1. Article type: Special Issue ESREL
2. Corresponding author info

Corresponding Author:

Coen van Gulijk, University of Huddersfield, SCEN/IRR, Queensgate, Huddersfield, HD13DH, United Kingdom

Email: c.vangulijk@hud.ac.uk

3. The case for IT transformation and Big Data for Safety Risk Management on the GB railways
4. Authors

Coen van Gulijk[1], Peter Hughes[1], Miguel Figueres-Esteban[1], Rawia El-Rashidy[1] George Bearfield[2].

[1]University of Huddersfield, Huddersfield, UK

[2]RSSB, London, UK

5. Abstract

Abstract

This paper presents the case for IT transformation and big data for safety risk management on the GB railways. This paper explains why the interest in data driven safety solutions is very high in the railways by describing the drivers that shape risk management for the railways. A brief overview of research projects in the Big Data Risk Analysis (BDRA) programme supports the case and helps understand the research agenda for the transformation of safety and risk on the GB railways. The drivers and the projects provide insight in the current research needs for the transformation and explains why safety researchers have to broaden their skill set to include digital skills and potentially even programming. The case for IT transformation of risk management systems is compelling and the paper describes just the tip of the iceberg of opportunities opening up for safety analysis that, after all, depends on data.

Keywords

Railway risk, Risk analysis, IT transformation, BDRA, SMIS+

The case for IT transformation and Big Data for Safety Risk Management on the GB railways

Coen van Gulijk, Peter Hughes, Miguel Figueres-Esteban, George Bearfield

Abstract

This paper presents the case for IT transformation and big data for safety risk management on the GB railways. This paper explains why the interest in data driven safety solutions is very high in the railways by describing the drivers that shape risk management for the railways. A brief overview of research projects in the Big Data Risk Analysis (BDRA) programme supports the case and helps understand the research agenda for the transformation of safety and risk on the GB railways. The drivers and the projects provide insight in the current research needs for the transformation and explains why safety researchers have to broaden their skill set to include digital skills and potentially even programming. The case for IT transformation of risk management systems is compelling and the paper describes just the tip of the iceberg of opportunities opening up for safety analysis that, after all, depends on data.

Introduction

Computer scientists are clear in their belief that the data revolution is coming of age: there is a firm belief that the enormous amounts of data collected will inevitably lead to a revolution in how management will be undertaken in the future [1, 2, 3]. Yet, to date, the potential benefit of this revolution has barely been investigated for safety and risk management.

The term *big data* has been created to describe the methods and techniques that process and extract meaning from very large amounts of data; despite being very widely used there is no common definition of the term [4,5,6]. The broad interpretation is that big data systems economically extract value from huge volumes of a variety of data sources very quickly (leading to the *Three Vs* definition of big data: *volume*, *variety*, and *velocity*). A sceptical view is that this definition we can imagine that big data is simply another fad to describe a step forward in the evolution of management decision-support tools or Business Intelligence systems. In this work, however, we take a more optimistic view where big data creates opportunities for intelligent systems. In fact, the design of purpose built IT systems is not the principal concern in this paper, it is the development of a form of machine-assisted interpretation (perhaps resembling intelligence) in the form of a software layer that bridges the gap between data sources and the theoretical and practical mechanisms to deliver safety on the railways. This bypasses the need to specify what the

data should look like to support the safety system precisely and perfectly. Instead the extraction of safety relevant information and safety lessons is guided by theoretical and practical safety principles applied on existing data. Data that was not necessarily purposefully designed for safety analysis or safety systems. With this approach data drives safety rather than the other way around. This creates opportunities in the development of new safety solutions but it comes at a cost that some safety issues are difficult to capture with existing data streams or data may simply be absent.

This paper describes the approach that shapes risk management for the GB railways that inexorably leads to data-driven intelligent safety solutions and thereby explains the heightened interest in data-driven safety solutions. The drivers are explained by re-iterating common principles for managing railway safety and setting them against the greater global trends in IT and big data on the railways.

Drivers for IT transformation and BDRA

Common principles for railway safety

System safety management is the application of technical and managerial techniques to the systematic, forward-looking identification and control of hazards throughout the life-cycle of a project or activity. It calls for structured and rigorous identification and analysis of hazards; as well as the establishment of processes for change management, decision-making, implementation of risk controls, and on-going monitoring of safety.

Principle 1: Serious accidents are not tolerated

The prevention of serious accidents is a key focus of all railway partners. Accidents involving trains are disruptive, costly and cause human suffering; all of which damage the rail industry as a whole. Extensive media coverage on train accidents inevitably shapes public opinion regarding railway safety: in general the public abhors rail accidents but tends to be unwilling to fund safety investment (through fare or tax increases) or tolerate operational restrictions (such as speed restrictions). In very broad terms, public abhorrence influences the design of legislation requirements and the level of operational safety performance. The codification of public opinion and trade-off between law, safety performance and operational performance is unique to each country in the world and may vary widely from one country to another. In UK ‘societal concern’ is outside of the scope of legal responsibility for safety. Nevertheless, official bodies, such as the Office of Rail and Road and the Railway Accident Investigation Board have some degree of freedom when it comes to focussing attention to inspection and prosecution within the framework of legislative requirements.

Principle 2: Railway engineering for safety is of high integrity

Through continuous and determined effort to improve safety, a very high standard has been set for the safety integrity of railway engineering; but this integrity comes at a high price, both in terms of the direct cost of equipment, but also the cost of safety

management systems to operate the railway. A challenge for any railway is to reduce these costs whilst still operating a safety and efficient railway. This challenge is often met with the application of advanced technological systems such as the ones described in this paper.

Principle 3: The railways must not become less safe than they currently are

The closer that rail safety moves towards absolute safety, the more difficult it becomes to achieve further improvements, or even maintain the status quo. Continuous development is constrained by increasing costs, technological limitations, and the fundamental need to keep railway traffic moving to support the economy at large. One way to keep ahead in this challenge is to employ research in safety and reliability management. The GB railways have consistently been the safest in Europe which makes it especially prone to a reversal of the safety record and drives it to the edge of technology for even better safety performance.

IT trends for railway safety

In many ways, the extensive use of data systems in business is not a novelty. However, with increasing maturity of business intelligence systems, comprehensive digitized safety management systems are becoming more common.

Trend 1: The global digitalization trend

The global effort for IT transformation of business is hard to capture in a few words. It is served by a huge academic society and a massive industry. Areas of attention include: hardware development; algorithms to improve data storage and access; algorithms to reduce processing time; novel concepts for high-performance computing; and elaborate enterprise software techniques. The added value of these areas is that IT systems become better at supporting businesses [1, 7, 8, 9]. Academic domains for creating solutions include research in software systems [e.g. 10], ontologies [11], artificial intelligence [12], and business process modelling [13]. Big data is a label for the global digitalization trend that powers global business change and drives the digitalization of railways and railway safety. The label may change over time, but the global effort for the development of digital systems does not.

These developments are entering the GB railways and offer substantial advantages to the industry as a whole. The TSLG [14] report addresses the ambitions of the GB railways to embark on this journey of technological advancement in IT.

Trend 2: Datafication of railways

Big data project are appearing in the railways. [15, 16, 17, 18]. A particular area of interest is asset management with RFID systems [19, 20, 21, 22]. Although the work contain some references to safe operation, the do not deal with safety or risk management. Another data-hungry domain is condition monitoring: detectors attached to trains and rails are collecting and analysing huge amounts of data that would have been unmanageable until relatively recent advances in technology [23, 24].

