CIT2016 – XII Congreso de Ingeniería del Transporte
València, Universitat Politècnica de València, 2016.
DOI: http://dx.doi.org/10.4995/CIT2016.2016.1825

Big Data for Risk Analysis: the future of safe railways

Miguel Figueres-Esteban
Research Fellow, University of Huddersfield, UK
Peter Hughes
Principal Enterprise Fellow, University of Huddersfield, UK
Coen van Gulijk
Reader, University of Huddersfield, UK

ABSTRACT

New technology brings ever more data to support decision-making for intelligent transport systems. Big Data is no longer a futuristic challenge, it is happening right now: modern railway systems have countless sources of data providing a massive quantity of diverse information on every aspect of operations such as train position and speed, brake applications, passenger numbers, status of the signaling system or reported incidents.
The traditional approaches to safety management on the railways have relied on static data sources to populate traditional safety tools such as bow-tie models and fault trees. The Big Data Risk Analysis (BDRA) program for Railways at the University of Huddersfield is investigating how the many Big Data sources from the railway can be combined in a meaningful way to provide a better understanding about the GB railway systems and the environment within which they operate.
Moving to BDRA is not simply a matter of scaling-up existing analysis techniques. BDRA has to coordinate and combine a wide range of sources with different types of data and accuracy, and that is not straight-forward. BDRA is structured around three components: data, ontology and visualisation. Each of these components is critical to support the overall framework. This paper describes how these three components are used to get safety knowledge from two data sources by means of ontologies from text documents. This is a part of the ongoing BDRA research that is looking at integrating many large and varied data sources to support railway safety and decision-makers.

1. INTRODUCTION

Over the past decades, safety science has developed safety risk models based on static data from failures of systems, causes of accidents, incidents or near misses to discover hazardous situations (Figure 1). Railways have embraced many of these models in order to support safety management and decisions-making (e.g. Heinrich's Law, Reason's Swiss Cheese Model and Bow-Tie models). In the GB railways, the Safety Risk Model (SRM) has been developed based on a set of fault and event trees for hazardous events (Marsh and Bearfield, 2008), which come from a combination of expert judgement and incident data from the rail industry's accident reporting system (Safety Management Information System, SMIS).

DOI: http://dx.doi.org/10.4995/CIT2016.2016.1825

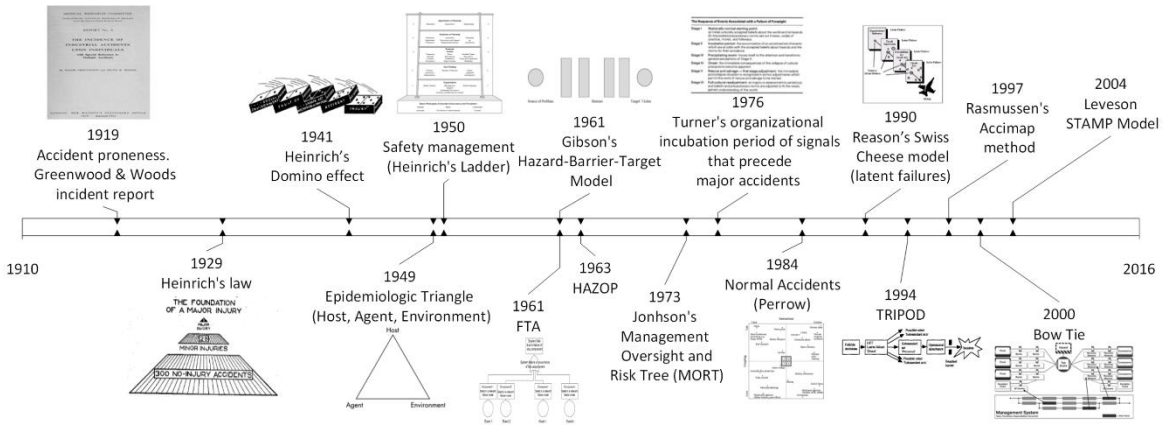

Figure 1. Evolution of safety risk models.

