University of Huddersfield Repository

Figueres-Esteban, Miguel, Hughes, Peter, EL Rashidy, Rawia Ahmed Hassan and Van Gulijk, Coen

Integrating data to support SPAD management

Original Citation

Figueres-Esteban, Miguel, Hughes, Peter, EL Rashidy, Rawia Ahmed Hassan and Van Gulijk, Coen (2017) Integrating data to support SPAD management. In: Sixth International Human Factors Rail Conference, 6-9 November 2017, London. (Unpublished)

This version is available at http://eprints.hud.ac.uk/id/eprint/33962/

Integrating data to support SPAD management

Miguel FIGUERES-ESTEBAN, Peter HUGHES, Rawia EL RASHIDY, Coen VAN GULIJK

University of Huddersfield, UK

Abstract. A stop signal passed without authority (aka SPAD) is one of the most serious types of incidents in railways since they potentially cause derailments or collisions. SPADs are complex incidents that have been usually analysed as human factors incidents. Human errors of train drivers such as slips or lapses have been prevalent in SPAD incident investigations. In the big data era, alternatives to the traditional methods can be used to support SPAD analysis of whatever kind. Railway systems produce a huge amount of data from a variety of data sources that can be used to get a better understanding of the factors involved in SPADs. This paper describes a first trial within the Big Data Risk Analysis program (BDRA) in order to combine unstructured data from SMIS/IFCS text records with structured data from of railway signals in order to support the SPAD management.

Keywords. Big Data, Risk Analysis, SPAD analysis, Data modelling

1. Introduction

The Office of Rail and Road (ORR) defines a signal passed at danger (SPAD) "when a train passes a stop signal without authority to do so" (Office of Rail and Road, 2017). Historically, SPADs have been involved in serious accidents and incidents such as Landbroke Grove disaster in 1999 that caused 31 deaths and more than 400 injuries (Stanton & Walker, 2011).

The railway industry has usually mitigated SPAD events adopting engineering solutions embedded in train-signalling systems or taking countermeasures from the investigation of human factors involved in the causes of SPAD (Filtness & Naweed, 2017; Naweed, 2013). However, despite the abundant technical and human factors research, the SPAD problem remains enigmatic due to its inherent complexity.

In the last decade, the GB railways have made considerable efforts to define a methodology to classify SPADs (RSSB, 2008) and improve SPAD management from the human factors perspective (RSSB, 2016). These efforts have produced different "data lakes" related to SPAD incidents such as semi-structured records from the Safety Management Intelligence System (SMIS) or bespoke human factors databases such the Incident Factor Classification System (IFCS).

In era of Big Data, even more data can be added to the data lakes. Technical railway systems are also producing a huge amounts of data that may be tapped into them (Van Gulijk, Hughes,

Figueres-Esteban, Dacre, & Harrison, 2015; Van Gulijk, Hughes, Figueres-Esteban, EL Rashidy, & Bearfield, 2017). The combination of these data sources with the SPAD data might provide a better understanding of underlying factors involved in SPADs, opening new opportunities to improve the SPAD management. But the question is how? This paper describes a Big Data approach to integrate SPAD data from SMIS and IFCS database with a railway signals database in order to support the SPAD management.

2. Method

2.1 Data sources

This trial uses four *.csv files from three data sources: SMIS, IFCS database and a database 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 Network Rail 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 SPADs. 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 sub-categories (Gibson, Mills, Basacik, & Harrison, 2015). The signal database is a sample of descriptions of signals that is part of the Ellipse Asset Management tool of Network Rail. Figure 1 shows the properties that were used to integrate the data.
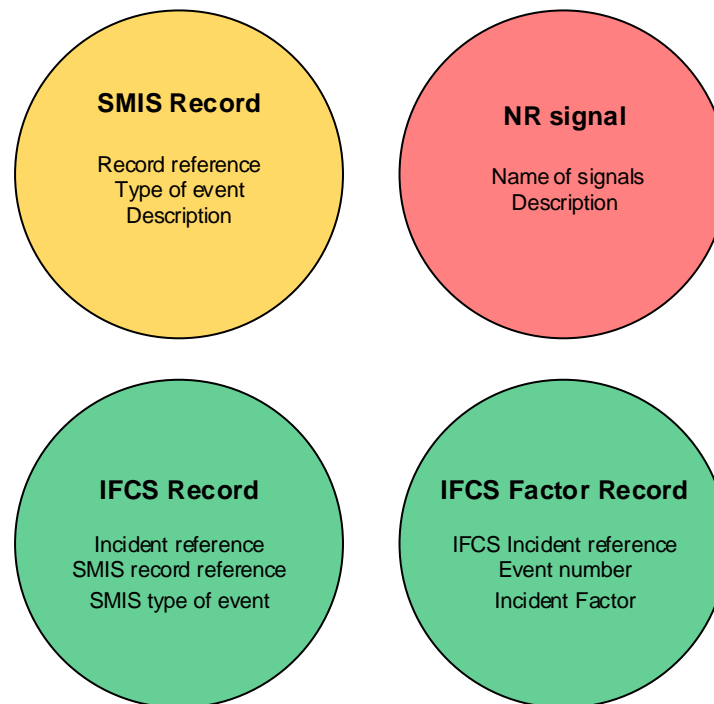

Figure 1. Description of the data sources used to support SPAD management.