European research projects such as “Intelligent integration of railway systems” (INTEGRAIL) and “Automated and cost effective railway infrastructure maintenance” (ACEM-rail) pave the way for the datafication of the railways [25, 26]. The direction that these projects take are further supported by underlying work that provides the data modelling tools required to manage big data projects for infrastructure such as the RailML framework that was developed by the UIC [27]. The railways embrace the digital transformation as an opportunity for improvement. Investigations into driverless trains, remote condition monitoring, digital ticketing and intermodal data interchange to support freight transport between ports and the railways are performed by railway organizations around the world. European infrastructure managers are making data-live feeds available to support this effort [28, 29, 30, 31].

Within the GB railways big data activities are underway, such as the ORBIS project in Network Rail [32] and rail condition monitoring [24].

Trend 3: Increased use of data driven safety risk controls

The traditional methods of safety risk management include well-tested techniques such as HAZOP, FMECA and fault tree analysis. Traditionally, individual risk models are fed by dedicated databases that provide cleansed data in a dedicated format for use in the risk model. These databases can be very large; for example, the SMIS database that is used to populate the Safety Risk Model for the GB railways contains more than two million records. Such databases are required to support risk-based decision making based on evidence.

But it is not just risk management systems that are evidence based. RAMS analyses form the backbone for the approval of technical systems on the railway (*viz.* EN50126, EN50128 & EN50129). When it comes to the implementation of advanced engineering systems, advanced tools help deal with the complexity of technical systems and safety-critical software systems but tend to be data-intensive. Examples include: Hip-Hops [33]; advanced safety case management [34]; software verification [35, 36, 37]. This development has inspired the development of advanced computer languages to represent risk scenarios [38]. Railway traffic management is a particularly challenging area and today’s signalling systems are based on fully integrated technical systems hosted by data-centres.

In the UK, the digital railway programme paces the digital transformation forward using ERMTS as a driver for the digital transformation of the railways [39, 40].

Trend 4: IT supported safety management systems

The Safety Risk Model [41,42] estimates risk from the full range of ‘hazardous events’ that might arise on the GB railways by estimating the frequency of these events, the likelihood of their various potential consequences and the severity of these consequences. Train accidents, and other high severity, low frequency events are estimated using detailed fault tree and event tree models, which were developed to provide a structured representation of the cause and consequences of potential accidents arising from the

operation and maintenance of the railway. The model is fed with data from the Safety Management Information System that contains more than two million data points. The SRM is supported by individual risk models that have been developed by RSSB and other railway companies in the UK for specific railway risks such as the Signal Overrun Risk Assessment Tool (SORAT) which analyses collision risk related to signalling layout design, and the All Level Crossing Risk Model, which looks at the risk from level crossings. It is the continued improvement of the SRM and other risk models that steers toward big data safety management systems. Thereby changing from IT supported systems to systems where IT provides interpretation and intelligence.

Trend 5: Big Data for safety science

An example of a big data safety system is the GeoSRM risk model to deal with localized risks [43]. The pilot model is based on the Safety Risk Model, but the fundamental difference between the two models is that the GeoSRM shows, on a map, how risk is distributed across the network, rather than generating a single national estimate for each type of event. A particular issue arose in the preparation of data: huge amounts of data were needed to populate the localised models, including not just safety incident data, but also timetable data and asset information. The GeoSRM pilot model has been proven with a subset of data dealing with risk profiles for derailment, suicides, and station slips, trips and falls for the 'Wessex' route in the South West of England.

Remember that we make a distinction between IT systems and data analytics in the sense that the latter depends on machine-assisted interpretation. The difference being that we do not specify what the data ought to look like for the purpose of safety management but we take the data as it is and construct interpretation algorithms to extract safety learning from it. A paper published in 2015 cautiously approaches the problem by focussing on the dangers of using big data and the potential security risks and risk associated with loss of data and limitations to interpretation of data [44] but a paper from that same year describes an analytic system to analyse human reliability [45]. The latter describes a computer based text interpretation engine to extract factors pertaining to training in nuclear power plants; an engine to populate a Bayesian belief net that is changed based on the findings of the interpretation and subsequently calculates the network to assess training quality on the nuclear site. A paper from the construction of a metro system in Wuhan describes how visual recordings and text-based commentaries are combined to detect and monitor unsafe behaviour [46]. An investigation in the UK extracted evidence from train data recorders to analyse the mental and physical demand on drives whilst they are driving trains [47]. Those researchers also created automated interpretation algorithms to extract safety relevant intelligence from a data-source that was not designed for that use. The final paper that relates big data and risk does not treat a safety solution but elaborates on developments for risk in relation to asset reliability and industrial systems reliability [48]. Though reliability is outside the scope of this paper, the overview provides three clear visions for risk research in the near future: technological advances in business intelligence (combining analysis methods and improve mining methods); system security and reliability (pertaining on the risks to the IT system itself); and advances in operational risk management, including the development of operational risk management frameworks.

Several safety software suppliers have embraced software solutions to deal with large data streams and harnessing potential for big data analytics. These solutions tend to use Bowties as their centre piece. In 2016, a paper by DNV describes the concepts for dynamic barrier management with an explicit reference to linking databases containing audits, barrier sensor data, control system data, incident data, maintenance records and personnel data to barrier monitoring systems [49]. Especially sensor data and control system data, which is not purposefully built for the safety control system, machine interpretation of mixed data sources is used. In 2016, CGE launched a cloud-based Bowtie solution to enable the big-data approach [50]. These developments have mostly come from the developments in the field of dynamic barrier management, an area that has received attention for a long time and also started incorporating machine interpretation e.g. [51].

The case for IT transformation of safety management

The case for IT transformation of safety management on the GB railways is made by the combination of the constant strive for safer railways and the relentless ingress of IT solutions in the railways and railway safety management. It is sensible to consider a systems approach to the IT transformation for safety management systems rather than working from individual technological solutions.

Safety activities must be integrated with all parts of the railway system for it to be effective and efficient [52]. A large part of the safety activities is delivered through safety management systems that aim to be holistic. This makes safety management an inclusive business process that, in principle, can be transformed into a big data business process.

This part of the paper demonstrated that the choice for investigating whether big data solutions that incorporate machine interpretation could benefit safety and risk for the railways in Britain. The arrival of big data management techniques for the railways provides a signpost to where risk management techniques may develop in the future. Yet the shape of things to come is unclear which leads to core research question for the BDRA research programme: how can big data techniques, exploiting machine interpretation of huge data sets be used most effectively to transform safety management systems for railway safety? The next part treats some projects in the BDRA research programme which shed some light into answering this question.

Introducing machine-assisted interpretation and intelligence: BDRA

The Big Data Risk Analysis (BDRA) research programme is a joint effort by RSSB and the Institute of Railway Research at the University of Huddersfield that investigates the potential for machine interpretation techniques for Safety. The objective is to investigate to what extent big data techniques, with a particular emphasis on machine-assisted interpretation, can support the current Safety Risk Model of RSSB and to investigate whether the modern data-analytics methods will change traditional risk analysis methods, and if so, how. The overview in the next two sections helps understand which alleys are investigated for the transformation of safety and risk on the GB railways. Six projects that are discussed, they were set up as relatively independent projects to investigate

different aspects of BDRA. Due to their differences, these projects provide a broad overview of the usefulness of big data to railway safety. For most of the projects, progress has been described in other papers so only a summary is repeated here. The projects discussed here are the following: OTDR based SPAD-Safety Indicator, RAATS, Learning from text-based Close Call records, visual analytics, ontology and SMIS+. The first project will be treated in more detail because it is not described elsewhere and it provides insight in skills that future safety analysts will have to embrace to perform safety data analyses.

