reference to those questions.

Is a Common Data Framework for UK Rail Needed?

Despite the financial and practical difficulties associated with updating what is still, arguably, a functioning set of IT systems, and with an eye to avoiding giving a flippant answer to a difficult question, the fact that 153 separate use cases for a system-wide data framework were suggested within the scope of a short workshop exercise shows that there is a real need to formally begin addressing the question of data management within UK rail. Leaving aside the politically-loaded issue of wider data sharing between railway undertakings for a moment, the workshop sessions have clearly shown that there is concern within the industry over the number of manual interfaces between current IT systems; a view that echoed the findings of the recent McNulty report into value for money within the industry (DfT, 2011). The removal of manual interfaces between existing software could of course be achieved by the development of custom, end-to-end XML links between systems, but would that be the correct way forward for the industry? Certainly, in the short-to-medium term at least, the creation of a limited number of custom interfaces has its attractions; each interface would be relatively quick and cheap to implement, exchanges would be very compact owing to the minimal need for metadata in a closed system, and it would be comparably easy to ensure that security concerns were addressed. Ultimately however, specialised end-to-end interfaces would almost certainly cause more problems than they would solve; long-term maintenance bills would be high owing to software changes needing to be made in an ever-increasing number of models over time, the flexibility to change the overall system architecture would be reduced because of the effort associated with accurately replicating a large number of external links, and the development of new systems would be increasingly hard to manage. The creation of a common data framework by comparison, in which existing systems were mapped to a single data model, would involve a not-insignificant up-front cost and would require a longer development period before an initial set of systems could be connected; however, the architecture of such a system would be far more modular, facilitating the replacement of old systems with new, possessing lower on-going maintenance costs, and removing the need to change existing systems.

The provision of a single, common interface to industry IT systems would have significant value to offer OEMs, "levelling the playing field" between suppliers when submitting bids to extend existing products or systems, and allowing the purchasers of such extensions much more flexibility in their selection of preferred bidder; this would ultimately lead to a more competitive bidding process and to better value for money for the purchasing body. Since a common IT systems interface should, in theory at least, allow any railway undertaking to buy solutions from any supplier, it is likely that the benefits of a common data framework would disproportionately favour small and medium enterprises (SMEs) over larger organisations with existing solutions, encouraging wider innovation within the industry.

It is important to note that, in the short term at least, a system-wide data framework for UK rail would not replace all existing data models. Many of these models have evolved over time to become very efficient in the situations they are currently used in, they are well understood and may also be very tightly coupled into the software that uses them, making their replacement difficult. Instead, a system-wide framework would map to those existing standards (see Figure 11), enabling their continued use in situations where they are appropriate but also making that data available to the newer system. The workshop sessions identified Gemini, RAVERS, SAP, SPEAR, TOPS, TRUST, CCF and Intelligent Infrastructure as important industry data sources for which mapping should be considered.

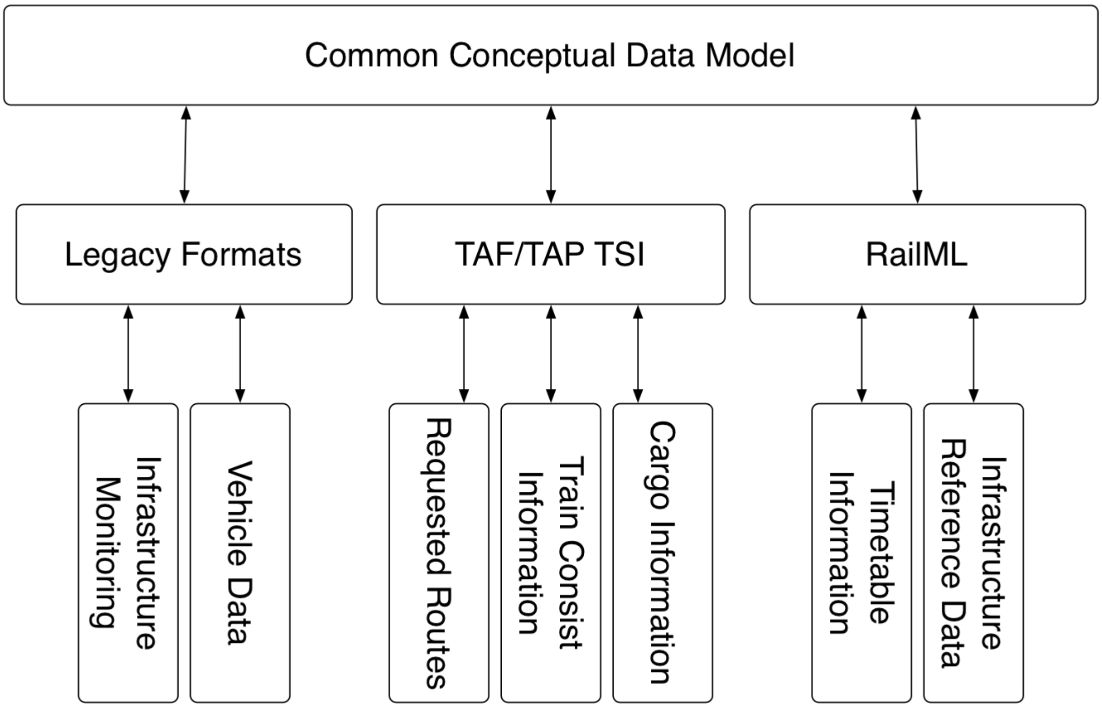

Figure 11: Mappings of existing standards to a system-wide data framework.

Who Should "Own" a Common Data Framework?