2.2 Data model

The integration of the data has been implemented in a NoSQL graph database. Graph databases store data in the form of nodes and relationships. Nodes represent a piece of information and links the relationships between those different pieces of information. The signal data has been integrated with a signal taxonomy that maps the standard name of signals with its different properties such the type of the signal. This taxonomy has been built based on the standard RS/521 "Signals, hand-signals, indictors and signs handbook". Figure 2 shows the data model that describes the relationships between the different data sources and the link with the signal taxonomy.
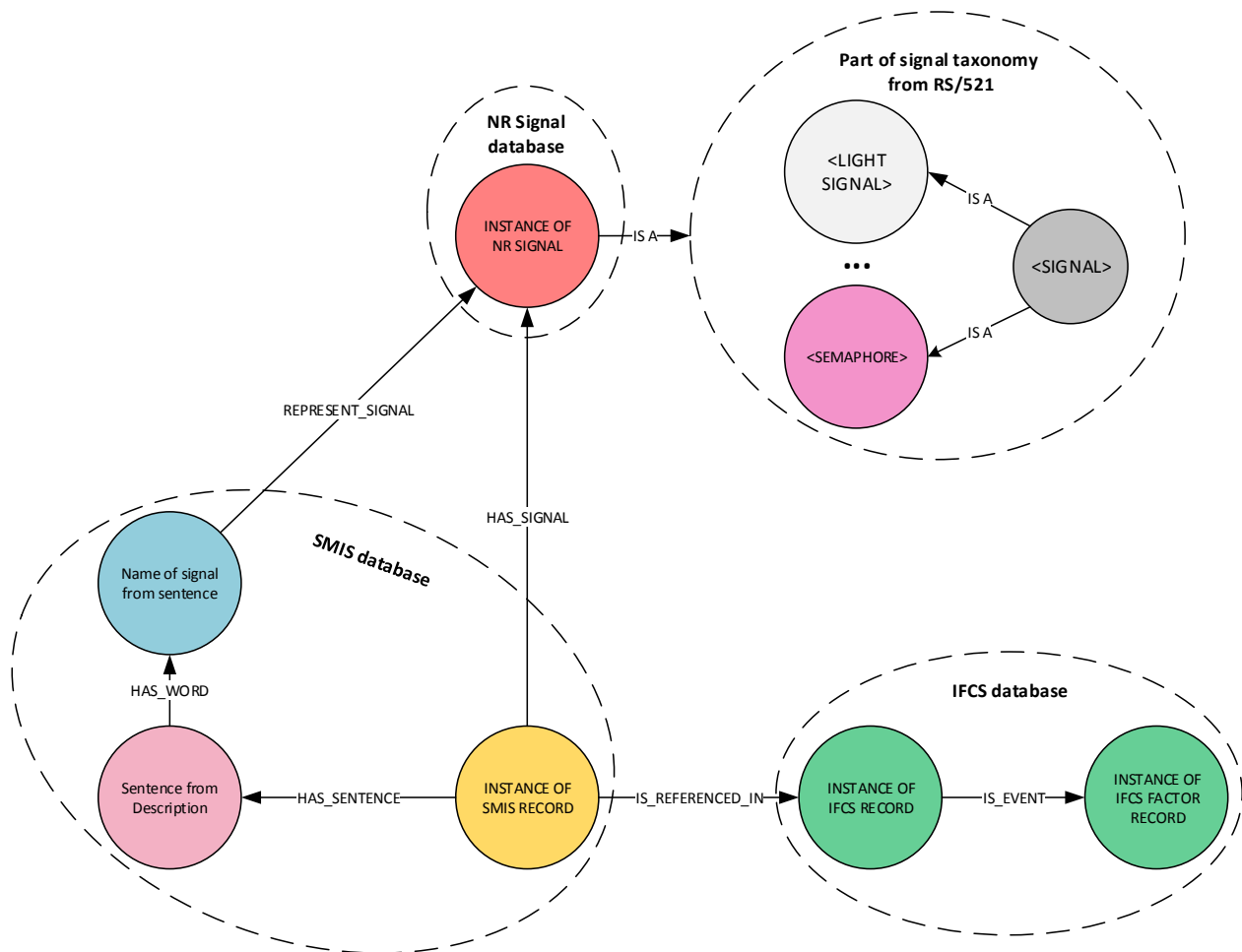

Figure 2. Data model that integrates the different data sources into a graph database.

