THE POTENTIAL FOR BIG DATA AND OCCURRENCE REPORTING FOR BETTER SAFETY MANAGEMENT

JEN ABLITT
THE EUROPEAN UNION AGENCY FOR RAILWAYS (THE AGENCY [1])

SUMMARY
Changes in the capability and cost of technology are driving an explosion in the use of intelligent, connected systems. Today, more *things* are connected to the internet than *people*. Tech giants like Facebook, Google and Amazon have led the advance and seem to know what and where we will buy, eat and visit before we do. Railways worldwide are investing in this technology to improve customer satisfaction, performance, ticket sales and maintenance planning. What else can we get out of this investment? What is the potential for this technology in predicting and preventing accidents? Can algorithms tame the chaos and manage the risk of accidents?

The paper will explain the Agency's [1] understanding of the popular term "Big Data" and how it can be used to model safety risk, both in real time operations, as well as planning safety management at a system level. The reasons motivating the Agency [1] to consider this technology alongside more traditional monitoring techniques will be set out. Some common myths about the potential for this technology will be exposed. Finally, readers will understand how they can find out more about this work and get involved with the wider Agency project to develop Occurrence Reporting for Europe.

INTRODUCTION

The Agency [1] is currently running a project to develop an occurrence reporting system for Europe. This is a long term, complex project, that is as concerned with roles, responsibilities, culture and the specific and controlled use of data, as it is with technology and IT systems.

Learning from accidents and incidents (occurrences) supports a risk-based approach to managing safety. Monitoring activities and learning and improving are an essential part of the Plan-Do-Check-Act cycle, at a company, sector, national and European level. Sharing the results of that monitoring allows learning from each other and an increasingly shared, harmonized and commonly accepted evidence-base.

Even though the potential benefits of making this kind of information available are evident, traditional manual reporting systems and procedures are costly and resource intensive, so we clearly need to proceed with caution. Today, new technology is changing the way that data is gathered and analysed, leading to cheaper systems and more useful information. Intelligent systems and "the internet of things", facilitated by cheaper, more available sensors and computing power, are capable of producing high volume, high variety, high velocity data, capable of adding real value. These techniques are already being widely applied to reduce maintenance costs, delays and increase passenger satisfaction.

WHAT IS BIG DATA?

Big-data is the new frontier for collecting and analysing data and for turning it into usable information. Big-data is a consequence of the increased potential for collecting data, the dramatic reduction in the price of storage devices and hugely increased computational power. For instance, modern smartphones can deliver more operations per second than the IBM "big-blue", the 1997 super-computer which is best known for winning against Gary Kasparov, a world chess champion, with a score of 2:1 in a 6 games chess match.

As an illustration, it is estimated that in 2014, the number of smartphone users[1] in the world was approximately 1.6 billion. In 2011 the number of connected devices overtook the global population of humans and it is estimated that by 2025 50 billions of sensors will be connected to the internet[2]. These figures provide an idea of the amount

[1] http://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/
[2] "Big data @ work" Thomas H. Davenport, page 11 – ISBN 97811422168165

of data generated, which can be used by super computers to infer information about smart phone users and the technical systems they interact with.

Defining big-data is difficult. In literature there are several tentative ways to define this new technology, all of them rely on the capability of big-data to handle: Volume, Velocity and Variety:

- Volume is the size of the data sets: the magnitude order is from Terabyte[3] to Petabyte[4];
- Variety means that big-data is capable of dealing with data coming from different sources and having different/no structure
- Velocity can be understood as the capability to handle data quickly (the speed data arrives) or to provide real-time information as output (the speed information is produced from the data).

The combination of these elements allows data to be used to draw conclusions about factors for which there is no data. This is achieved by identifying correlations, modelling those correlations using algorithms and then continually improving those algorithms using a combination of historical data, live streaming data and data on how conclusions are drawn and used from the data. For example, Oystercard (London's travel smartcard) data for buses and metro travel was used to infer information about complex bus journeys, even though no data is collected on *leaving* a bus. Of course, these techniques allow for anomalous and even bizarre conclusions, so that expert knowledge and oversight (for example, from people knowledgeable about London's bus network and geography) will always be required. This is important because, as explained below, human intelligence is capable of providing much richer information, able to interpret complex circumstances in terms of abstract competing goals (for example, the ability to understand that fare evasion is not the biggest problem in a scenario where a man on fire jumps a ticket barrier!).