The issue of ownership of a rail data framework should in actuality probably be phrased as two separate questions, who owns the common data framework (model & architecture) itself, and who owns the data that it relies on?

It is quite likely that ultimately the framework would fall under the control of the Rail Systems Agency proposed by the McNulty Report (DfT, 2011), or at least a very similar body. Ongoing development activities would also require a steering group to ensure that the framework continued to meet the needs of the industry. For licensing purposes the data model, and potentially the architecture, should be kept as open as possible; at the very least developers would need to be allowed to produce derived projects without penalty, otherwise stakeholders could not create their own, local extensions to the model.

The ownership of any data is in some ways a far more difficult issue than the ownership of the framework. Common sense suggests that stakeholders voluntarily allowing access to their data over a system-wide framework must retain the rights to that data, and therefore be able to stop sharing that data at any time. However, uncertainty over the continuing existence of data sources might well discourage stakeholders from using the framework. A mandated data sharing policy would solve this problem, however it cannot be ignored that the ownership of data is of much more financial significance to some railway stakeholders than it is to others, and as such enforced data sharing may be politically very difficult to implement.

Who Would be Responsible for the Quality, Consistency and Timeliness of Data?

It should be noted that perhaps one of the greatest advantages of a common data framework would be that it would allow details that are already 'known' by the industry from personal experience or as a gut feeling to be proved using data from multiple different sources; a good example of this would be in the determination of best maintenance practice using data from different TOCs that are operating and maintaining vehicles of the same type. Use cases of this type would however, be heavily reliant on the consistency of data across the various contributing systems. Ensuring data quality is, generally speaking, a difficult issue for computer science and solutions usually involve either a "majority vote" taken between multiple providers offering the same information, or a user-assigned "reputation" for a particular data provider, where the provider with the highest reputation is assumed to be correct. Conceptual data models like ontologies allow, at least in theory, for this processes to be taken a step further, with multiple items of data being used to corroborate each other via inference operations; however, that type of technology is still a matter of active research and in the short term at least any system-wide data framework for the rail industry must rely on simple consistency checking techniques and a rubbish-in, rubbish-out approach.

The problem of inconsistency is likely to be particularly pronounced in the case of a system-wide data framework for the UK rail industry, where the current industry viewpoint, that data entered into IT systems is sacrosanct, must be challenged - particularly where data is known to have been exposed to manual system interfaces in the past. The question of who would ultimately be responsible for the quality of data accessed via a common framework is largely one of the mechanism by which provision of data takes place; if railway undertakings were required to allow access to their data by a regulating body (for example the DfT), then it is reasonable to assume that each undertaking could be held responsible for the quality of its own data, since they would have a duty to provide it. In the case of voluntary provision of data however, the issue is more complex; obviously, it would be irresponsible to allow data that was know to be inaccurate to be accessed via the framework and a certain due diligence in terms of data checking would have to be assumed, but at the same time making the data provider responsible for the quality of their data in that context would only serve to discourage stakeholders from sharing their information, thus limiting the usefulness of the framework across company silos. Instead, under a voluntary data provision scheme, it would be more appropriate for the risk to lie with the end-user of the system, and as such it would be their responsibility to validate any critical results carefully.

Who Should be Able to Access Which Information?

It was noted in the workshops that the rail industry in the UK generally places a lack of emphasis on data, its exchange, and its uses, within commercial agreements. While in a utopian society every stakeholder would make their data available to all via a system-wide data framework, the realities of business mean the best that could likely be achieved is a situation in which stakeholders make their data openly available unless there is a good commercial reason not to. There would then, be several possible different approaches to securing information within a system-wide data framework for UK rail. The first would treat all data as open, essentially creating a situation where stakeholders only made available that data which they were happy for all other parties to see; this would have the advantage of being very simple to use, but would likely result in very small amounts of data being shared (unless it was mandated otherwise) and would make it difficult for stakeholders to use data they were not sharing in conjunction with data accessed via the framework. The second approach, would be to allow security to be applied to data based on specific stakeholders or groups of stakeholders (such as all TOCs), this would be far more complex to apply than the first policy, and would have the potential to make the system-wide framework something of a joke if not carefully regulated, but it would at least ensure that sensitive data was not exposed to competitors. The final approach would be to allow stakeholders to make data accessible by the system either public (accessible to all) or private (accessible only to members of their own organisation). This approach would work in a very similar way to the first, apart from the fact that it would allow employees of the stakeholder providing the data to use their own private data, in combination with other stakeholder's public data, achieving a small additional advantage. Approach three would work well in a situation where a regulating body forced stakeholders to make all their information available unless it was deemed to be commercially sensitive.

The level of IT competence within the industry was identified during the project workshops as a potential barrier to the adoption of a system-wide data framework, and in an industry which is traditionally driven by heavy engineering rather than IT, simplicity of use and access to information must be a key aspect of any IT initiative. The early provision of interfaces to a wide range of to existing software would be a key component in the success of a system, and care would need to be taken to ensure that these focus on industry-used data manipulation packages such as Microsoft Excel, rather than just on software more common used for data processing in academic & scientific contexts.

Dissemination and Impact of Feasibility Study Results

Dissemination of the project findings to both the industry and academia is now underway. The following section outlines the dissemination work that has taken place to date, along with tasks that are planned as a result of the feasibility study.

Journal Papers

So far one journal paper reviewing ontology engineering methodologies has been published (Easton, Davies & Roberts, 2011), with a further two papers in preparation for submission before the end of 2011. Of these, one will use the results of the project workshops as the basis for a discussion of the requirements for an industry-wide common data model, along with the most significant barriers to its adoption; this paper will aim for publication in the *Journal of Rail Transport Planning & Management*. The second will present an expanded version of the outputs of the prototype modelling exercise, and is aimed at the Elsevier journal *Expert Systems with Applications*.

