Manifestation of ontologies in graph databases for Big Data Risk Analysis

M. Figueres-Esteban, P. Hughes, R. EL Rashidy & C. van Gulijk
Institute of Railway Research, University of Huddersfield, Huddersfield, UK

ABSTRACT: Big Data Risk Analysis (BDRA) intends to combine the huge volume of information that railway systems produce from a variety of data sources for safety and risk management. One of the most challenging issues is how safety scientists can use big data techniques. This is especially important in the light of data coming from different systems that hold information about critical events, hazards or controls. Yet, the integration of complex safety-related data is not just another IT problem. Fundamentally, it requires the expertise of safety experts in order to make sense of the data. A solution lies in the use of graph databases and ontologies to access this data for safety purposes. This type of database allows for the handling of huge amounts of data whilst it is still accessible to safety experts that are not gifted programmers. This approach opens up big data for safety management and enables a plethora of possibilities for future safety research.

1 INTRODUCTION

A BDRA safety system is defined as an enterprise safety management system that performs the following:

- Extracts information from mixed data sources.
- Processes it quickly to infer and present relevant safety management information.
- Combines applications to collectively provide sensible interpretation.
- Uses online interfaces to connect the right people at the right time.

In order to:

- Provide decision support for safety and risk management.

This definition guides the development of BDRA systems that are of use to companies that work on the GB railways. BDRA aims to use big data analytics techniques for safety (Van Gulijk et al. 2018; Van Gulijk et al. 2017).

One of the key challenges of BDRA is to store and process that massive amount of data and manage the heterogeneous knowledge from different information systems to obtain safety insight. A solution lies in the use of graph databases that are controlled by ontologies to represent a common framework of understanding and integrate data. (Figueres-Esteban et al. 2016; Van Gulijk et al. 2016). The method explained in this paper opens up Big Data for safety scientists. The method is straightforward but powerful and does not rely on gifted programmers. In theory, the database is infinitely scalable so it is hard to predict the limitations of the approach.

2 DATABASES AND BIG DATA

In the last decades, relational databases (aka SQL databases) have dominate the market of databases until coming up a standard way. They are structured in tables for access and have been very efficient when it comes to rapid and efficient access to data. Nevertheless, in big data environments where huge amounts of information have to be stored and integrated from new unknown sources, relational databases become unwieldy (Sadalage & Fowler 2013).

A solution to bypass this problem is to omit the relational table by simply storing data in a system that, for lack of a better example, finds its analogy in an infinitely scalable library card catalogue (Van Gulijk et al. 2018). Databases that work in that way are called NoSQL databases.

2.1 *GRAPH DATABASES*

In a relatively novel development these NoSQL databases have been enriched with a sensible visual interface based on graphs. They are simply called graph databases. A graph database is database management system that store data in the form of a property graph (Robinson et al. 2013). Safety scientists will recognise a property graph as a collection of nodes and links; as we often see them in our work.
The organization of the data in graphs is extremely useful in terms of understanding (Figueres-Esteban, Van Gulijk, et al. 2015; Figueres-Esteban, Hughes, et al. 2015). Graph databases allow to represent different types of data models into a common space in order to integrate diverse type of data (EL Rashidy et al. 2017). This issue is a key aspect in order to implement ontologies that represent the knowledge of technical domains such as railways, risk and safety.

3 KNOWLEDGE REPRESENTATION IN RAILWAYS

Railways are a complex systems that represent a rich tapestry of different types of organisational knowledge, created for different purposes and people with different expertise, skills and competences in many different contexts. Bringing together all the data that railways produce means to make sense of heterogeneous knowledge from different information systems.
The most common technique used by computer scientists to represent a common framework of understanding and manage the knowledge is an ontology. A formal, and broadly accepted, definition of an ontology is provided by (Gruber 1995): "*An ontology is an explicit specification of a conceptualization*."
There are different types of ontologies such domain and application ontologies depending on their specificity of the knowledge (Guarino 1997).
In the railway domain, the FP6 European Integrail project (http://www.integrail.eu/) and the RailML community (http://www.railml.org) proved the utility of ontologies in the communication and integration of data through railway information systems (Van Gulijk & Figueres-Esteban 2016).

4 ONTOLOGIES AND GRAPH DATABASES FOR BDRA

Different ontology languages and frameworks have been developed to support the implementation of an ontology (Corcho et al. 2003). The challenge is that a single ontology should be understood by people and machines.
One of the most used frameworks is showed in the left side of Figure 1. Different data structures represented in formats such as XML, JSON and CSV can be integrated through ontologies implemented in RDF/OWL languages. These languages support the application of Artificial Intelligence (AI) in order to reason with the represented knowledge. The approach that this work is taking bypasses complicated ontology languages and replaces it with a relatively straightforward visual interface in a graph database. This is where we omit the need for gifted programmers.