The ability to analyse unstructured, "messy" data can unlock invaluable meaning. Again, relying on very big data sets, machines can learn to find patterns, and therefore meaning, in free text reports, social media comments, handwriting and news headlines etc. One theory is that with enough data, even if not well planned or coordinated, it should always be possible to learn *something* about the underlying activity. Of course, this is no excuse for poorly designed data projects!

Perhaps of particular interest for this audience, and indeed the Agency, is that we are on the verge of being able to reliably analyse free text in different languages.

For the purpose of our work, big-data can be defined as the combination of the necessary hardware and software which is able to handle, at sufficient speed, the input of structured and unstructured data into a system/model/algorithm. This is able to turn data into information to improve the system itself and to provide timely information to end-users.

WHY IS THE AGENCY INTERESTED?

For high frequency unwanted events (hazards), that produce on their own a robust volume of data, simple trend analysis can support better decision-making and management. For example, slips, trips and falls or rolling stock flat wheels that could be automatically detected by a Wayside Train Monitoring System.

For complex, low frequency / high consequence events, such as major train accidents, there are simply too few incidents and accidents to provide predictive intelligence. New and improved methods of monitoring and managing the risks of these types of accidents are needed in order to prevent them. Pooling data across Europe will allow much larger data sets and lead to more robust and useful conclusions. Our existing data suggests rail

[3] To put it in some perspective, a Terabyte could hold about 3.6 million 300 Kilobyte images or maybe about 300 hours of good quality video. A Terabyte could hold 1,000 copies of the Encyclopedia Britannica. Ten Terabytes could hold the printed collection of the Library of Congress.

[4] 1 Petabyte could hold approximately 20 million 4-door filing cabinets full of text. It could hold 500 billion pages of standard printed text.

safety is improving, but a series of high profile, multi-fatality accidents, in Member States with otherwise strong safety performance, is driving a need to improve the way we manage these particular risks.

To understand the risks of these events, analysis and monitoring of the causes and possible consequences of these events is required. A variety of techniques exist to map and weight (model) these relationships, including the traditional fault tree and bow tie analysis. Collection and monitoring of the high frequency contributory causes of these events can help to predict, prevent and target activity toward areas of greatest risk. We see a clear link here to the requirements in the European Common Safety Method for Monitoring[5], which requires operators and infrastructure managers to establish a monitoring system capable of providing early warning that the SMS will not achieve its intended outcome and support decision-making. Traditionally, this modeling activity requires costly, time consuming and sometimes unreliable reporting of contributory causes, as well as expert knowledge and judgement to interpret the data. Nevertheless, before Europe moves to legislate and impose (cost effective) solutions for Europe, it is vital that we understand what improvements may be possible in the near future.

Collecting data
For the purposes of our work, we have categorised data collection as automatic or manual.

Data collection is automatic when the data acquisition is triggered by a specific event detected by sensors (such as trains traversing the route on a specific point) and then collected and stored by means of technical equipment, without any human intervention.

Manual reporting can be done using technical systems or IT equipment (tablets, mobile phones, etc.) but it is always done by humans. The decision to report is not triggered by sensors but is made by human beings according to their perception of reality. This introduces a subjective element.

To date, automatic and manual reporting are complementary, in other words, you need both types of data collection to build an accurate picture for most observed activities. Automatic systems *can* mean cheaper, quicker and more reliable reporting. For instance, the actual axle load of a freight wagon could be measured by humans but it will require the use of a specific weigh scales and then a reporting procedure. A Wayside Track Monitoring System makes this measurement and reporting much easier.

On the other hand, humans are still necessary to detect and report new risks or unexpected occurrences – human intelligence is far more flexible and resourceful, understanding complex objectives and context to recognise valuable information. It is not therefore possible or desirable to entirely replace manual reporting made by humans with automatic reporting systems, and thanks to new technology able to scan and analyse free text reporting, it may not be necessary to do so.

Automatic reporting is characterised by:

- strong data reliability and structure, data is collected in a systematic way and structured according to the design of the system;
- need of technical sub-systems and the related supportive infrastructure;
- need of a strict occurrence identification, the proper sensor has to be installed to detect the desired event.

Manual reporting is characterised by:

- decision-making/contribution of human beings;
- subjective perception of reality which may lead to inconsistent and less structured information;

[5] Commission Regulation (EU) No 1078/2012 of 16 November 2012 on the common safety method for monitoring to be applied by railway undertakings, infrastructure managers after receiving a safety certificate or safety authorisation and entities in charge of maintenance OJ: L320/8 of 17/11/2012

› potential use of open text reporting, which is more difficult to analyse but could be high information density because it could also include the circumstance under which a specific occurrence took place.