Conference Papers and Standardisation

A conference paper on "the development of domain ontologies for the railway industry" was presented to the World Congress on Railway Research in May 2011. This has led to an invitation to participate in one of the International Electrotechnical Commission's technical committees (TC9 - Electrical Equipment and Systems for Railways).

Grant Applications

Two grants for further funding in the area are being prepared as a result of this work and if successful could provide the necessary funding for the first phase of the conceptual model development. The first is in response to the Engineering and Physical Sciences Research Council's (EPSRC) "Autonomous and Intelligent Systems" call, and aims to produce an advanced system for condition monitoring data collection and analysis, back-ended by an ontology data model and with a case study in the railway domain. The second grant application will be submitted under the EU's FP7 ICT call 8, and will propose the development of a generic data management framework based on the complementary technologies of ontology and software agents. While the outputs of this work would be applicable to any large-scale infrastructure system, such as the water, gas or electricity distribution networks, the case study used during the project itself, as in the case of the EPSRC grant, will be built around railway applications.

Conclusions

The feasibility study discussed in this report aimed to determine whether a cross-industry, semantic data exchange framework would be of value to the UK railway industry with respect to better asset management and patronage, leading to reduced CO_2 emissions from the transport sector. In its exploration of this question, the study team organised two industry workshops and created a selection of prototype data models that supported use cases provided by industry stakeholders.

The project workshops showed that there are a number of concrete applications for a common data model within the UK rail industry. Many of these applications are system-wide and therefore contain not only elements of more automated data exchange between current industry IT systems, but also of wider data sharing between industry stakeholders. While the development of a common data framework for UK rail will be crucial to facilitating such exchanges of information, it will need to be supported by a philosophical movement within the industry towards more extensive data sharing in order to achieve its potential benefits and encourage the model shift this study was looking for.

The distribution of the applications produced across the 5 Cs of Cost, Customer, Capacity, Carbon and Commission/Cooperation indicated that, as expected, a common data model for UK rail would have little direct impact on the levels of CO_2 produced by the UK transport sector. Indeed only 6 of the 153 applications suggested were identified as being primarily related to Carbon. Instead, the majority of applications were attributed Cost, Commission/Cooperation and Customer, and focussed on areas such as asset management, cross-border services and more integrated planning around events and incidents. Many of the applications proposed had obvious links to improved customer service (better information during disruptions, end-to-end ticketing), cheaper railway operation leading to reduced fares (reduced maintenance, improved delay recovery leading to reduced penalties, asset life extension, data to support business processes such as renewal decisions) or greater availability of routes (reductions in planned and unplanned maintenance, better management of freight flows and vehicle location) and it is easy to see how such improvements could help to increase patronage by drawing passengers away from short-haul air travel thereby indirectly leading to a reduction in CO_2 emissions.

The perceived barriers to the adoption of a common data framework within the industry were seen to be problems resulting from the ownership of data, the need of less asset-rich stakeholders, such as TOCs, to protect any advantage their data provided them with, questions over the general levels of IT competence within the industry, and the difficulties associated with generating a coherent, stakeholder-level business case for a framework that delivered whole-system benefits. There was a general feeling amongst many of the stakeholders that the only way to overcome these difficulties and move the idea of a common data framework forward, would be for either RSSB or the DfT to champion the initiative; potentially to the extent of mandating the sharing of particular types of data via franchise agreements.

The breadth of application areas suggested during the first project workshop was very promising in terms of potentially paying for the eventual development of a model, and suggested that once a degree of core functionality was in place, benefits could be found in many areas of the industry. This supported the proposed development roadmap, which is based around a ‘lite’ data model, paid for by government or EU funding, with industry stakeholders supporting the future development of specific sub-domains that relate to their business activities once a clear business case can be determined.

Priority areas for framework development were seen to be the provision of geographical data at varying levels of granularity, and real-time operational data. Many of the application areas involved the linkage of this type of information to other, currently available datasets, such as condition monitoring information from fixed and mobile assets, and determining the current state of rolling stock with respect to its maintenance plan. The provision of historical information, while important, was not seen as a critical component of an initial data framework.

The prototype model development activities performed during the project have highlighted several key practical concerns that will need to be considered during any future model implementation. First and foremost amongst these is the difficulty associated with gathering sufficient knowledge on the railway domain to be able to create the initial ‘straw man’ model, and it has become clear as a result of the feasibility study that wider engagement with industry than was expected will be needed in order to perform this phase of the work. The structuring of the domain knowledge within a wiki has proven very valuable to the model development activities; both in terms of providing easy access to the information and as the potential basis for an audit trail showing why particular modelling decisions were made. Whether the wiki should only be populated with information from the group standards documents is however debateable, and it is likely that a future development activity would benefit from keeping the knowledge repository as open as possible; this would allow the writing of pages relating to particular domains at focussed workshop sessions, or even through a crowdsourcing process, where members of the industry can independently contribute and edit material related to their particular areas of expertise. A community-driven knowledge acquisition process would also serve as a useful lead into a later, DILIGENT style, ontology engineering exercise for the fine-tuning of the straw man, fostering a sense of involvement with the project and of ownership of parts of the model within the industry; this should be sufficient to encourage individuals to learn the skills needed to assist with the further development of the model.

In conclusion, this study found that a common data model for UK rail, preferably combined with wider data sharing amongst stakeholders, has the potential to both improve asset management within the industry and help encourage a shift in transport mode away from short-haul air travel and towards the railways; either of these outcomes would result in a net reduction in CO_2 emissions from the transport sector. The study has also shown that, while challenging, the implementation of such a model is possible, and that potential funding mechanisms do exist for those sections of the model for which it is hard to build a business case in the current, devolved railway system.