A SMIS record contains semi-structured data. The signal involved in the SPAD has been extracted from the text description of the record. The text is broken down into sentences and words. The words that represent a signal are identified and linked to the pertinent instance of a signal. In our model, the IFCS database is divided into two different nodes, the instances of IFCS records and the nodes that represent the different incident factors. The IFCS record is linked to the related SMIS record.

3. Results

Figure 3 demonstrates the linkage inside the database for a single SPAD event.
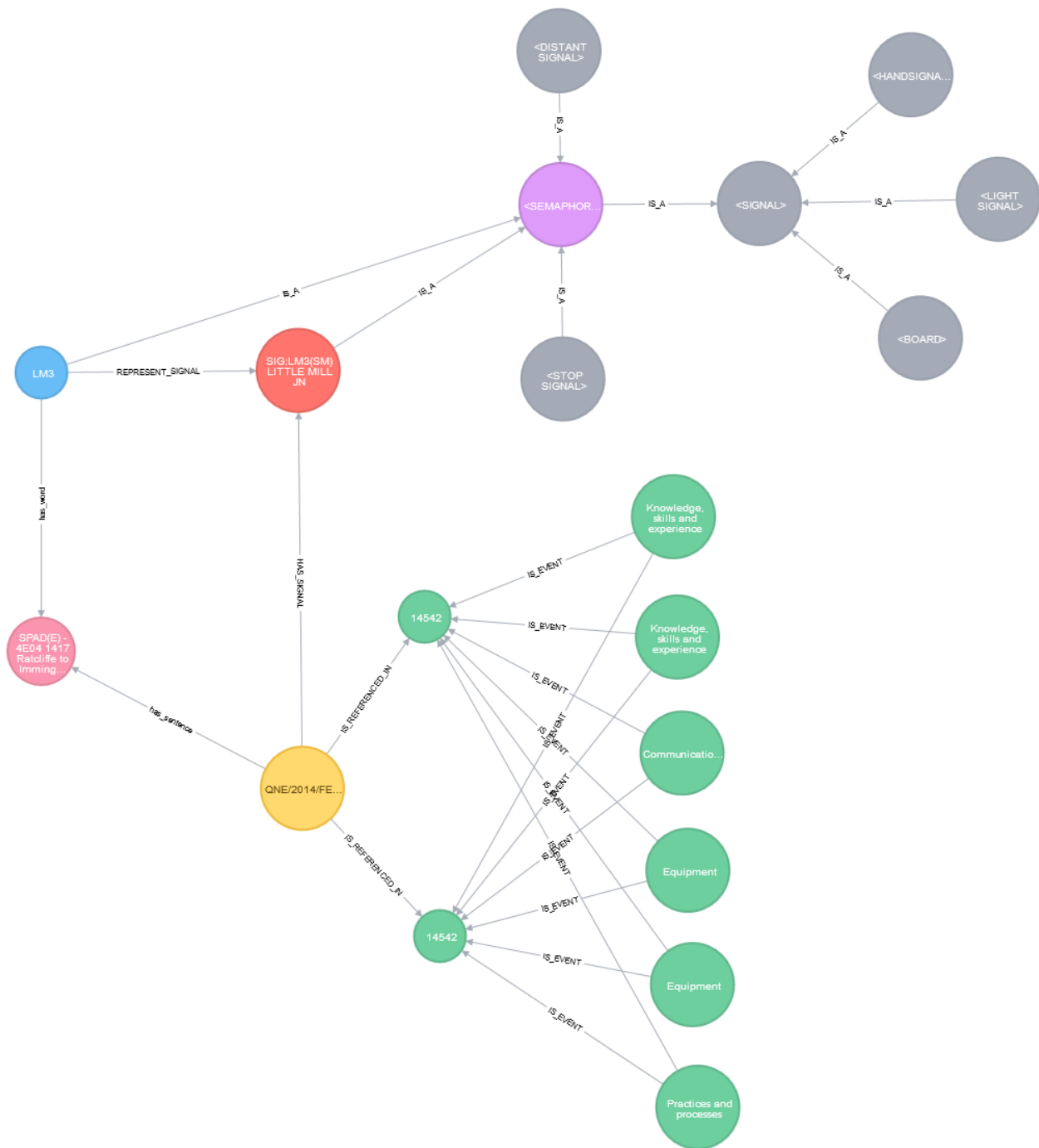

Figure 3. Part of the database that integrates three different types of database (NR signals, SMIS database and IFCS database) with the signal taxonomy of the standard RS 521.