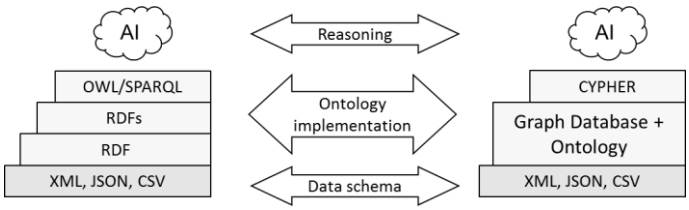

Figure 1. Transformation of the stack of ontology languages for BDRA.

This paper demonstrates how to use the framework showed in the right side of Figure.1.

5 METHODOLOGY

The paper describes the implementation of a railway domain ontology by safety experts in order to connect three different data sources to an event related to safety management.

The current BDRA project focuses on understanding SPAD risks (passing red signals) but for the benefit of explaining the method we focus on part of that risk: the "signal obscured" hazard. This means that safety records related to obscured signals and instances of a signal database have to be found and linked to enrich the analysis of these type of events. The methodology has three basic steps:

a) Selection of data sources and storing data in a graph database.
b) Building the signal domain ontology.
c) Implementing the signal ontology for the integration of data.

5.1 *Data sources*

This trial uses four *.csv files extracted from three information systems: three files of text records from the SMIS and IFCS systems of Railway Safety and Standard Board (RSSB) containing around 100,000 incidents and a table of signals from the Ellipse Asset Management tool of Network Rail (NR) containing 40,000 descriptions of signals.

SMIS is a database for recording safety-related events that occur on the rail network in Britain (RSSB 2017). Railway stakeholders such as NR or train/freight operators enter about 75,000 events per year such as derailments and SPADs. In this exercise, we are just using records related to obscured signals. IFCS is a database that focuses on human performance and underlying causes of rail incidents. These underlying caused are classified using 10 Incident Factors that are breakdown by different levels of subcategories (Gibson et al. 2015). The table of signals is a sample of descriptions of signals that is part of the Ellipse Asset Management tool of NR. Figure 2 shows the properties that were used to integrate the data.

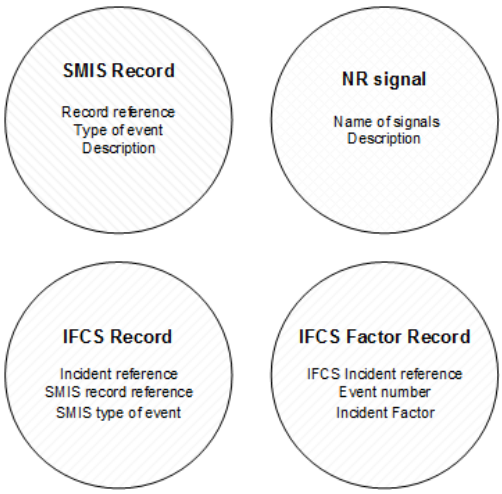

Figure 2. Description of the data sources used to support data integration.

5.2 *Signal domain ontology*

The purpose of the signal domain ontology is to align data structures of the information systems with an accepted reference framework by railways. For this exercise, the reference framework for the signal domain has been the railway signal standard in UK (RSSB 2015).

The sources showed below have been used to build the ontology:

- The Signals, handsignals, indicators and signs. Handbook RS/521 Issue 3 (December 2015).
- The data model of the SMIS+ program.
- The schema of the table of signals.

The standard RS/521 provides a classification and description of all railway signals in UK. The data model of the SMIS+ includes a taxonomy of railway signals that is aligned with other reporting systems. The data structure of the table of signals does not provide a signal taxonomy but it can be extracted from the field "Item Name". An example of value of this field is "EZ220 - SIG HEAD - COLOUR LIGHT – LED". Note that programmers don't have the expertise for this exercise, even if they are gifted. The interpretation and consideration of safety-aspects lies within the remit of safety experts.

5.3 *Implementation of the ontology and data integration*

The data from the information systems were stored in a Neo4j graph database. Each row of the data files represents a node in the graph database and each node has as many properties as columns the data file has. In this first step, the database has no structure and it just stores data under a label (data nodes).

In the same database, the signal domain ontology was implemented in a graph data model (ontology nodes). Using the properties of the ontology nodes and analysing the property of the data nodes that stores the text of the record, the links between each node were created.

The signal obscure event (event node) was connected to data nodes of signals and these ones were connected to the data nodes of SMIS/IFCS records.

6 RESULTS

Table 1 shows an excerpt of the extracted ontology from the table of signals. This ontology was mapped with the explicit ontologies of the standard RS/521 and the SMIS data model. Table 2 shows an excerpt of the mapping table. Figure 3 shows a piece of the final signal ontology.