References

Batres, R., West, M., Leal, D., Price, D., Masaki, K., Shimada, Y., Fuchino, T. & Naka, Y. (2007). An upper ontology based on ISO 15926. *Computers and Chemical Engineering*, 31, 519–534.

Blomqvist, E. (2005). Fully automatic construction of enterprise ontologies using design patterns: Initial method and first experiences. In R. Meersman & Z. Tari (Eds.), *On the move to meaningful internet systems 2005: CoopIS, DOA, and ODBASE* (pp. 1314-1329). Springer-Verlag.

Blomqvist, E., & Sandkuhl, K. (2005). Patterns in ontology engineering: classification of ontology patterns. In *Proceedings of the seventh international conference on enterprise information systems (ICEIS 2005)* (pp. 413-416).

Brown, W., Malveau, R., III, H. M., & Mowbray, T. (1998). *AntiPatterns: Refactoring software, architectures, and projects in crisis*. John Wiley & Sons, Inc.

Cardoso, J. (2007). The semantic web vision: Where are we? *IEEE Intelligent Systems*, 22(5), 84–88.

Cassidy, P. (2008). Toward an open-source foundation ontology representing the Longmans defining vocabulary: The COSMO ontology OWL version. In *Proceedings of the third international ontology for the intelligence community conference* (pp. 45–48).

Clark, P., Thompson, J., & Porter, B. (2000). Knowledge patterns. In *Proceedings of the seventh international conference on principles of knowledge representation and reasoning (KR-2000)* (pp. 591–600). Morgan Kaufmann.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science 5 (2): 121–152.

Department for Transport / Office of Rail Regulation. (2011). Realising the potential of GB rail: Final independent report of the rail value for money study (ISBN: 978 1 84864 124 2). Retrieved from http://assets.dft.gov.uk/publications/report-of-the-rail-vfm-study/realising-the-potential-of-gb-rail.pdf

Devedzic, V. (1999). Ontologies: borrowing from software patterns. *Intelligence*, 10(3), 14–24.

Dietrich, J., & Elgar, C. (2005). A formal description of design patterns using OWL. In P. Strooper (Ed.), *Proceedings of the 2005 Australian software engineering conference (ASWEC'05)* (pp. 243–250). IEEE Computer Society.

Ding, Y., & Foo, S. (2002a). Ontology research and development. Part 1 - a review of ontology generation. *Journal of Information Science*, 28(2), 123–136.

Easton, J. M., Davies, J. R., & Roberts, C. (2011). Ontology Engineering The "What's", "Why's", and "How's" of Data Exchange. *International Journal of Decision Support System Technology (IJDSST), 3*(1), 40-53.

Fernández, M., Gómez-Pérez, A., & Juristo, N. (1997). METHONTOLOGY: From ontological art towards ontological engineering. In *Proceedings of AAAI97 spring symposium series, workshop on ontological engineering* (pp. 33-40).

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design patterns: Elements of reusable object-oriented software.* Addison-Wesley.

Gangemi, A. (2005). Ontology design patterns for semantic web content. In Y. Gil, E. Motta, V. Benjamins, & M. Musen (Eds.), *The semantic web - ISWC 2005* (pp. 262-276). Springer-Verlag.

Grűninger, M., & Fox, M. (1995). Methodology for the design and evaluation of ontologies. In *Proceedings of the workshop on basic ontological issues in knowledge sharing (IJCAI-95)*.

Maedche, A., & Staab, S. (2000a). Discovering conceptual relations from text. In W. Horn (Ed.), *Proceedings of the 14th European conference on artificial intelligence (ECAI 2000)* (pp. 321–325).

Maedche, A., & Staab, S. (2000b). Mining ontologies from text. In R. Dieng, & O. Corby (Eds.), *Knowledge engineering and knowledge management. Methods, models and tools: 12th International conference (EKAW 2000)* (pp. 189-202). Springer-Verlag.

Matuszek, C., Cabral, J., Witbrock, M., & DeOliveira, J. (2006). An introduction to the syntax and content of Cyc. In *Proceedings of the 2006 AAAI spring symposium on formalizing and compiling background knowledge and its applications to knowledge representation and question answering.* Stanford, CA.

Niles, I., & Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of the international conference on formal ontology in information systems* (pp. 2–9).

Pease, A., Niles, I., & Li, J. (2002). The suggested upper merged ontology: A large ontology for the semantic web and its applications. In *Working notes of the AAAI-2002 workshop on ontologies and the semantic web*.

Pinto, H.S., & Martins, J. (2004). Ontologies: How can they be built? *Knowledge and Information Systems*, 6(4), 441–464.

Pinto, S., Staab, S., Sure, Y., & Tempich, C. (2004). OnToEdit empowering SWAP: a case study in supporting DIstributed, Loosely-controlled and evolvInG Engineering of oNTologies (DILIGENT). In C. Bussler, J. Davies, D. Fensel, & R. Studer (Eds.), *The semantic web: Research and applications (ESWS 2004)* (pp. 16–30). Springer-Verlag.

Pinto, H.S., Staab, S., & Tempich, C. (2004). DILIGENT: Towards a fine-grained methodology for DIstributed, Loosely-controlled and evolvInG Engineering of oNTologies. In R. de Mantaras, & L. Saitta (Eds.), *ECAI 2004: Proceedings of the 16th European conference on artificial intelligence* (pp. 393–397). IOS Press.