A BDRA project: Leading SPAD safety indicator from OTDR

SPAD risks

A SPAD, a Signal Passed At Danger constitutes a serious breach of safety. When a signal is at danger, showing a red aspect, a train does not have the authority to proceed because the line is occupied. To some extent, it is similar to a red light in a road; it immediately puts the vehicle at risk from colliding into another one. In GB railway signalling, however, the driver has warning systems at their disposal. For the route considered in this work the warning signals work as follows. If the line ahead is clear, a signal shows a green aspect, indicating that it is safe to proceed. A double yellow indicates that the next signal is showing a yellow signal but the train may proceed as normal. A single yellow indicates that the driver has to prepare to stop at the next signal that could be red. A red signal indicates that stopping is obligatory. The driver is supported by a signalling system that is relevant for this discussion: AWS. An AWS horn is triggered with a magnet on the rails (typically at about 180 meters before the signal, depending on the line speed) that indicates that the next signal is a restrictive aspect (not green). The driver has 2.4 seconds to acknowledge the horn by pushing a button, if the button is not pressed, the train will brake automatically, thereby supporting the driver in preventing a SPAD.

OTDR data

On train data recorders (OTDR) are used to collect data from trains, to assess how they are driven and the state of various train systems during its journey. Examples of data collected include power and brake controller position, driver acknowledgement of signalling system warnings, whether the doors are open, the operation of driver's reminder appliance and the emergency bypass switch systems and the operation of the brake system.

OTDR data are also used in:

- Incident/accident investigation,
- Automated train condition monitoring, for example, TAPAS condition monitoring system processes data recorded by OTDR to identify the required maintenance for trains [53],

- Automated driver assessment, for example, TAPAS and Churros process OTDR data to estimate a number of speed indictors such as the speed at which power Notch 4 is selected when accelerating.

This project extends the use of OTDR data to leading indicators for SPAD risk.

Method: data cleansing

OTDR raw data from a single class of trains is received in their native or 'raw' data format. This format is optimized for compactness and has to be reformatted before it can be analysed. The initial handling examines types and format of data channels (that appear in rows) and corrects them if needed. The two corrections are correcting notation of time and filling out missing data. The latter is helpful in the analysis as some channels only record changes of state (and don't record any data whilst the state remains the same) and some channels do not record zeroes. The changes have been manually checked against a manual extraction process using Excel. To automate the correction of data format, an R script illustrated by figure 1 has been written. Stage 1 simply turns the raw data into a CSV file format. The second stage converts relative journey into seconds: "+ 01h24mn26s6" to "5066.6" seconds. Stage 3 corrects for slightly differing times from different data channels. This is required because a relative journey time record appears more than once with different groups of variables (i.e. for the same time record, there are more than one input line from different data channels). A process was carried out to compress all variables occurred at the same time in a single data row. In Stage 4, all the missing data were filled using different logical processes, for example;

- Merging speed information from two different columns of the data source into a single column;

- Filling the missing values of train distance with calculated distance based on the available time duration and train speed.

This particular error checking is specific to train class used in this study. Other classes that use other data systems to record OTDR will need similar error handling and cleansing routines to make it useable.

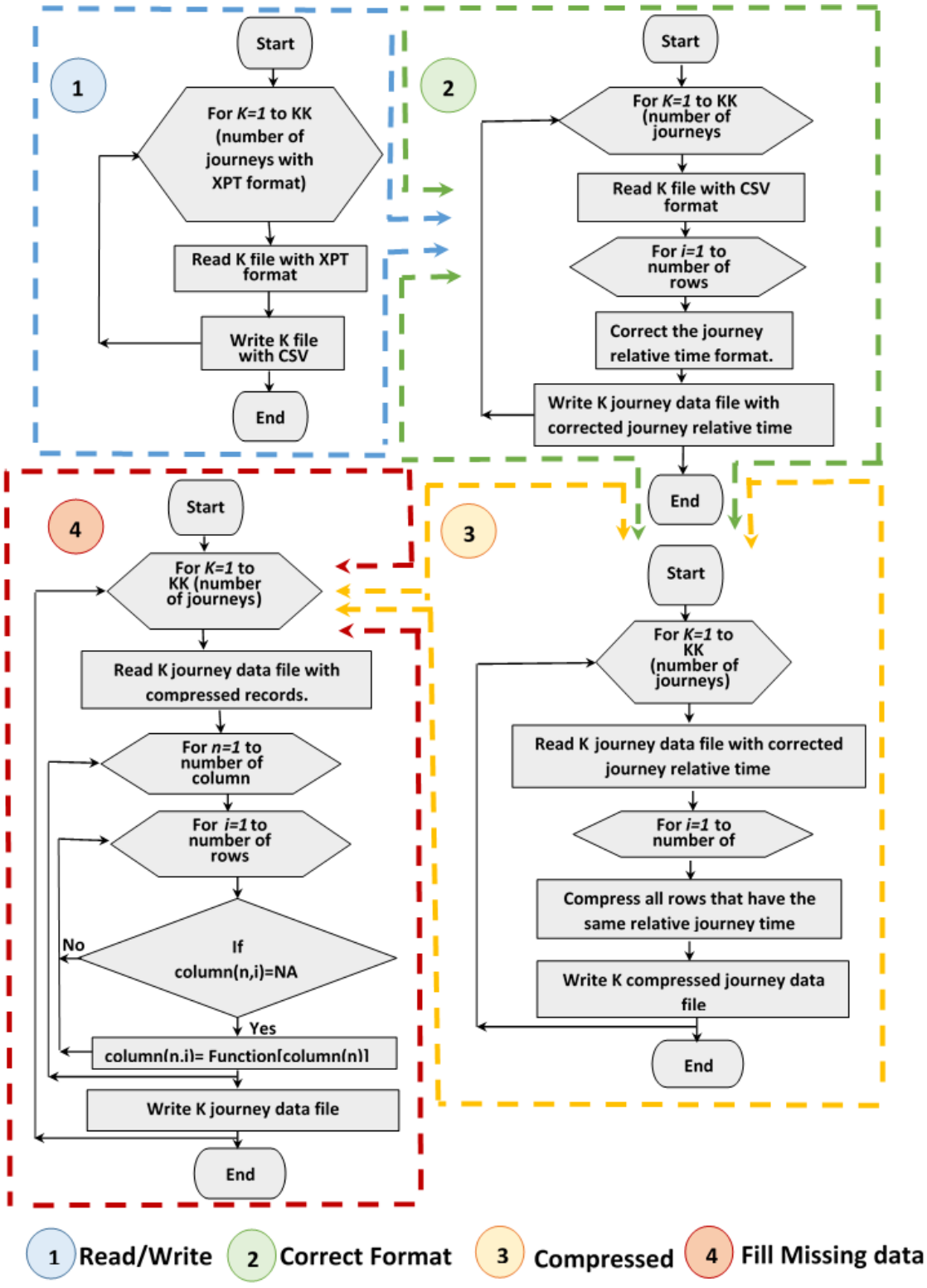

Figure 1. data cleaning and processing.