Analysing data

New software techniques, made possible by the increased availability of Big Data and increased computing power, mean that the data itself can be used to identify correlations and causal relationships between monitored events. These approaches rely on technologies such as neural networks, on which self-learning predictive models can be built. In this way, previously unknown or poorly understood indicators or causes of complex accidents can be identified, tracked and managed to reduce accident risk. This is particularly useful to understanding non-intuitive relationships, such as the increased likelihood of human error if workload is *too low,* creating boredom and inattention, or relationships between advert breaks in popular television shows and power surges !

It is increasingly clear that accidents, and big accidents in particular, have complex causes (the "Swiss cheese" theory). Use of a broad and unstructured range of data could help to unearth these complex causes: human observation, audits, manual reporting, culture climate surveys, spare part procurement records, staff training records etc. This data may give us information about previously hard to monitor safety factors, such as the efficiency and effectiveness of processes and positive as well as negative human contribution to accidents. This could create the conditions necessary to model very large fault trees to target interventions.

EXISTING USE OF THE TECHNOLOGY AND METHODS IN TRANSPORT

The transport industry has implemented big-data primarily to monitor and improve the quality of service and maintenance of assets. Only a few apply this technology in the domain of transport safety and then, mostly real time operational safety management.

One of the most widely used applications of the technology is in private road transport. Private car drivers can be supported by GPS applications able to provide real-time information on traffic, accidents and other disruptions. Applications such as Google Maps, Apple Maps or Waze rely on the position of the users (provided by private smartphones) to calculate average speed and detect traffic jams. Waze is also a good example of how manual and automatic reporting can be mixed together, in fact Waze users can voluntarily report accidents, road works or other types of dangers/disruptions. These reports, combined with those provided automatically by smartphone sensors, is used to improve the quality of the information provided on traffic and itineraries.

Similarly, this technology used to support real time transport decision-making (by car drivers) is also supporting transport planners to optimise journeys and minimise walking distances, as in the example of Transport for London, given above.

Stockholmståg, a train operator in Sweden, has developed a new algorithm able to forecast delays and better support their management. According to the press[6], traffic controllers can be alerted to possible delays 2 hours before they occur. This predictive algorithm gives the traffic controller the chance to be proactive and manage the traffic in order to preserve the quality of the service. The algorithm is based on machine learning which uses historical data to identify events which led to train delays. When the system detects the same type of pattern, an alert is raised to the traffic controller in order to make timely interventions.

This application of the technology is interesting because it makes use of historical data to detect unknown causes of delays (machine learning) but also because it allows the traffic controller to simulate the effectiveness of possible solutions.

Using intelligent systems to monitor specific risks is being used to make timely and effective interventions in some European countries, for real time safety management. For instance in Switzerland and Finland there is a

[6] http://www.railwaygazette.com/news/technology/single-view/view/commuter-prognosticator-avoids-delays-which-havent-happened-yet.html

comprehensive Wayside Train Monitoring System already implemented and fully working. Those systems are able to provide data on:

- Contact force between wheel and rail, which provides information on:
  - Actual weight of the rolling stock;
  - Load balance;
  - Geometry of the wheel;
- Loading gauge (envelope);
- Temperature of axle boxes;
- Temperature of wheels;
- Temperature of the brake discs;
- Pantograph and catenary monitoring.

This is an automatic reporting system, able to inform traffic controllers in real time on specific parameters of the rolling stock, giving them the possibility to make timely decisions. This helps in avoiding accidents, improving the availability of the infrastructure and reduce maintenance costs for both railway undertakings and infrastructure managers.

But in order to really investigate and manage all the causes of accidents, the technology will need to be capable of monitoring, and contributing to, internal safety risk management processes in real-time. The usual internal monitoring methods based on audits and inspections, supplemented by manual reporting, often does not generate enough data and intelligence to support proactive interventions. A possible approach is to infer information from a wide range of more readily available indicators.

As presented to you last year (Van Gulijk, 2015)[7], Huddersfield University, working in partnership with RSSB, are working to develop a comprehensive approach:

- Data is collected and analysed using big-data techniques;
- Open text reports are analysed using text mining software;
- Information is extracted by the data set and used to feed risk models;
- The SMS of the operators have been modelled using enterprise architecture. This means that the SMS is mapped and the responsibility of the risk control measures are allocated clearly in the organisation. Each responsible person has a dashboard with the list of the controlled and uncontrolled risks;
- The combination of big-data analytics, risk models and enterprise architecture to map the SMS allows organisations to monitor their risks in real time.