New technological railway systems are producing massive amounts of data which could be related to safety. The GB railways is, for example, dealing with how to manage complex rail assets by means of the ORBIS and the Future Railway programs (NR, 2016; RSSB, 2016). These ambitious programmes bring a cultural change in the railway organisations to improve the acquisition, storage and use of asset information for a truly digitised railway. Big Data analytics techniques based on these complex data could support safety management by providing valuable information on the precursors of hazardous events. The aim of the Big Data Risk Analysis (BDRA) programme at the University of Huddersfield is to investigate whether Big Data processing techniques can support the current railway risk models and safety decision-making and change traditional risk analysis, and if so, how (Van Gulijk et al., 2015).

Our research to date has identified three basic components of a BDRA system: data, ontology and visualisation.

Data is the basis of BDRA. Converting and integrating various types of data from heterogeneous information systems is the starting point Big Data analysis. Combining these data sources can provide a powerful insight into how the operation and failure of equipment contribute to failures and accidents on the railway. An example of data sources are the Close Call (CC) records, Train Describer (TD) Messages or On-Train Monitoring and Recording (OTMR) data.

Ontology is “an explicit specification of a conceptualization” (Gruber, 1993). It is a representation of a domain knowledge that supports the use of different databases in a meaningful way and enables data analysis: it can be compared to a search engine which holds the right search keys to produce results that can be interpreted or analysed by the human operator. In this case the domain is safety and risk management for the GB railways, the concepts are the ways in which the components within the domain combine and interact to create the emergent behavior of the overall system (Van Gulijk and Figueres-Esteban, 2016).

Visualisation is a powerful, and arguably the only useful, tool for understanding large quantities of data by humans. Hundreds of different visualisation techniques are available to provide visual output of data. However, visualisation is not just for representing results. Modern visual analytics tools also provide the ability to interact with the data to perform

DOI: http://dx.doi.org/10.4995/CIT2016.2016.1825

data analysis (Figueres-Esteban et al., 2015a, 2015c).

2. CASE STUDY DESCRIPTION

This section provides an overview of the key components being used in the development of a trial for practical implementation of BDRA. The aim of this trial is to extract information from the Close Call database and Rail Notices related to *Possession events*, that is, Close Call events and notices related to the arrangements of planned blockages of a line section for infrastructure engineering works.

In the GB railways, the Rule Book is the safety document that describes operational rules for railway staff to enable safe engineering works, and therefore it describes the operational rules related to possessions. This safety document is complemented by numerous standards from the infrastructure manager (Network Rail) which also comply with the requirements of European regulations (e.g. Technical Specification for Interoperability relating to the 'operation and traffic management'). The details of the possession's arrangements are published and described in the Rail Notices system. The case study attempts to extract information from the Close Call database using the knowledge domain described in the documents and standards. Figure 2 shows the BDRA components involved in the case study.

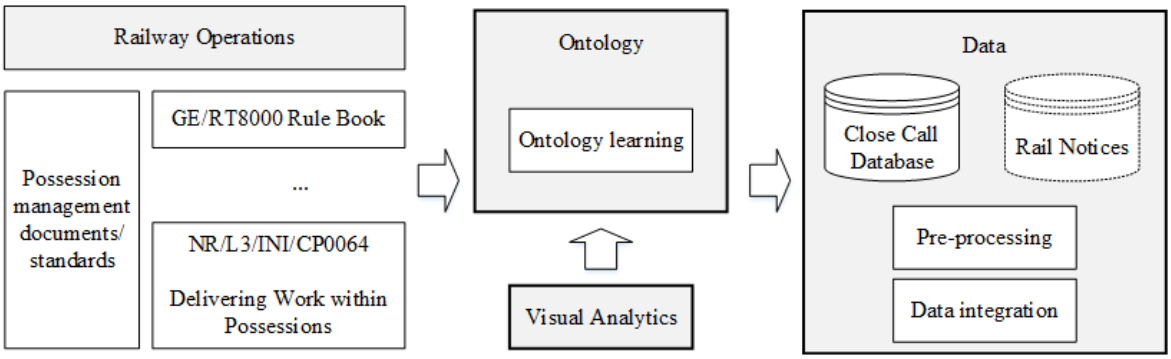

Figure 2. Basic components of BDRA in the Possession case study.