Method: approaching a red aspect

The train speed when the driver receives the last AWS horn prior to a red signal is considered a leading indicator for SPADs. A high speed when approaching a red signal may cause a SPAD or lead to a full brake application to stop the train at the correct location, about 20 metres away from the red signal. An algorithm was developed to read the train speeds when the driver receives the last AWS horn before a red signal. The algorithm differentiates between trains stopping at a station or a red signal by checking door release as illustrated in figure 2. When the train stops without a subsequent door release (so, when the train is approaching a signal outside a train station), the AWS horn event is identified from the AWS channels measurements. If there are a number of AWS horns during the period under investigation the time of last horn is extracted along with the train speed. The procedure was not optimized for computational speed; this is a subject of further work.

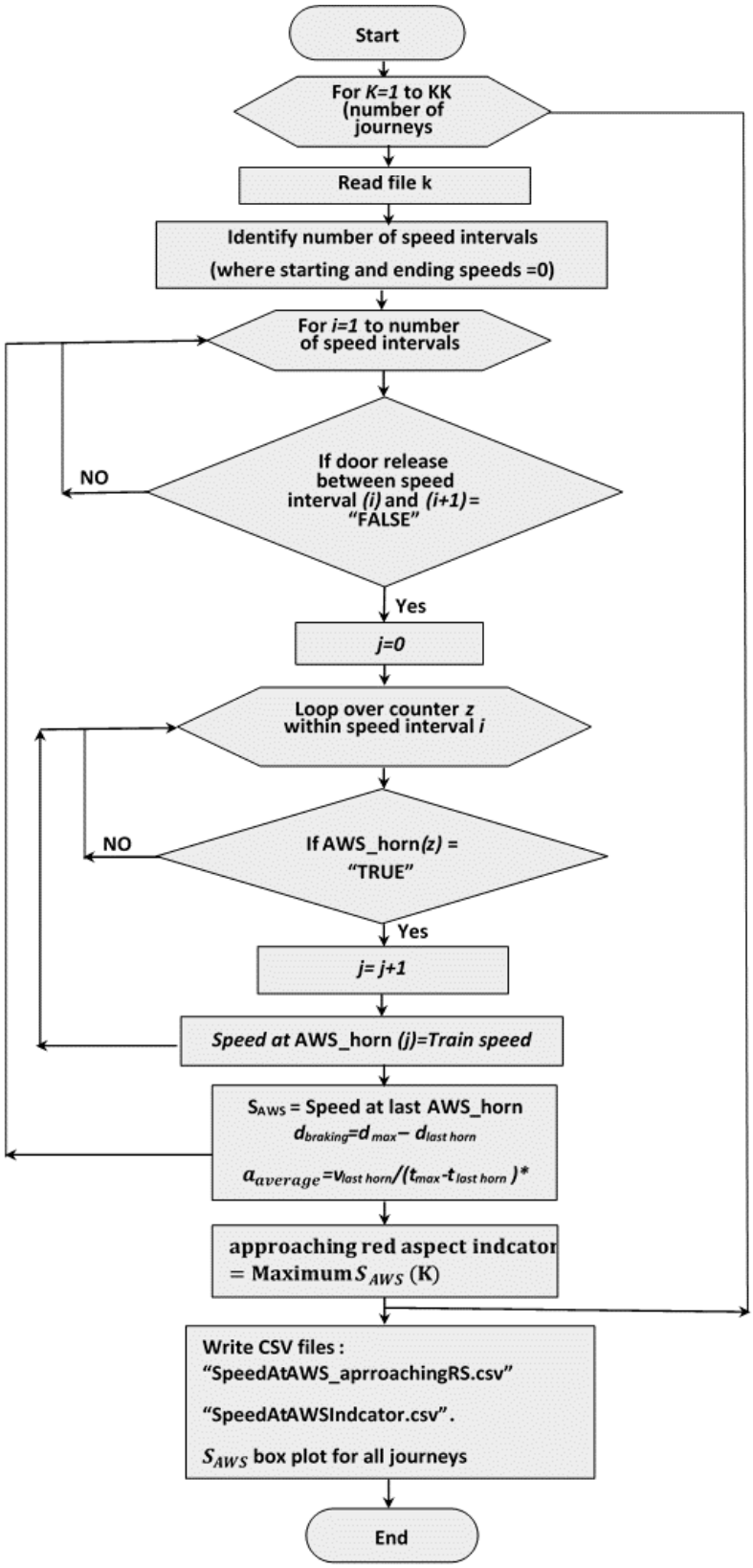

Figure 2. Speed at last AWS horn.

Results and discussion

A number of routes were investigated and the variation in the train speed at receiving the AWS last horn. Table 1 shows an excerpt of this data. The table shows the presents a number of services, how many red aspects it approached during that service and the maximum speed at AWS horn recorded for any red aspect approached at the red aspect. For one of the approaches, the speed is higher than the recommended 20 mph but only by a small margin. Further analysis showed that this train could come to a standstill at the red signal with the lowest braking step if it were immediately activated.

Journey Number	Number of red Signals	Maximum train Speed approaching a red aspect
1	1	16
2	0	NA
3	2	20
4	2	11
5	1	14
6	4	15
7	1	11
8	1	14
9	1	11
10	2	22
11	2	13
12	1	14

Table 1. Maximum speed at AWS horn prior to red aspect on a single route.

Though the results of this analysis do not seem very spectacular it has important consequences for the GB railways, especially when the analysis is scaled-up to include all trains in GB in which case it provides a national leading indicator for SPAD risks. When the indicator keeps rising on a national level, it is worth investigating the cause of the national rise. The number is potentially also useful for smaller parts of the GB railways: operating companies can compare their safety performance, high-risk routes can be

identified and particularly troublesome signal locations can be redesigned. Railway partners in GB now have to consider the desired use of this, and other digital safety solutions going forward in the future.

Also, the seemingly straightforward result hides the fact that the development of safety-inspired data systems is not a straightforward task at all. Data from different sources, recording different channels and storing them in different ways have to be harmonized, technical flaws have to be corrected (such as different times recorded in different data-channels) and harmonized, and the system has to be scaled-up to a national digital system. In theory digital interoperability language, such as RailML [27], can be used for that but it is technically challenging and requires constant review by safety experts to ensure that the desired safety outcome is met. This means that safety experts have to upgrade their skill set with knowledge about digital systems and, preferably, programming if they are to assess modern safety solutions.

Succinct description of published BDRA projects

Red aspect approach to signals (RAATS)

In the RAATS project, SPAD risks are assessed by identifying how many times trains approach a signal when it is displaying a red aspect [54]. Traditionally, this frequency is estimated from counting samples on trains. This project increases the sample set to all connected signalling systems, which covers about 70% of the GB railway network.

The source of the information used in the RAATS software is Train Describer (TD) data [56]. A Train Describer is an electronic device connected to each signalling panel which provides a description of each train (its ‘headcode’) and which section of track (or ‘track section’) it currently occupies. RAATS software reads the TD live-feed, stores it in a database, calculates which trains actually approach a red aspect and presents the data in a graphical interface or creates an excel file for further analysis. The approaches to signals for a single signal can be analysed over a period from a single day to a period of a year. Alternatively the user can choose to analyse all signals in a TD-area or indeed all the signals in the database.

Figure 3. RAATS GUI showing signal ET776.