Sommerville, I. (2007). *Software engineering* (8th Ed.). Addison-Wesley.

Staab, S., Erdmann, M., & Maedche, A. (2001). Engineering ontologies using semantic patterns. In *Proceedings of the IJCAI-01 workshop on e-business and the intelligent web*.

Sure, S., Staab, S., & Studer, R. (2009). Ontology engineering methodology. In S. Staab, & R. Studer (Eds.), *Handbook on ontologies* (second ed., pp. 135–152). Springer-Verlag.

Tempich, C., Studer, R., Simperl, E., Luczak, M., & Pinto, H.S. (2007). Argumentation-based ontology engineering. *IEEE Intelligent Systems*, 22(6), 52–59.

Uschold, M., King, M., Moralee, S., & Zorgios, Y. (1998). The Enterprise Ontology. *The Knowledge Engineering Review*, 13(1), 31–89.

Zhou, L. (2007). Ontology learning: state of the art and open issues. *Information Technology and Management*, 8(3), 241–252.

Appendix A – Suggested Applications for a Railway Data Framework

Group	Commission (44)	Cost (44)	Customer (28)	Capacity (23)	Carbon (8)	Misc (6)
Table 1	Standard time table between bus and train and EU Allocation of roles between actors (national and regional) Knowledge management Standard vocabulary for train systems and faults to allow data to be transferred between TOCs at Franchise change Increase knowledge of railway such as crossing boundaries of trains and infrastructure	Integration orchestration Maintenance planning Defect management Data sharing to optimise performance (capacity availability) Reduce cost of data – capture, storage, dissemination Maintenance life cycles Understand real asset usage for	Passenger able to understand 'railway' data and turn this into information that passengers needs especially during disruptions Transport planning times of disruption due to severe weather or national emergencies Data for information for while journey decisions / information –	Better data informed timetabling Assist real time train control System of system composability Standard interlocking control interface to enable easier move lost (?) effective control How much capacity does exist Unit formation – vehicles in a consist	Efficiency by hierarchy Energy usage Accurate energy metering Real time optimal driving control (adaptive on current state)	Could assist devolution to route structure by making it easier to share info within routes and still maintain control / oversee at centre

	With devolution of NR tracking that activities across devolved areas are common and standardized Interoperability reference One industry geographical data model Reference for stakeholders Cross industry trends Interoperability requires data sharing across EU and beyond	renewal planning Position within a maintenance programme of a vehicle Extending vehicle life Integrating condition monitoring data for track and train and visa versa Franchise tenders Data for whole system models for maintenance and renewal decision tools Increase productive by enabling people to easily discover and access data Better data - less management time	freight and passengers Govt gets more view on how railway is working Ticketing solutions Customer information services Organisation of "need to know" (or business point of view for confidentiality of information) Integrated transport systems – train to train, train to bus etc. Asset configuration Component fitments status –			

		spent on reconciliations Providing more complete whole life costing at high resolution by giving access to current data Failure analysis (root cause) Reduce failures asset train by sharing data between domains Cross function business solution sharing (modifications, step changes in maintenance) Standard output of data from OTMR	what is fitted to what (esp. 'swung' components) Miles run by vehicle Improved safety – swung components, vehicles crossing leases with poor data etc. Safety improvements Delay causation			

Table 2	Information display based on end user roles Rail operations integration Inter organisation incident management e.g. Olympic coordination Consistent approach to train location / tracking Freight logistical flows + tracking goods across different European countries Train delays impacting on future crew duties Crew management -> train cancellation i.e. loss of crew	Performance management Performance measuring KPIs / cross industry Asset lifecycle integration Fault reporting (DRACAS) Asset supply chain integration Providing a wide range of data types to test new algorithms Mapping train performance metrics to infrastructure impact (Next lot on border with customer)	End to end tickets for cross modal journeys Consistency of information to TOCs / pubic re service disruption due Signalling infrastructure problems -> need for re-scheduling, re-routing etc. How we mange delays – keep everyone in the loop	Combination of vehicle specs and track clearances to check route compatibility for cross Europe services Train consist information (which vehicle numbers in which order) Sharing of train service characteristics between TOCs / FOCs and operations control (formulations, coach class etc.) Journeys taken -> future planning	Driving optimisation i.e. train regulation and advisory speeds to avoid wasted energy Energy usage – how to improve it, how to pay for it, train class	

		Integration of track condition data with vehicle data to identify problem interfaces Sharing remote condition monitoring data about trains / infrastructure in order to make better quality decisions (both technical and operational) Improved maintenance opportunities – reduced costs, trending or measurements (deterioration) utilisation (???? from time based) How track defects (bumps, rough rides)				

		are reported and managed				
Table 3	Defining geography Defining assets / infrastructure Communication protocols standardisation Centralised safety component tracking system from supplier / repairer to end user – common model and framework Centralised national maintenance management data system – common model and framework Defect reporting and	Improved data quality (data only being entered once) Improved productivity (reduced inputs) Lifecycle planning 9for assets) in whole life costs Improve effectiveness of capital planning – govt (regulatory) -> industry -> customer Value for money planning – reduced no of applications	Performance management and improvement – including KPI Customer information Improve the quality of working life and job satisfaction	Improve safety and capacity by having a system wide view of the organisation Real-time timetabling/scheduling (end to end including train consistency, crew and non available network		

	corrective action systems at a national level from supplier to end user – common model an framework – improved data quality reduced cost Easier to extract data – data flowing form different systems Join up core business processes across the industry Improve alignment with business processes ((quality of information) Objective communication (clarity of targets and expectation within and across industry)	that require to be supported hosted + developed Reduce wasteful it/ in expenditure				
Table 4	What are our most critical assets	European wide economies of scale in rail data format	Faster fine grained distribution of timetable changes	How good are our assets Where is my rolling stock	Use of historic real better data to improve predictions and	Classed as 'all' How have our