2.1. Data

Railway Close Calls are free-text descriptions of hazards provided by railway workers. The non-numeric data is an example of the variety of BDRA. Where standard numerical processing techniques can be used for data sources such as TD or OTMR data, these are not available for CC data. Close Call records are provided through a number of interfaces: entered directly into a mobile app by a worker, through a web interface, or typed by a telephonist who receives a call from a railway worker. The Close Call System collects about 150,000 text-based records each year on potentially dangerous events, including events related to possession arrangements.

A potential complementary semi-structured data source is the Rail Notices system. It is a web-based application which allows users to search, view, create and manage notices and alerts relating to a variety of rail industry processes, in particular, related to Weekly Operating Notices that describe possession arrangements. This information is provided by means of a combination of free-text descriptions and structured data.

DOI: http://dx.doi.org/10.4995/CIT2016.2016.1825

The challenge for BDRA is to uncover how this free-text information can be understood and integrated in a way that is meaningful within a railway safety context. Natural Language Processing (NLP) and data integration techniques based on ontologies are used to extract safety information of railway operations related to possessions.

2.2. Ontology

A key component of this case study is the development of an ontology that represents the knowledge domain of possession management to support the text analysis and data integration from Close Call and Railway Notices. In this case it is not possible to reuse other ontologies because of the singularity of the topic, so it is necessary to build the ontology from scratch. The process to build an ontology is also called Ontology Learning. There are different approaches to build ontologies (Corcho et al., 2003). In this case study is used a corpus-based methodology. This method helps to acquire ontologies from a set of documents written in natural language that explain the knowledge of the domain. This approach is a semi-automatic method that implies a human interactive terminology extraction and natural language processing. Moreover, this method allows population of a lexicon related to the meaning of the extracted concepts. At this stage it is possible to represent a low-level ontology in a UML sequence or class diagram (Figure 3) that support the development of the formal ontology. A formal ontology describes the concepts, the different types of relationships and the axioms of the domain. This formal ontology has to be represented in a knowledge representation language for further use.