Figure 1 shows the RAATS user interface. The pie chart shows the results for a single signal: ET776 which is located on the up Cowdenbeath line at Redford. The figure shows that at 23%, of trains approach the signal at red in the period of the 17th of August 2014 to the 13th of October 2014 which is a high percentage compared with the average. The bar chart shows the signals with the highest train approach frequencies (top ten) in the EA signalling area in Edinburgh (bottom left: Select TD). The names of the signals are not visible in this figure.

In this way, RAATS software provides intricate details about the number of trains approaching a signal at danger and helps identify high-risk signals. This information can be used in subsequent risk analyses for signals.

The full scientific description is given by Zhao [54].

Safety learning from Close Calls

A close call is a hazardous situation where the event sequence could lead to an accident if it had not been interrupted by a planned intervention or by random event [56]. Network Rail workers and specific sub-contractors within the GB railway industry are asked to report such events in the ‘Close Call’ database. Close Call reports are freeform text reports where anyone can enter a situation that, in their view, could have led to an

accident. This leaves the reporter with more freedom to report what they think are dangerous situations and could, in theory, lead to a richer data-source for railway safety issues. The Close Call Database collects approximately 150,000 entries annually. Due to the large number of records, it is impractical to manually review the records and therefore computer-based techniques have been developed to extract safety relevant information from them.

Since the key safety information is embedded in text, Natural Language Processing or NLP is used. NLP techniques have been an emerging area of study over the past two decades [57, 58, 59, 60]. One of the key problems is the inherent ambiguity in written language. These include jargon, abbreviations, misspelling and lack of punctuation. Processing of Close Call data by extracting information from free text involves five processes:

• Text cleansing, tokenizing, and tagging;

• Ontology parsing and coding (creation of a taxonomy of related words);

• Clustering (creation of groups of records that are semantically similar);

• Text analysis and;

• Information extraction.

As this process description suggests, a sensible automated text analysis is complicated. The exact procedure is elsewhere [60]. This paper highlights two investigations for the information extraction process.

The first investigation is the identification of incidents with track workers. The SMIS database (GB reportable incident database for railways) shows that incidents with track workers take place most frequently in the hours between 11:00 and 15:00. The analysis was performed to investigate whether the same pattern is present in the Close Call database. An automated search query was programmed to retrieve the protection/possession arrangements events in the Close Call database as function of time-of-day. The results are compared SMIS data. The relative distributions of these events by time of day are shown in figure 4. The figure illustrates that the SMIS incident database and Close Call reports follow similar trends during the day. Unfortunately, the times at which reports are made trend for all close calls are similar to the times reports are made for protection arrangements, which suggests that reporting bias may interfere.

The high fraction of Close Call events between 00:00 and 01:00 is due to a default of the reporting system that sets the time-stamp to 00:00 when the time of the incident is not entered by the person making the entry. This correction is made more frequently with the Close Call database than the SMIS database since there is less rigorous quality control on Close Call reports.

The second investigation is a similar problem but now focussed on trespass. The question was whether trespasses take place at certain times of the day or equal probability

throughout the day? Figure 5 shows the frequency of occurrence for trespass based on automated identification of trespass events in the Close Call database. Note that trespass does not occur with equal frequency throughout the day. The trend seems that they occur more frequently during working hours. What causes this trend is as yet unexplained but similar to the possession entries, reporting bias may play a role.

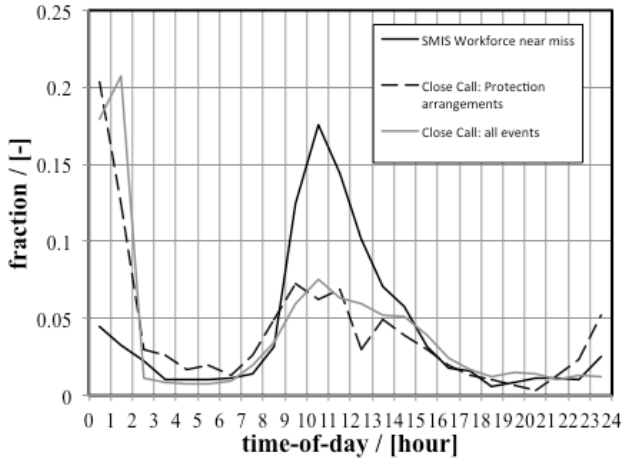

Figure 4. Frequencies of workforce incidents in SMIS and Close Call.

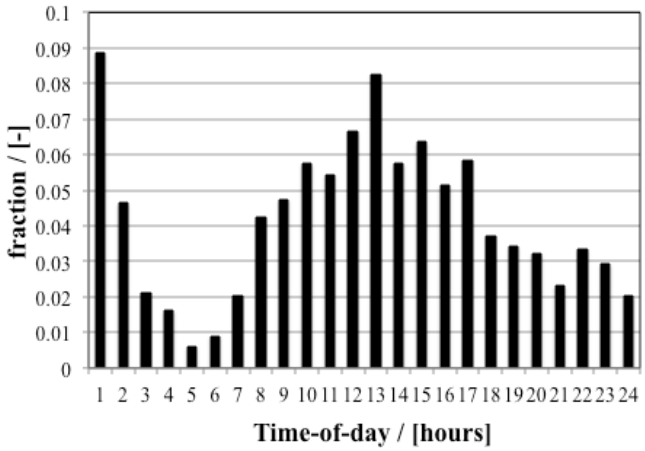

Figure 5: Frequency of trespass as function of time-of-day.

The journal paper describing this project in detail is published elsewhere [61].

Visual analytics

Visual analytics (VA) enhances the analysis and discovery of information from data; it engages creative interpretations in humans beyond those that a computer can detect automatically [62]. VA enables interactive learning with computers, which makes it quick and efficient and potentially enables better decisions [63, 64, 65].

The investigation featured here analyses a data set of 500 records that were a random sample of 12,171 Slip, Trip & Fall Close Call records. The text was pre-processed to eliminate anomalies that could obscure the text analysis. For instance, essential multi-words that represent one concept are detected to used to create a unique tokens (e.g. TRIPPING_HAZARD_ and NETWORK_RAIL_) and information of places, codes, numbers or measured entities are condensed into tags (e.g. GEO_PLACE and _CODE_).

The cleansed text was transformed into a network building its word-per-word co-occurrence matrix. The nodes of the matrix are words of the text, tokens or tags. This matrix maps the nodes of a network into pairs of nodes and it is a common input for visualisation tools. Figure 6 shows part of the word-network that was created; it identifies missing covers and trough lids as causes for slip-trip and fall incidents. The journal paper describing this project in detail is published elsewhere [66].