	Share good practice Reduce duplication Info models (integrated) provide better basis for design + new requirements e.g. r + m data for asset types Automated gathering and distribution of disruption information in a standard incident format within rail and multimodal Avoiding re-inventing the wheel Supplier rail standard developments Which are the bet sites for trials Duty, representativeness Quality of data which will be available for	Cost of operations, maintenance interventions 1,3,2 what assets will I need in the future	to passengers How to optimise traffic patterns Availability of fare information to support advanced passenger information systems	How to optimise track and rolling stock How are my assets degrading Maintenance optimisation – the best schedule maintenance intervention and geometry adjustments to minimise risk of failures unplanned intervention Rail breaks – cracking – characteristics, sites, steel supply, age, traffic, other rails which might be similar maintenance interventions What trains could we utilise on a route Sharing of passenger congestion information to support passenger information and behavioural change applications	support transport modelling Overall equipment asset effectiveness – utilisation of track rolling stock, preparation of time at less than designated speed, quality of journey How are our assets utilised	assets failed What is our inventory of assets Better records of asset status saves effort in need for asset survey Asset records capture essentials of functions interfaces – key for considering impact of change

	assessment Influence location? ??? Specific steel sleepers – how to evaluate quantify advantage and ideal / most suite locations When can I maintain fixed assets Agreeing and aligning timetables for information					
Table 5	Interfacing between local regional centres (immediate operations) and national one (coordination intercity) Product comparison	Optimising supply chain Maintenance plans Defect mapping Rail grade developments Maintenance information – how to ensure that train	Operating standards e.g. where to find supporting information for station management Safety standards i.e. from a people competency	Data exchange between IM and TOCs for train information and passenger information to Adhesion information to help with SPADs, wheel flats, delays, problems (lowering speed across the network due to small LA) Interfacing interlocking with		Marked as all – metallurgical investigations

		is able to achieve its mission (diagnostic results) – exchanges between maintenance TOC IM	perspective Seat reservation free seat location available – bookable from platform Train movements – accurate movements and capacity of trains	control centres – different protocols – different data		

Appendix B – In-workshop Elaborations of Key Scenarios

Table 1

Scenario: Understanding vehicle status and history (e.g. franchise transfer)

Data types

Data	Notes	Source
Energy metering		
Gemini	Mileage	ATOS Gemini appliances
RAVERS (rail vehicle record system – legacy)		TRSMG, ATOS
Maintenance scheduler systems (SAP, SPEAR etc)		
OTMRS		
Rolling stock library (RSL)	Mileage, configuration, unit level, rolling stock fixed asset reg.	
Maintenance history including status of modifications for each		

vehicle		
Configuration management	Component, what manufacturer	
Component history,	Third party suppliers, bar-coding	
Paper records		
Risk register	Measuring risk to not achieve the mission – analysis of train interfaces and trends	
Maintenance history	TOC, ROSCO, third party, maintainer, system supplier	
Component records	Traceability of all components fitted to vehicles	
Vehicle acceptance	RVAR all vehicle asset register – rail vehicle acceptance register	NR
Asset configuration (Mod status)		TOC, ROSCO, third party maintainer systems
Vehicle history (whole life)		
Avoid imparting risk		

Drawings (PADS)		
FAMS	NR owned fleet – register, maintenance	NR
Safety system history	Component tracker etc	ATOC
Spares availability data for each vehicle		
Trend analysis		TOC bespoke systems
Vehicle failure history		BUGLE and TRUST
Vehicle reliability data		
ROSCO maintenance systems		
Prior lessee system		

Benefits

- Improved ppm
- Safety
- Warranty
- Tracking fitment and notifications to reduce potential for dip in performance as stock is transferred

Barriers

- IT competence
- Poor documentation
- Lack of open / standard interface
- High level collaboration (directors)
- Legacy
- Incentives for sharing
- Costs benefits
- Who owns what?

Table 2

Scenario: Track to train condition monitoring

Data types

Data	Notes	Source
Train formation	Unit and components, vehicle number, orientation	
Headcode	Vehicle ID, automatic?, down to level of wheels	Ops
Location of measuring equip	To within 10-20m	
Properties measured, / units		
Vehicle normal parameters		
Trended data from historical		
Appropriate operator and maintenance group		
Maintenance schedule / records		
Measurement time	Close enough to timetable time	CCF

Stakeholders

- NR
- TOCs
- FOCs
- Train maintainers and suppliers

Benefits

- More reactive maintenance
- Reduced delays minutes costs
- Reduced failures better customer experience
- Improved safety
- Increase asset life and availability

Barriers

- Need to agree on the data type / flow transmitted
- Business case - why should NR spend money to help TOCs
- Commercial arrangements cost / benefits in the right organisation

Table 3

Scenario: Operations planning (long term and real-time)

Data types

Data	Notes	Source
Routes		NR, DfT
Paths		NR, DfT
Capacity (actual / future)	Simulations	
Availability -> Network -> Equipment -> People		
Access rights		
Technical and physical constraints		
Demand		
Network operation		

Benefits

- Optimise performance
- Utilise network effectively
- Asset productivity
- Customer experience (delay)
- Operational costs

Barriers

- Structure of the industry
- Govt direction
- Inflexibility
- Contracts
- No feedback cycle

Table 4

Scenario: Increasing route utilisation (increasing usage of existing track; optimising track to support increased usage)