Table 1. Excerpt of the signal taxonomy from the table of signals.

Item name			
First token	Second token	Third token	Fourth token
EZ220	SIG HEAD	COLOUR LIGHT	LED
EZ101			1 ASPECT
EZ102			2 ASPECT
EZ103			3 ASPECT
EZ104			4 ASPECT

Table 2. Excerpt of the mapping between the signal taxonomies of the RS/521 standard, SMIS and the table of signals.

RS/521	SMIS	Table of signals
SPAD indicator	SPAD indicator	ES100
Limit of shunt signal	Limit of shunt	EZ160
Point indicator	Points indicator	BR100, EZ170

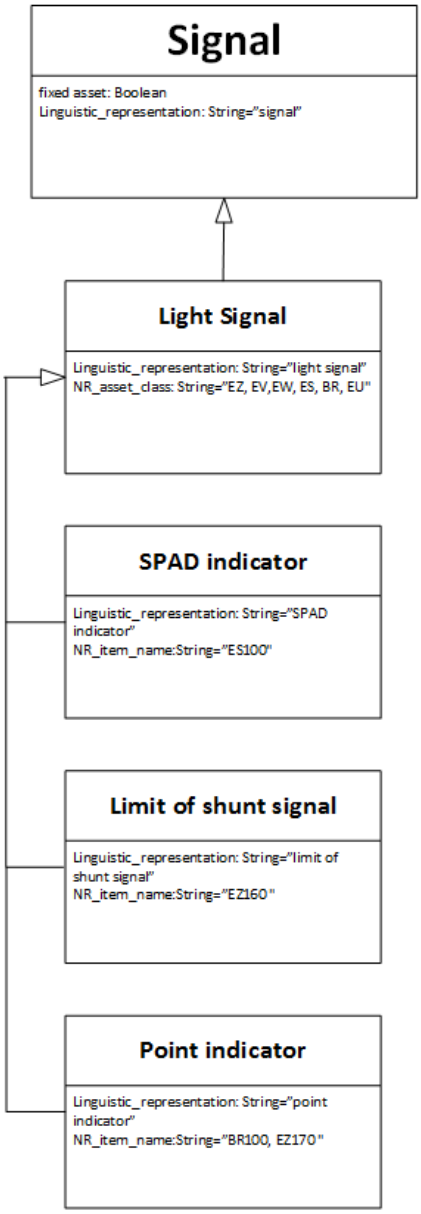

Figure 3. Excerpt of the UML diagram that represents the signal ontology.

Figure 4 shows an excerpt of the graph database that contains part of the implementation of the signal ontology and instances of signal and SMIS/ IFCS records. The ontology is connected to the signal nodes that are connected to the SMIS/IFCS records and the signal obscure event.

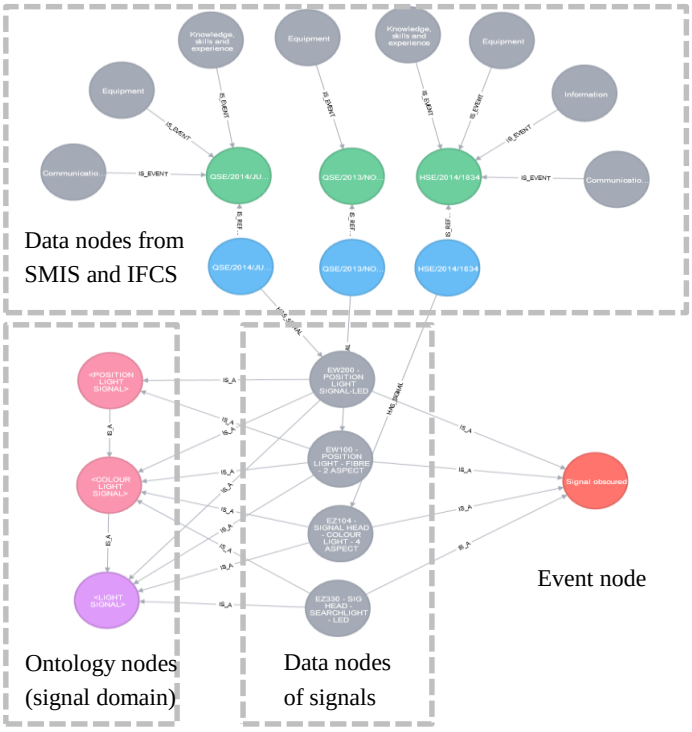

Figure 4. Excerpt of the graph database that integrates different types of instances of data with the event "Signal obscured" by means of the signal domain ontology.

7 DISCUSSION

This paper shows that NoSQL graph databases guided by ontologies enable safety scientists to work with big data techniques without the intervention of IT experts. Some programming is required but most of it is not much more complicated than excel macros or Matlab. The expertise of safety experts, however, is fundamentally required to build safety and railway domain ontologies to support the integration of data for further safety analysis.