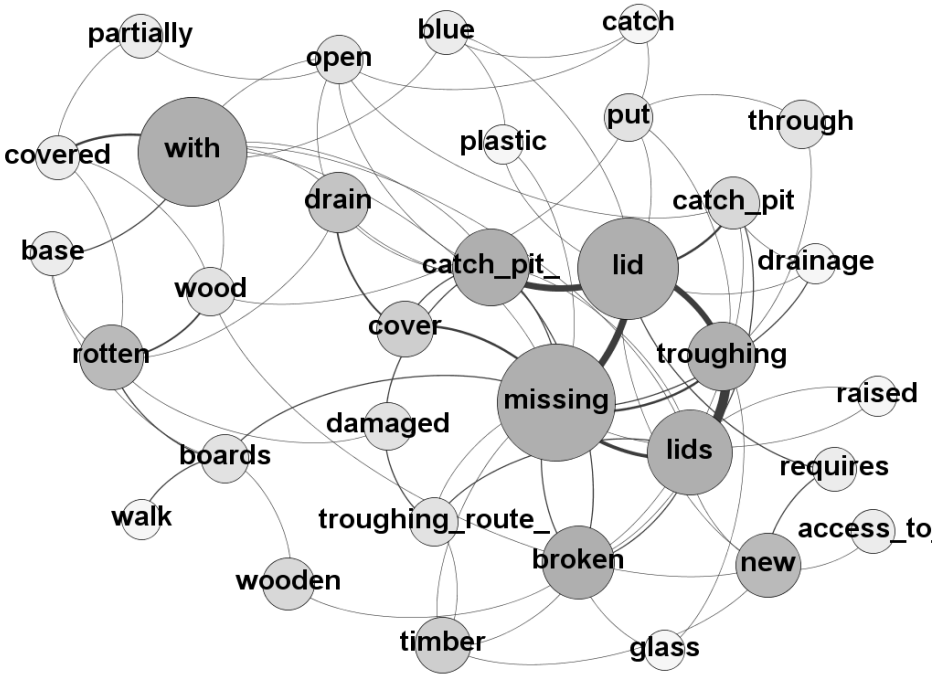

Figure 6: Visual analytics for risk identification

Ontologies

Railways have primarily used ontologies for describing railways systems in order to support traceability of information in data integration processes and design engineering processes [67, 68, 69, 70, 71]. The combination of information from different railway

data sources can provide valuable insights about the railway system and its engineering processes; however it requires a complete and structured data in order to analyze them. For instance, in the TD feed, a piece of track is referred to as a block number but in Close Call, a distance from a station indicates that same piece of track. So it is necessary to define a common vocabulary between the systems or knowledge domains that allows the communication between them. Ontologies are key enablers for the exchange of information in data-fusion since they help to maintain the context (semantics) when information is used in conjunction. Moreover, ontologies can be shared by railway operators that are in competition, helping to distinguish between shared information (such as accident reports) and intellectual property that should not be shared This work was not published in a journal but in a technical report [72].

The RAATS project uses a hard coded ontology to extract relevant information from the live stream TD feed. Concepts that it uses for extracting the right information include: train head codes (e.g. 1F98), timestamps (unix timestamp, e.g. 1458649435), TD areas (e.g. MP), signal ID's (e.g. MP1201) and Berth ID's (e.g. 1201). Concepts that it uses for analysis include: GRN (not a Red Aspect Approach incident), CSS (Cleared Stopped at signal), CAS (Cleared Approaching at Signal) and CNV (Cleared Not Visible) to describe the various types of red aspect approaches. The dissemination ontology includes: time windows; geographical areas; input concepts and analysis concepts. The ontology is a lightweight ontology with high specificity and direct application in a software system. Figure 7 shows the input ontology alongside a single data entry (derived from Zhao, 2016).

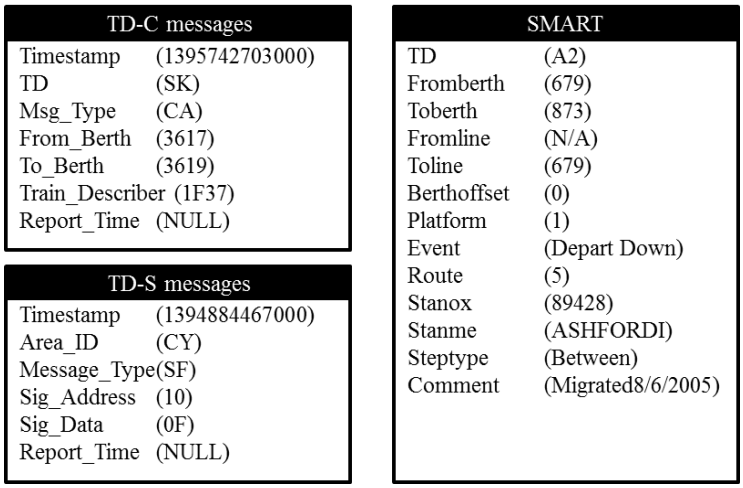

Figure 7. Input taxonomy for RAATS, example data in brackets.

SMIS+

Alongside the BDRA research programme, RSSB is in the process of completely modernising the Safety Risk Model and the tools around it to revolutionise safety risk management and decision support. The SMIS+ programme is underway which provides an opportunity to systematically capture the data that is required to perform localised

analyses, in a format that can be readily used to support risk modelling and analysis. The programme upgrades SMIS and creates a new cloud-based on-line system exploiting commercial off-the-shelf, state of the art, enterprise safety management software. The system has been specified taking into account the data needs of the safety risk model and the GeoSRM, as well as through consideration of a huge number of Railway specific 'bow-tie' models developed by RSSB, following significant work on bow-ties across the Rail Industry in Great Britain. The system incorporates the collection of information around the safety incidents currently reported into SMIS, but will also capture data relating to Close Calls as representatives for precursor incidents or breaches in safety management controls. The SMIS+ programme will offer industry a more intuitive and powerful tool to report and track all safety-related incidents in a new common format. The SMIS+ platform offers a platform for the development of advanced business intelligence (BI) features will also allow stakeholders to create their own local safety performance dashboards. A more detailed description is given elsewhere [73].

Challenges and research issues

This part of the paper focussed on brief descriptions of research projects in the BDRA programme. The projects demonstrate potential of the big data techniques for the safety and risk sciences this paragraph highlights challenges and research issues.

Invariably, more data enables a richer risk analysis. More data holds the promise of better evidence to support safety and risk analyses which, when processed adequately, informs decision-making and prognostics. However, extracting the right information is not straightforward. On an operational research level the challenges are mostly related to a skills gap. To be able to implement big data techniques, safety researchers need to understand data and databases. For example, RAATs has to deal with almost 800,000,000 messages per year. Parts of the data are corrupted parts of the data are conflicting and parts are simply absent. Safety scientists need to learn to work with a variety of huge databases that invariably contain imperfections. We found that extracting data could not simply be delegated to programmers; researchers need to understand databases to work with them or to instruct them effectively. In a similar way, safety scientists need to understand data visualization, analytics software, natural language processing and ontologies. Put it another way, safety scientists have to become proficient with basic computer science, IT systems and especially databases.

A richer data-environment is essential for a number of concepts in Safety Science. First, there is the concept that more information from near misses could strengthen the evidence base for risk analyses (which may be viewed as the contested 'iceberg' theory that saw its first iteration with Heinrich's triangle). The Close Call database was set up for that purpose. With 150,000 text records per year and automated natural language interpretation its scientific foundations can be investigated more rigorously than before. Interestingly, the Safety II approach also depends on integrating more data, even if that is on successes rather than failures. From the data-analytics approach, however, the problems are similar: a justified and effective integration of different data sources is

required. Providing the rationale and justification for integration the integration of imperfect data sources is a scientific challenge for safety research in the near future. Another area where efficient integration of databases is required is dynamic barrier management; it depends on online systems that gather data from different databases that could be physically distributed over the world. Integration of such data depends on sensible risk ontologies and clear visualization (e.g. in a bow-tie). Though the concept is relatively straightforward, its technical development is challenging even with modern Enterprise IT systems. Safety scientists will have to develop knowledge management systems based on ontologies that define and map out the relationship between concepts in safety and translate them into machine-readable systems.