Data types

Data	Notes	Source
Track geometry, gauging, signalling		
Rolling stock	Max speed envelope tonnage condition	
Track asset info	Material, batch, repair	
Timetable data	Trust	
Real time history data		
Demand (passenger and freight)	Current and forecasts	
Constraints		
Tonnage data		
Climate, weather		

Stakeholders

- NR
- TOC
- TOC
- DfT
- Local government

Benefits

- Increased utilisation
- More money and passengers
- Better passenger info
- More predictable and reliable

Barriers

- Data standards (lack of)
- Fragmentation
- Supplier led
- Uncertain future
- Complexity
- Inertia

Table 5

Scenario: Supply chain data

Data types

Data	Notes	Source
Supplier database		
Installation date		
Maintenance schedules		
Track design geometry		
Train information (type, weight)		
Delays (track possession)		
Asset history		
Train schedule		
Operational environment		

Asset location
Recorded incidents

Benefits

- Better equipment for purpose
- Better knowledge of assets
- Better product design

Barriers

- Data not available or in enough detail
- Confidentiality
- A "shared vision"
- Volumes and sources of data
- Getting data to the right place at the right time
- Interpreting a lot of data

Appendix C – Example Vocabulary Lists from the Development of the Vehicle and Rail Fault Ontologies

Vehicle Ontology

Vehicle Type Terms

Passenger

Freight

Departmental (On-track machine, recovery vehicle, inspection vehicle)

Wagon (tank wagon)

Coach (saloon, sleeper, dining car)

Driving van trailer

Locomotive

Guard's van

Bogie / non-bogie

Capacity (number of occupants, weight)

Propulsion System Terms

Diesel

Electric

Steam

Traction motor (power, efficiency)

Electric traction motor (voltage, current)

Traction inverter

Diesel traction motor (fuel capacity)

Transmission (gear ratios)

Vehicle Power Supply System Terms

Overhead / pantograph

Third rail conductor shoe

Four rail conductor shoe

Main power supply (voltage)

Auxiliary power supply (voltage, number of batteries, capacity)

Brake System Terms

Power brake

Parking brake

Emergency brake

Air brakes (EP assist, EP control, direct air brake, automatic air brake)

Vacuum brakes

Brake discs

Brake pipes

Brake pads

Train Door / Door Control System Terms

Passenger door

Train crew door

Emergency exit

Power/manually operated

Door leaf

Door handle

Door lock

Hinge

External door status indicators

Door control system (selective door opening, automatic selective door opening, correct-side door enable, fully automatic selective door opening)

Fire Protection System Terms

Automatic fire detection (smoke, heat)

Automatic fire alarm (sounder, visual)

Fire extinguisher (fixed, portable)

Fire blanket

Fire containment (fireproof barrier)

Train Control Systems

European Train Control System (ETCS, level 0 – 3)

Train Protection & Warning System (TPWS)

Driver Only Operation (DOO)

Driver’s Safety Device (DSD)

Driver’s Reminder Appliance (DRA)

Automatic Train Operation (ATO)

Automatic Train Control (ATC)

Automatic Warning System (AWS)

Other System Terms

Tilt system

Cab radio

Sanding system

Closed circuit television

Passenger information system

Ventilation system

Lighting system (standard, emergency)

Suspension system

Coupling system (buffers, drawgear, gangway, electrical interconnect)

Other Possible Vehicle Components

Head / tail lamp

Vehicle body (side, floor, roof, frame)

Seat (back, cushion, armrest, trim)

Bogie (bogie frame, motor bogie)

Lifeguard

Compressor

Catering facilities

Horn (two note, three note)

Whistle

Boiler

Firebox

Train data recorder / OTMR

Window

Windscreen

Demister

Wiper

Underframe equipment

Cab / driving position (driving position, auxiliary driving position)

Vestibule

Wheelset

Axle

Axle bearing (plain, cylindrical, roller)

Axlebox

Wheel (tyred, monobloc)

Wheel centre

Wheel rim

Wheel tread

Wheel flange

Yaw damper

Jacking point

Lifting point

Toilet

Waste tank

Track circuit assister

Cab controls

Rail Fault Ontology

Rail Ends

Head-Profile Defects

- Crushing at rail end

Head-Inherent Defects

- Progressive transverse cracking (tache ovale) at rail end
- Horizontal cracking of head at rail end
- Longitudinal vertical cracking of head at rail end

Rolling Contact Fatigue

- Shelling at rail end
- Gauge corner cracking or head checking at rail end
- Squat at rail end
- Wheelburn at rail end

Web

- Horizontal cracking at the web-head fillet radius at rail end
- Horizontal cracking at the web-foot fillet radius at rail end
- Longitudinal vertical cracking (piping) at rail end
- Star-cracking of fishbolt holes

Head-Damage

- Battered rail end

Untestable

- Untestable rail end

Midrail

Head-Profile Defects

- Short-pitch corrugation
- Long-pitch corrugation
- Sidewear
- Abnormal vertical wear
- Crushing
- Localised head loss
- Insufficient rail depth
- Lipping

Head-Inherent Defects

- Progressive transverse cracking (tache ovale)
- Horizontal cracking of head
- Longitudinal vertical cracking of head
- Surface defects
- Local batter of the running surface

Rolling Contact Fatigue

- Shelling of the gauge corner
- Shelling of the running surface
- Gauge corner cracking
- Head checking
- Field flow or field cracking
- Tongue lipping
- Squat
- False flange damage
- Isolated wheel burns
- Continuous wheel burns