[7] Big Data Risk Assessment, the 21st century approach to safety science, Van Gulijk, C., Figueres-Esteban, M. & Hughes, P.

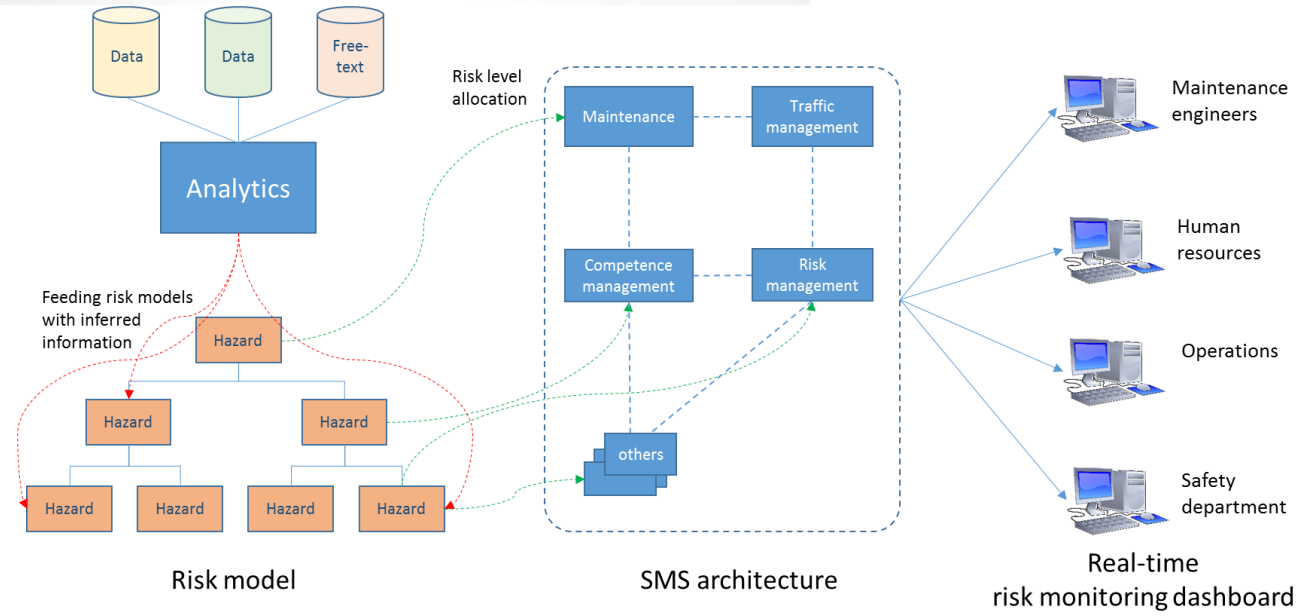

Figure 1 – Big data for Risk Assessment

The European Aviation Safety Agency is also investigating the potential for this technology to be applied to support safety management systems, itself based on a programme of the Federal Aviation Administration in the United States, which is now in its 9th year[8]. Following a feasibility study in 2015, EASA is now working on a proof of concept (pilot) phase planned for 2016 – 2019, with a limited number of stakeholders, which will require significant infrastructure. EASA have worked hard to establish the trust, willingness and support of aviation stakeholders to share this breadth of data. The intention is to integrate this project into their wider Data4Safety programme, with full implementation planned for 2019.

SOME MYTHS

We have seen that there is a strong push to invest in automatic monitoring and reporting systems, in many instances replacing tasks traditionally done by staff. Innotrans was certainly full of impressive new technologies for remote condition monitoring, predictive maintenance and more intelligent support to safety management.

The advantages for this are clear:

- the possibility to detect occurrences, which are not detectable by human beings;
- better data quality and structure;
- the efficiency of automatic systems, which can provide more data at less cost.

Nevertheless, no technology is perfect and it is still worth considering the possible costs and benefits of such systems. Automatic systems need to be designed for a specific purpose, properly installed and maintained. Moreover, their life-cycle is not only related to the obsolescence of the equipment but also to the type of occurrence to be detected – if an organisation wants to review the indicators that are monitored, new technology might be required. A system designed to report on overloading will not be able to report fire in rolling stock.

As already stated, human beings are more flexible in this regard. They can be “retrained” to identify different occurrences and also understand the context and purpose for the monitoring task. Of course, they are also limited in what they can measure.