Safety Scientists will be confronted with new methods and solutions and will have to assess their scientific validity for the safety sciences. The word cloud in figure 6 is an example of a new method: a well established clustering technique now contains words that have meaning for safety and risk analysts. Safety scientists will have to contemplate whether such techniques are valid for use in safety management systems and if so, under what conditions they are acceptable. This is just one example but computer scientists continue to develop tools and solutions that safety scientists will have to consider at some point in the near future.

Finally, safety scientists will be confronted with Artificial Intelligence solutions in the near future. AI presents a special problem for safety solutions in the sense that it is a black-box approach: users do not know what goes on inside. This creates an issue with trust: can we trust AI to make safety-critical decisions on a management level? And if AI made a decision that turned out to cost lives, who is liable? One way of dealing with AI is to assess its performance against benchmark systems but which systems would be sufficiently reliable to do that?

Conclusion

This paper presents the case for opening our minds to big data analytics, machine-assisted interpretation and business intelligence in safety management. It describes the drivers that shape the research, initial experiences with data systems for safety and sheds light on challenges and scientific issues. The case is compelling and worthy of our attention but it also shows that safety analysts have to adjust their skill set to be future-proof. In that sense, this work contributes to the way forward in the integration of computer science and the safety sciences.

What remains is to define what BDRA systems actually are. We suggest the following:

BDRA systems are IT solution systems that:

- extract information from data with high volume, variety and velocity to
- interpret the data quickly with a collection of software applications to
- extract relevant safety and risk intelligence to populate

- online interfaces to connect the right people at the right time in order to
- provide decision support for safety and risk management.

The process of redesigning traditional safety management systems or their digitized counterparts we shall call Safety Management System Transformation or SMS Transformation. The design of software solutions for SMS Transformations we shall call Safety Enterprise Architecture.

We suggest that research focuses on three key areas: design and specification of safety data models and safety databases that handle high volume, variety and velocity data which involves using distributed file systems to manage the scale of the data; safety ontologies and visual analytics that function as facilitators for the fusion of data-sources and machine-assisted interpretation; and artificial intelligence solutions to extract safety information from big data.

We believe that this paper describes just the tip of the iceberg of opportunities opening up for safety analysis that, after all, depends on data.

Acknowledgements

This work was funded through the RSSB-Huddersfield strategic partnership according to the MoU of 8 August 2013.

References

1. Chen H, Chiang RHL and Storey VC. Business intelligence and analytics: from big data to big impact. *MIS Quarterly* 2012; 36: 1165–1188.

2. McAfee A and Brynjolffson E. Big Data: the management revolution. *Harvard Bus. Rev.* 2012; 2012 oct: 61–67.

3. Watson HJ and Marjanovic O. Big data: the fourth data management generation. *Bus. Int. J.* 2013; 18 no. 3: 4–8.

4. Davenport TH. *Big data at work*. Boston, Harvard Business School Publishing Corporation, 2014.

5. Mayer-Schönberger V and Cukier K. *Big data: a revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt, 2013.

6. Gandomi A and Haider M. 2015. Beyond the hype: Big data concepts, methods, and analytics. *Int. J. Information Mgt.* 2015; 35(2): 137–144.

7. Jimenez–Redondo N, Bosso N, Zeni L et al. Automated and Cost Effective Maintenance for Railway (ACEM–Rail). *Procedia Soc. and Bhv. Sciences* 2012; 48: 1058–1067.

8. Kaur ER. 2015. Big Data is a Turnkey Solution. *Procedia Comp. Sci.* 2015; 62: 326–331.

9. Shingler R, Fadin G and Umiliacchi P. 2008. From RCM to predictive maintenance: The InteGRail approach. In: *4th IET International Conference on Railway Condition Monitoring*, Derby, UK, 18 – 20 June 2008. pp 1–5, London: IET.

10. White T. *Hadoop, the definite guide.* 3rd ed. USA: O'Reilly media Inc. 2012.

11. Huang C, Calzolari N, Gangemi A, Lenci A, et al. *Ontology and the lexicon, A natural language processing perspective*. Cambridge, Cambridge University Press, 2010,

12. James G, Witten D, Hastie T and Tibshirani R. *An introduction to statistical learning*. New York, Springer, 2013.

13. OMG, Business process modeling and notation (BPMN) version 2.0. Report, OMG, 2010

14. TSLG The future railway, the industry's rail technical strategy. Report, RSSB, UK, 2012.

15. Attoh-okine N. Big Data Challenges in Railway Engineering. In: *2014 IEEE International Conference on Big Data* (ed. TY Lin), Washington, USA, 7–9.

16. Tan X and Ai B. 2011. The issues of cloud computing security in high-speed railway. In: *IEEE International conference on electronic mechanical engineering and information technology*. Harbin, China, August 12–14 2011, pp 4358–4363, IEEE.

17. Li H, Parikh D, He Q et al. 2014. Improving rail network velocity: a machine learning approach to predictive maintenance. *Trans. Res. part C* 2014; 45: 17–26.

18. Thomas P. The role of big data in railroading. *Railway Age* 2014; August 2014: 44.

19. Yan B and Yu W. 2009, Application of RFID technology in railway track inspection. In: *IEEE First international work-shop on education technology and computer science*. Wuhan, China 7-8 Mar 2009, pp 526–529, IEEE.

20. Zhang X and Tentzeris M. 2011. Applications of fast-moving RFID tags in high-speed railway systems. *Int. J. Engng. Bus. Mngmnt.* 2011; 3 no. 1: 27–31.

21. Makalar B and Roy BK. 2014. Survey of RFID applications in railway industry. In: *IEE First international conference on Automation, Control and Systems.* Hoogly, India, 1-2 feb 2014, pp 1–6, IEEE.

22. Kour R, Karim R, Parida A and Kumar U. 2014. Applications of Radio Frequency Identification (RFID) technology with e-maintenance cloud for railway system. *Int. J. Syst. Assur. Eng. Manag.*; 5(1): 99–106.

23. Lee JS, Choi S, Kim S et al. A Mixed Filtering Approach for Track Condition Monitoring Using Accelerometers on the Axle Box and Bogie. *IEEE Trans. Instrumentation and measurement*; 61(3): 749 – 758.

24. Buckett G and Kimkeran S. Asset Data, A New Beginning, A New AURIZON: OmniSurveyor3D. In: Proceedings of Ausrail Conference. Melbourne, Australia, 24 - 26 November 2015, Ausrail.

25. INTEGRAIL. Intelligent integration of railway systems, http://www.integrail.info (accessed on 26 July 2016).

26. ACEM rail. Automated and cost effective railway infrastructure maintenance, http://www.acem-rail.eu (accessed 26 July 2016)

27. RailML. Data exchange by railML: As easy as changing the train, https://www.railml.org/en (accessed on 26 July 2016).

28. Deutsche Bahn. Open-Data-Portal, data.deutschebahn.com (accessed 26 July 2016).

29. Network Rail. Digital Railway, http://digitalrailway.co.uk (accessed 26 July 2016).

30. SBB. Open data, www.sbb.ch/opendata (accessed 26 June 2016).

31. SNCF. Open data, data.sncf.com (accessed 26 June 2016).

32. ORBIS. The ORBIS programme - Transforming Network Rail's approach to asset data capture, https://www.youtube.com/watch?v=u5p1nLLSZ_M (accessed 26 July 2016).

33. Papadopoulos Y, Walker M, Parker D et al. Engineering Failure Analysis & Design Optimisation with HiP-HOPS, *J. Eng. Failure Analysis* 2011; 18: 590 – 608.