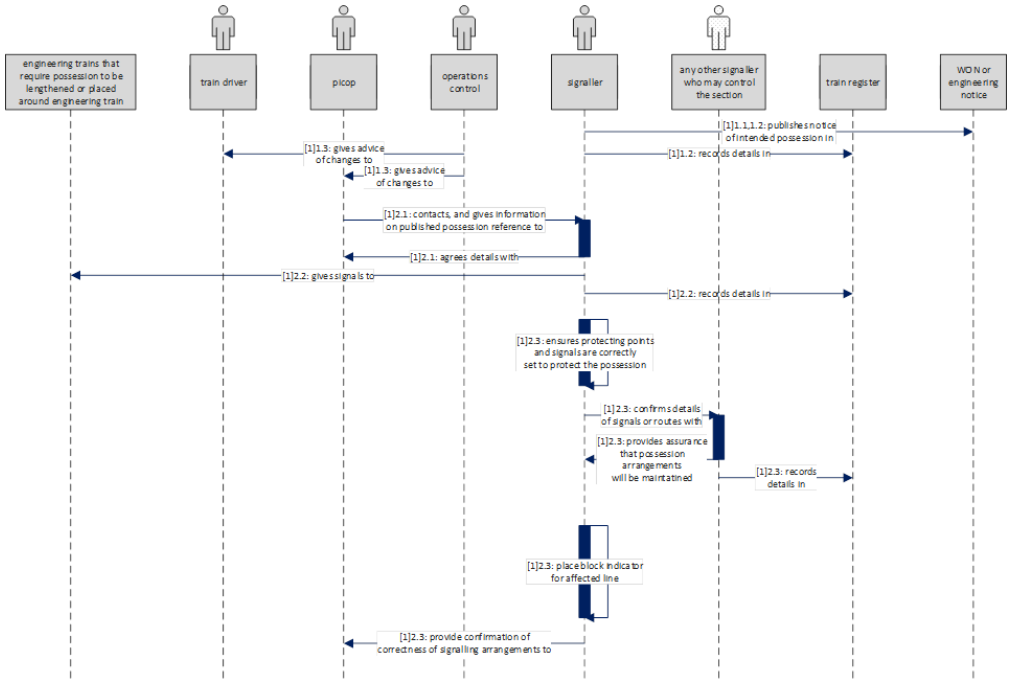

Figure 3. Part of a UML sequence diagram of the Rule Book, issue 6: Possession of a running line for engineering work.

DOI: http://dx.doi.org/10.4995/CIT2016.2016.1825

2.3. Visualisation

Visual Analytics can support text analysis by means of network text analysis (Figueres-Esteban et al., 2015b). In this trial, we use a directed graph of words with part-of-speech tag attributes to support the term extraction in the Ontology Learning process. Figure 4 shows how it is possible to detect multi-word terms, for example detonator protection or block marker, and key concepts from the Rule Book related to possessions (e.g. PICOP or signaller).