Web

- Horizontal cracking at the web-head fillet radius
- Horizontal cracking at the web-foot fillet radius
- Longitudinal vertical cracking (piping)
- Corrosion of web
- Cracking around holes other than fishbolt holes
- Diagonal cracking away from any hole
- Flame cut holes
- Lap

Head-Damage

- Bruising
- Rail head arcing damage
- Rail head deformation

Foot

- Tamper damage
- Rail foot arcing damage
- Longitudinal vertical cracking in rail foot
- Rail foot corrosion
- Gall

Untestable

- Untestable rail

Welding and Weld Repairs

Head-Inherent Defects

- Progressive transverse cracking (tache ovale) at a weld

Rolling Contact Fatigue

- Gauge corner cracking or head checking at a weld
- Squat on a weld
- Wheelburn on a weld

Web

- Horizontal cracking of web at a weld
- Cracking through fishbolt holes at a weld

Weld Process Defects

- Early/late tap
- Porosity
- Lack of fusion
- Inclusion
- Black hole
- Rail misalignment
- Weld collar misalignment
- Hot tear
- Oxidation
- Bad trimming/grinding

Weld Repair and Attachment Defects

- Transverse cracking of the rail head at a weld repair
- Detachment or shelling of the weld repair
- Transverse cracking under electrical connection

Switch Derailment Hazards

Head-Profile Defects

- Sideworn stock rail associated with a little-used switch blade
- Stock rail and switch blade both sideworn
- Stock rail headwear associated with a less headworn switch blade
- Switch blade damage
- Switch blade with a sharp gauge corner profile

Defects Associated with Machined Rails

Head-Profile Defects

- Lipping on reduced section rail

Head-Inherent Defects

- Progressive transverse cracking (tache ovale) in reduced section rail
- **Rolling Contact Fatigue**
- Shelling in reduced section rail
- Gauge corner cracking or head checking in reduced section rail
- Squat in reduced section rail

Web

- Cracking associated with stretcher bar connection
- Other cracking of the web in reduced section rail

Foot

- Transverse vertical cracking at machined rail angle

Cast Crossings

Head-Profile Defects

- Misshapen crossing nose profile
- Crossing nose wear
- Lipping on crossing

Head-Inherent Defects

- Transverse cracking of casting – crossing vee
- Transverse cracking of casting – wing rail
- Transverse cracking of casting – leg ends
- Transverse cracking of casting – elsewhere in the box section
- Longitudinal vertical cracking of casting

Rolling Contact Fatigue

- Shelling of the gauge corner of a casting
- Shelling of the running surface of a casting
- Gauge corner cracking or head checking on a casting
- Squat on a casting
- False flange damage on a casting
- Wheelburn on a casting

Web

- Horizontal cracking in the upper fillet radius of a casting
- Crack at a hole in the web in a casting

Foot

- Vertical transverse cracking of the foot in a casting

Elsewhere in the casting

- Longitudinal cracking of the apron
- Horizontal cracking of the flangeway wall
- Longitudinal cracking of the base of the flangeway
- Flash butt weld at stainless steel insert

Appendix D – Conceptualisation of Vehicle Vocabulary Terms

Vehicle

Vehicle Use

- Passenger
- Freight
- Departmental

Vehicle Type

- Wagon
- Coach
- Driving Van Trailer
- Locomotive
- Guard's Van

Vehicle Internal Supply Type

- Main
- Auxiliary

Brake System Type

- Power
- Parking
- Emergency

Door Usage Type

- Passenger
- Train crew
- Emergency exit

Door Operation Type

- Power
- Manual

Vehicle Component

- Armrest
- Axle bearing
- Axle trust pad
- Axlebox
- Battery
- Body side
- Bogie
- Bogie frame
- Boiler

- Brake disc
- Brake pad
- Brake pipe
- Buffers
- Cab
- CCTV camera
- CCTV monitor
- CCTV recorder
- Compressor
- Conductor rail shoegear (3, 4-rail)
- Coupler
- Demister
- Door
- Door control
- Door handle
- Door status indicator
- Door traction interlock
- Drawgear
- Driver's reminder appliance
- Driver's safety device
- Driving position
- Fire alarm (audible, visual)
- Fire blanket
- Fire detector (heat, smoke)
- Fire extinguisher
- Firebox
- Gangway
- Horn
- Jacking point
- LCD display
- Lamp (head, tail)
- Lifeguard
- Lifting point
- Lighting system
- Microphone
- Motor (diesel, electric)
- OTMR
- Pantograph
- Passenger information system
- PA system
- Roof
- Sanding system
- Seat
- Seat reservation computer
- Speaker
- Suspension
- Thrust pad axle

- Tilt system
- Toilet
- Track circuit assistor
- TPWS
- Transmission
- Tyre
- Ventilation system
- Vestibule
- Waste tank
- Wheel centre
- Wheelset
- Whistle
- Window
- Windscreen
- Windscreen wiper
- Yaw damper
- **Traction Package**
  - Diesel
  - Electric (AC/DC)
  - Hybrid
- **Driving Position**
  - Full
  - Auxiliary
- **Train Control System**
  - European Train Control System
  - Automatic Train Operation
  - Automatic Train Control
- **Internal Power Supply**
  - Main power supply
  - Auxiliary power supply
- **Door Control System**
  - Selective Door Opening
  - Automatic Selective Door Opening
  - Fully Automatic Selective Door Opening
  - Correct-side Door Enable
- **Air Brake Type**
  - Automatic air brake
  - Direct air brake
  - Electro-pneumatic assist air brake
  - Electro-pneumatic control air brake
- **Axle**
  - Plain bearing axle
  - Roller bearing axle
- **Wheel Type**
  - Monobloc wheel
  - Tyred wheel
  - Road wheel