This work demonstrates that graph databases can store complex data structures as single nodes, which helps safety scientists navigate through their data. Figure 4 displays different nodes that represent data from signals and SMIS/IFCS records regardless of the internal structure of the data source.

Domain ontologies can be straightforwardly implemented in the database as a data model to integrate data. These ontologies support the analysis of data nodes that use different semantics about a railway domain. This semantic alignment allows to interconnect data nodes each other or connect them to specific events related to safety management. However, ontologies are far from being populated automatically and require safety expertise and human effort to build them (Figueres-Esteban & Van Gulijk 2016). Table 1 and Table 2 shows the results of this effort in order to align three different data sources with a railway standard in a single ontology that represent the signal domain (Figure 3).

The alignment of different types of data with a domain ontology and events related to safety has important benefits. Firstly, the ontology provides framework in order to query signals. In this case, the standard RS/521 was selected as reference framework. Secondly, the integration of data allows to connect all the information available in the data sources. For example, linking data nodes of signals to the signal obscure event and SMIS/IFCS records allows to filter records by specific types of signals in order to improve the safety understanding related to obscured signals.

8 CONCLUSIONS

This paper demonstrates how safety scientists can enter the realm of big data. It demonstrates that the challenge of storing large amounts of data from diverse railway data sources to extract safety learning requires safety experts that can work with graph databases.

Graph databases allow to store data regardless of the structure of the data. But more fundamentally, it allows the co-location of domain ontologies to integrate different data sources and extract safety learning.

In theory, the database is infinitely scalable so it is hard to predict the limitations of the approach.

9 REFERENCES

Corcho, O., Fernandez-Lopez, M. & Gomez-Perez, A., 2003. Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data & Knowledge Engineering*, 46, pp.41–64.

Figueres-Esteban, M. & Van Gulijk, C., 2016. *Ontology learning for BDRA. Report 110-124 II*, Huddersfield.

Figueres-Esteban, M., Van Gulijk, C. & Hughes, P., 2015. *Visualisation and Risk Communication in Railway Big Data Risk Analysis (BDRA): Literature Review. Report 110-113*, Huddersfield.

Figueres-Esteban, M., Hughes, P. & Van Gulijk, C., 2016. Ontology network analysis for safety learning in the railway domain. In L. Walls, M. Revie, & T. Bedford, eds. *Risk, Reliability and Safety: Innovating Theory and Practice*. London: Taylor & Francis Group, pp. 2937–2942.

Figueres-Esteban, M., Hughes, P. & Van Gulijk, C., 2015. The role of data visualization in railway Big Data Risk Analysis. In *Safety and Reliability of Complex Engineered Systems - Proceedings of the 25th European Safety and Reliability Conference, ESREL 2015*.

Gibson, W.H. et al., 2015. The incident factor classification system and signals passed at danger. In *Fifth International Rail Human Factors Conference*. RSSB, pp. 22–31.

Gruber, T.R., 1995. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43, pp.907–928. Available at: http://dx.doi.org/10.1006/ijhc.1995.1081.

Guarino, N., 1997. Understanding, building and using ontologies. *International Journal of Human Computer Studies*, 46, pp.293–310.

Van Gulijk, C. et al., 2018. Introduction to IT transformation of safety and risk management systems. In *Handbook of RAMS in railway systems: Theory and Practice*. CRC.

Van Gulijk, C. et al., 2017. The case for IT transformation and Big Data for Safety Risk Management on the GB railways. *Journal of Risk and Reliability*.

Van Gulijk, C. & Figueres-Esteban, M., 2016. *Background of Ontology for BDRA. Report 110-124 I*, Huddersfield.

Van Gulijk, C., Hughes, P. & Figueres-Esteban, M., 2016. The potential of ontologies for safety and risk analysis. In L. Walls, M. Revie, & T. Bedford, eds. *Risk, Reliability and Safety: Innovating Theory and Practice*. London: Taylor & Francis Group, pp. 1315–1322.

EL Rashidy, R. et al., 2017. A Big Data modeling approach with graph databases for SPAD risk. *Safety Science*.

Robinson, I., Webber, J. & Eifrem, E., 2013. *Graph Databases*, California, United States of America: O'Reilly Media, Inc.

RSSB, 2017. Safety Management Intelligence System (SMIS). Available at: https://www.rssb.co.uk/risk-analysis-and-safety-reporting/reporting-systems/smis.

RSSB, 2015. Signals, handsignals, indicators and signs. Handbook. RS/521 Issue 3. , p.60.

Sadalage, P.J. & Fowler, M., 2013. *NoSQL Distilled: A brief guide to the emerging word of plyglot persistence*, New Jersey: Pearson Education, Inc.