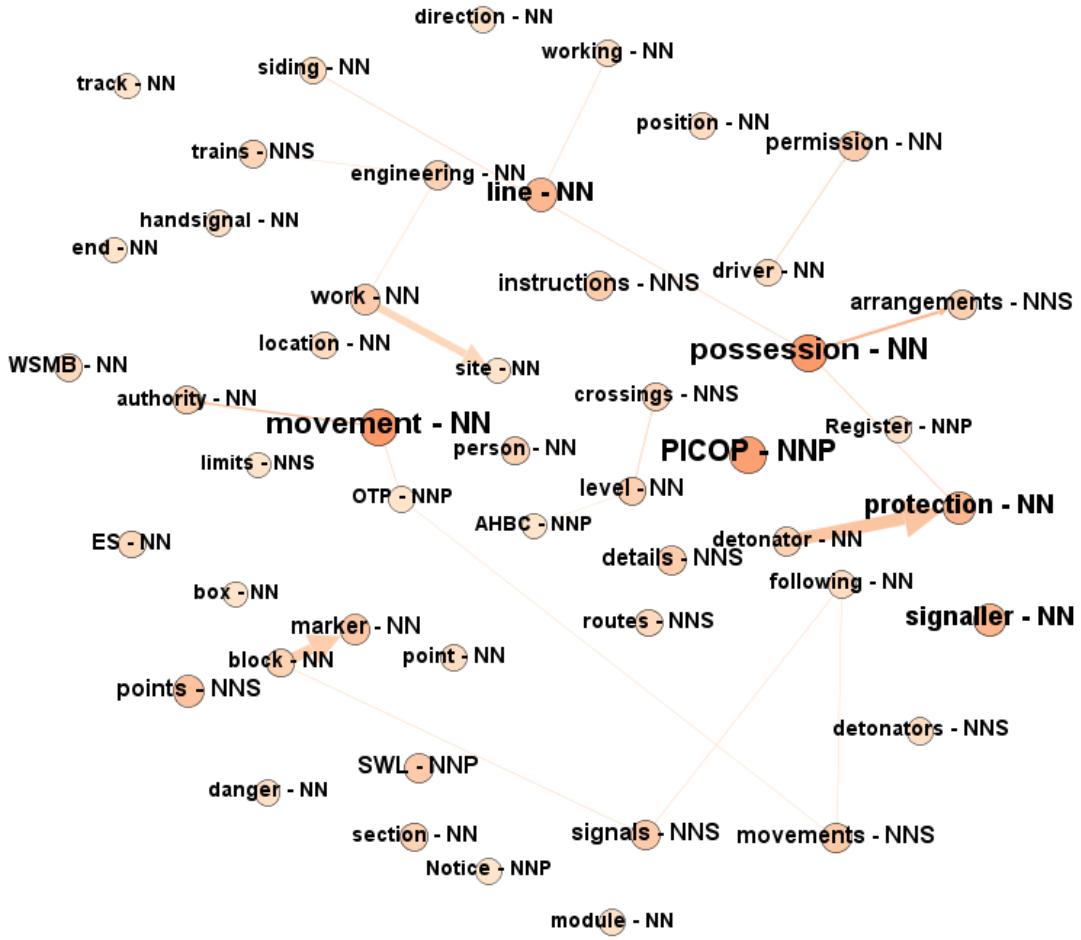

Figure 4. Network of Nouns and Proper Nouns (NN, NP, NNS and NNPS) of the Rule Book, issue 6: Possession of a running line for engineering work. Filtered graph by degree ≥ 5.

3. CONCLUSION AND NEXT STEPS

The BDRA programme aims to investigate how data from large and heterogeneous data sources can be tamed to provide meaningful information about railway safety. This paper describes the key elements of BDRA for practical implementation of a case study: getting insight from data about the railway operations related to possession. Close Call database and Rail Notices are potential data sources to extract safety information about possession events.

DOI: http://dx.doi.org/10.4995/CIT2016.2016.1825

This paper demonstrates the extraction of terms from the Rule Book to create an ontology to represent the domain knowledge of possession. Visual Analytics is supporting the Ontology Learning process for term extraction from text, as well as the creation of a lexicon linked to the concepts of the ontology.

The next stages of our research will construct a formal ontology and represent the ontology in a knowledge representation language. The ontology will be used for the natural language interpretation of hazards related to possession railway operations that are reported in the Close Call System. Moreover, it will allow to integrate the information from Close Call and Rail Notices to improve the railway operations related to possession.

ACKNOLEDGEMENTS

The work reported in this paper was undertaken under the Strategic Partnership between the University of Huddersfield and RSSB.

REFERENCES

Corcho, O., Fernandez-Lopez, M., Gomez-Perez, A., 2003. Methodologies, tools and languages for building ontologies. Where is their meeting point? Data Knowl. Eng. 46, 41–64. doi:10.1016/S0169-023X(02)00195-7

Figueres-Esteban, M., Hughes, P., Van Gulijk, C., 2015a. The role of data visualization in Railway Big Data Risk Analysis, in: *Proceedings of the 25th European Safety and Reliability Conference, ESREL 2015*.

Figueres-Esteban, M., Hughes, P., Van Gulijk, C., 2015b. Visualising Close Call in railways: a step towards Big Data Risk Analysis, in: *Proceedings of the Fifth International Rail Human Factors Conference.*

Figueres-Esteban, M., Van Gulijk, C., Hughes, P., 2015c. Visualisation and Risk Communication in Railway Big Data Risk Analysis (BDRA): Literature Review.

Gruber, T.R., 1993. A Translation Approach to Portable Ontology Specifications, in: Knowledge Creation Diffusion Utilization. pp. 199–220. doi:http://dx.doi.org/10.1006/knac.1993.1008

Marsh, D.W.R., Bearfield, G., 2008. Generalizing event trees using Bayesian networks. Proc. Inst. Mech. Eng. Part O J. Risk Reliab. 222, 105–114. doi:10.1243/1748006XJRR131

NR, 2016. NR 2016 [WWW Document]. URL http://www.networkrail.co.uk/aspx/12210.aspx (accessed 2.8.16).

RSSB, 2016. RSSB 2016 [WWW Document]. URL http://www.rssb.co.uk/future-railway-programme (accessed 2.8.16).

Van Gulijk, C., Figueres-Esteban, M., 2016. Background of Ontology for BDRA.

DOI: http://dx.doi.org/10.4995/CIT2016.2016.1825

Van Gulijk, C., Hughes, P., Figueres-Esteban, M., Dacre, M., Harrison, C., 2015. Big Data Risk Analysis for Rail Safety?, in: *Proceedings of the 25th European Safety and Reliability Conference, ESREL 2015*.