The nodes with the label between angle brackets such as <SIGNAL>, <LIGHT SIGNAL> or <SEMAPHORE> represent concepts of signals from the standard RS 521 in the form of a signal taxonomy. The red node with the label "SIG:LM3(SM) LITTLE MILL JN" represents an instance of a railway signal (a particular railway signal that is placed in the network) that is linked to a node that represents a type of signal of the signal taxonomy. The orange node with the label "QNE/2014/…" represents a SMIS record about an SPAD incident. It is linked to an instance of signal (LM3 signal) since it represents the signal that was passed in danger. The nodes with the label "SPAD(E)-4E04…" and "LM3" represent the sentence of the SPAD description that includes the name of the signal and the name of signal respectively. The SMIS record and the node with the name of the signal are linked to the corresponding signal. Finally, the green nodes represent IFCS records. The nodes with the label "14542" represent the IFCS record number. There are two nodes because this particular SPAD instance is classified in two different categories, one of them SPAD. The other nodes with the labels "Knowledge, skills and experience", "Communications", "Equipment" and "Practices and processes" represent the different events that are involved in the record and its related incident factor.

4. Discussion and Conclusion

The digital transformation of railways is providing a massive quantity of data that opens new ways to support safety management systems including human factors. However, this amount of data cannot be managed using the traditional data analysis based on small samples of data that usually comes from interviews, observations or surveys. Big data techniques are necessary to extract some insight from huge sets of data. Because there is no direct interaction with data, a data model has to be built in order to conduct the data analysis (Grolemund & Wickham, 2014).

We have proposed a data model (Figure 2) in a database to integrate all the available data from the SMIS database, a database with all the railway signals from the British network and the IFCS database. This model has been developed for SPAD incidents. At this point, only a few properties have been used to build the database such as text descriptions of records and names of signals and reference numbers. A taxonomy of signals from the standard RS 521 has been implemented to map the different ways of naming signals into standard concepts such as <SEMAPHORE> or <LIGHT SIGNAL>.

The added value of building this data model is that it is possible to conduct data analysis without interacting with the raw data. The graph database integrates around half a million of nodes of data from the three databases and provides the semantics of a standard to avoid ambiguity with the names of the signals. That means that all the properties of the databases can be shared or queried for search of patters that might support the SPAD management. For example, the SMIS and IFCS records describe the SPAD incident and the signal that was passed, but there is no additional information about that signal. The integration of the SMIS and IFSC databases with a signal database allows to add new information to SPAD incidents such as the type of signal according to the RS 521 standard and its technical aspects (e.g. mechanical signal or LED technology).

This trial is just a small proof of concept of how human factors experts can take advantage of new techniques that come with the digital transformation of the railways. We believe it is possible to deal with all the available data, both in volume and variety, for defining the right semantic models

to analyse SPADS. This transformation opens new possibilities to get a richer insight in human factors by means of the link of the multiple huge data that produce the railway systems.

Acknowledgement

References

Filtness, A. J., & Naweed, A. (2017). Causes, consequences and countermeasures to driver fatigue in the rail industry: The train driver perspective. *Applied Ergonomics*, *60*, 12–21.

Gibson, W. H., Mills, A., Basacik, D., & Harrison, C. (2015). The incident factor classification system and signals passed at danger. *Fifth International Rail Human Factors Conference* (pp. 22–31). RSSB.

Grolemund, G., & Wickham, H. (2014). A Cognitive Interpretation of Data Analysis. *International Statistical Review*, *82*(2), 184–204.

Van Gulijk, C., Hughes, P., Figueres-Esteban, M., Dacre, M., & Harrison, C. (2015). Big Data Risk Analysis for Rail Safety? In L. Podofillini, B. Sudret, B. Stojadinovic, E. Zio, & W. Kroger (Eds.), *Safety and Reliability of Complex Engineered Systems* (pp. 643–650). London: Taylor & Francis Group.

Van Gulijk, C., Hughes, P., Figueres-Esteban, M., EL Rashidy, R., & Bearfield, G. (2017). The case for IT transformation and Big Data for Safety Risk Management on the GB railways. *Journal of Risk and Reliability*.

Naweed, A. (2013). Psychological factors for driver distraction and inattention in the Australian and New Zealand rail industry. *Accident Analysis & Prevention*, *60*, 193–204.

Office of Rail and Road. (2017). Signals passed at danger. Retrieved from http://www.orr.gov.uk/rail/health-and-safety/infrastructure-safety/signals-passed-at-danger

RSSB. (2008). SPAD Risk Ranking Methodology Handbook.

RSSB. (2016). How to manage SPAD risk better. A guide for directors.

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

Stanton, N. A., & Walker, G. H. (2011). Exploring the psychological factors involved in the Ladbroke Grove rail accident. *Accident Analysis & Prevention*, *43*(3), 1117–1127.