34. Kelly TP, and McDermid JA. A systematic approach to safety case maintenance. *RESS* 2011; 71: 271–284.

35. Lewis RO. *Independent verification and validation*. New York: John Wiley, 1992.

36. Bernard B, Bidoit M, Finkel A et al. *Fundamentals of software engineering, systems and software verification*. 2nd edition, Upper Saddle River, Prentice Hall, 2001.

37. Kolmorgen VP and Huerlimann D. RailML - a standard interface for railway data. *European Railway Review* 2005; 4: 80 – 83.

38. Issad M. Kloul L and Rauzy A. SCOLA, a scenario oriented modeling language for railway systems. *INSIGHT* 2015; 18(4): 34–37.

39. Goodall W, Fishman T, Dixon S and Perricos C. Transport in the digital age, disruptive trends for smart mobility. Report, Deloitte, UK, 2015.

40. ERMTS. The European rail traffic management system, www.ermts.net (accessed 26 June 2016).

41. Dennis C. Development and Use of the UK Railway Network's Safety Risk Model. In: *Proceedings of the Twelfth Safety-critical Systems Symposium*, Birmingham, UK, 17–19 February 2004: 69-89, Springer.

42. Muttram RI. Railway Safety's Safety Risk Model. *Proc. Inst. Mech. Engrs. Part F: J. of Rail and Rapid Transit* 2002; 216: 71-79.

43. Sadler J, Griffin D, Gilchrist A, Austin J et al. GeoSRM – Online geospatial safety risk model for the GB rail network, *IET Intelligent Transport Systems* 2016; 10(1): 17-24.

44. Damiani E. Toward Big Data Risk Analysis. In: 2015 IEEE International Conference on Big Data, Santa Clara, USA, 29 October – 1 November 2015, 1905–1909, IEEE. DOI: 10.1109/BigData.2015.7363966.

45. Pence, J. Mohagheig, Z., Ostroff, C., Dang V., Kee E., Hebenak R. and Billings M. Quantifying organizational factors in human reliability analysis using the big data-theoretic algorithm. In: Proceedings of the International Topical Meeting on

Probablistic Safety Assessment and Analysis, Sun Valley, USA, 26 – 30 April 2015, 650 - 659.

46. Guo S.Y., Ding L.Y., Luo H.B. and Jiang, X.Y. A Big-Data-based platform of workers' behavior: observations from the field. *Accident Analysis and Prevention* 2016, 93: 299–309. DOI: 10.1016/j.aap.2015.09.024

47. Walker G. and Strathie A. Big data and ergonomics methods: a new paradigm for tackling strategic transport safety risks. *Applied Ergonomics* 2016, 53: 298-311. DOI: 10.1016/j.apergo.2015.09.008.

48. Choi T., Chan H.K. and Yue X. Recent development in Big Data Analytics for business operations and risk management. *IEEE Trans. Cybernetics* 2017, 44: 81-92. DOI: 10.1109/TCYB.2015.2507599.

49. Pitblado R., Fisher M., Nelson B., Flotaker H., Molazemi K. and Stokke A. Concepts for dynamic barrier management. *J Loss Prevention Proc. Ind.* 2016, 43: 741–746. DOI: 10.1016/j.jlp.2016.07.005

50. Bowtie Server, http://www.cgerisk.com/software/risk-assessment/bowtieserver (accessed february 2017).

51. Paltrinieri N., Khan F. and Cozzani V. Coupling of advanced techniques for dynamic risk management. *J Risk Research* 2012, 18(7): 910-930.

52. Bearfield G, Holloway A and Marsh W. Change and safety: decision-making from data. *Proc. IMechE Part F: J. Rail and rapid transit* 2013; 227(6): 704–714.

53. Green S.R., Barkby S. and Puttock A. Automatically assessing driver performance using black box OTDR data. Railway Condition Monitoring and Non-Destructive Testing. In: proceedings of 5th IET Conference, London, UK, 29 – 30 November 2001: IET.

54. Zhao Y, Stow J and Harrison C. Estimating the frequency of trains approaching red signals—a key to improved understanding of SPAD risk. *IET Intell. Transp. Syst.* 2016, 11 (1), 1-8.

55. Network Rail. Data feeds, http://www.networkrail.co.uk/data-feeds/ (accessed 26 July 2016).

56. Gnoni MG, Andriulo S, Maggio G and Nardone P. Lean occupational safety: an application for a near-miss management system design, *Safety Science* 2013; 53: 96–104.

57. Allen JF. *Natural language processing*. New York: John Wiley and Sons Ltd, 1994.

58. Wu J and Heydecker BG. Natural language understanding in road accident data analysis, *Adv. Eng. Software* 1998; 29: 599–610.

59. Dale R, Moisl H and Somers H. *Handbook of natural language processing*, New York: CRC Press, 2000.

60. Xu H, Stenner SP, Doan S, Johnson KB et al. MedEx: a medication information extraction system for clinical narratives, *J. Am. Med. Informatics Ass.* 2010; 17(1): 19–24.

61. Hughes P., Figueres Esteban M. and Van Gulijk C. Learning from text-based close call data. In: proceedings of ESREL 2015: Safety and reliability of complex engineered systems. Zürich, Switzerland, 14-22 September 2015, Taylor & Francis.

62. Keim D, Andrienko G, Fekete JD, et al. Visual analytics: Definition, process, and challenges. In: Information Visualization - Human-Centered Issues and Perspectives, Dagsthul Castle, Germany, May 28 to June 1 2007, 154–175, Springer.

63. Card SK, Mackinlay JD and Shneiderman B. *Readings in Information Visualization: Using Vision to Think*. London: Academic Press, 1999.

64. Grolemund G and Wickham H. A Cognitive Interpretation of Data Analysis. *Int. Stat. Rev.* 2014; 82: 184–204

65. Tory M, Möller T. Human factors in visualization research. *IEEE Trans. Vis. Comput. Graph* 2004; 10: 72–84.

66. Figueres-Esteban M., Hughes P. and Van Gulijk C. Visual analytics for text-based railway incidents reports. *Safety Science* 2016; 89: 72-76.

67. Delgoshaei P and Austin M. 2012. Software Patterns for Traceability of Requirements to Finite State Machine Behavior. *Proc. Comp. Sci.* 2012; 8: 214–219.

68. Easton JM, Davies JR and Roberts C. Railway modelling - The case for ontologies in the rail industry. In: KEOD 2010 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development, Valencia, Spain, 25-28 October 2010, 257–262, Springer.

69. Lewis R. *A semantic approach to railway data integration and decision support*. PhD thesis, University of Birmingham, UK, 2012.

70. Pavković N, Tečec-Ribarić Z and Sviličić T. Traceability Case Study on Rail Vehicle Control Unit Development Project. In: Proceedings of the 12th International Design Conference DESIGN, Dubrovnik, Croatia, 21 – 24 May 2012, University of Zagreb.

71. Tutcher J. Easton J. and Roberts C. Enabling Data Integration in the Rail Industry Using RDF and OWL - the RaCoOn Ontology. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, Epub ahead of print 26 July 2016. DOI: 10.1061/AJRUA6.0000859.

72. Van Gulijk C. Background of ontology for BDRA. Report: University of Huddersfield/IRR 110/124, 2015.

73. Dacre M., Harrison C. and Hunt M. Risk analysis for GB rail: today and tomorrow. *Safety and Reliability* 2016; 36(3): 166-183.