It is also worth at this stage underlining that humans will always be needed to understand and interpret information provided by Big Data systems. Correlation is not always causation, so that relationships in data do

[8] There are similar programmes running in Latin America and Asia Pacific

not always have meaning. Coincidences do still happen! Expertise will always be vital in understanding the complexities of context. Finally, in a human centred world, people will always be involved in deciding what to do and how to implement improvements.

WHAT'S NEXT AND HOW CAN YOU GET INVOLVED?

The Agency will be launching a study to understand the potential for these technologies in rail at a European level, in terms of better ways of collecting the data we need and better analytics to predict and manage accident risk. We will be looking to our stakeholders and manufacturers of these technologies to work with us. Put simply:

- What is currently collected and what could be available and / or relevant? (see Annex)
- How much data does your organisation collect?
- Are you getting all the possible value and use out of this collection?

CONCLUSION

The Agency believes that these new technologies could ultimately save costs and make rail more competitive. These savings are not limited to cheaper and easier collection and analysis of data. Safety improvement brings considerable savings, both in terms of the costs of serious accidents, but also the costs of the everyday precursor incidents that eat into the sector's profit margins – every derailment, unscheduled maintenance, replaced component and delay incur costs that can be driven out of the business with better management.

Nevertheless, there is no simple answer and improvement will not be achieved with the flick of a switch. Collecting and sharing data takes trust, willingness, shared understanding and commitment. It is not yet known if railway operations, and their attendant risks, are capable of being managed by pooling data across different companies, networks and territories.

Finally, we must avoid the danger of undervaluing the human contribution to operating safe railways. Computing power and technology is there to support the decisions of humans, who remain at the centre of rail operations.

ANNEX

Data collected by the infrastructure managers:

- Data for the internal monitoring of the SMS:
  - CSIs, to be reported to comply with the EU legislation;
  - Indicators defined by the NOR, to be reported to comply with national legislation;
  - Internal indicators, defined by the operators to monitor their own performance and improve the SMS.
- Data describing the Infrastructure[9] and its conditions of use:
  - Data describing the rail and its conditions of use:
    - Rail temperature;
    - Track geometry;
    - Rail profile;
    - Rail corrugation.
- Data describing the rail fastening systems and their conditions of use;
- Data describing the track sleepers and their conditions of use;
- Data on trains running on the infrastructure, for each train:
  - From the TSI OPE collected through the TAF/TAP TSI[10]:
    - Train identification;
    - Identification of reporting point;
    - Line on which the train is running;
    - scheduled time at reporting point;
    - actual time at reporting point (and whether depart, arrive or pass — separate arrival and departure times must be provided in respect of intermediate reporting points at which the train calls);
    - number of minutes early or late at the reporting point;
    - initial explanation of any single delay exceeding 10 minutes or as otherwise required by the performance monitoring regime;
    - indication that a report for a train is overdue and the number of minutes by which it is overdue;
    - former train identification(s), if any;
    - train cancelled for a whole or a part of its journey.
  - EVN of the engine;
  - EVN of each vehicle composing the train;
  - Weight of the train;
  - Length of the train;
  - Running speed;
  - Maximum speed;
  - Type of goods.
- Data on the track-side control-command and signalling system:
  - State of each signal of the infrastructure;
  - Availability of block sections;
- Wayside Train Monitoring Systems:
  - Contact force between wheel and rail, which provides data on:
    - Actual weight of the rolling stock;
    - Load balance;
    - Geometry of the wheel;
  - Loading gauge (envelope);
  - Temperature of axle boxes ;
  - Temperature of wheels;
  - Temperature of the brake discs;

[9] TSI Infrastructure – Reg.1299/2014
[10] TSI OPE – Reg. 2015/995

- Pantograph and catenary monitoring.

› On-board monitoring systems installed on infrastructure maintenance rolling stock.

Data collected by RUs

› Data for the internal monitoring of the SMS:
  - CSIs;
  - Indicators defined by the NOR;
  - Internal indicators.
› Asset management:
  - Rolling stock:
    - Data collected via pre-departure checks;
    - Maintenance reports;
    - On-board monitoring systems:
      - On-board diagnostic systems;
      - On-board recording devices.
› Operational staff:
  - Competence and behaviours;
  - Medical.
› Operations[11]:
  - the detection of passing of signals at danger or “end of movement authority”;
  - application of the emergency brake;
  - speed at which the train is running;
  - any isolation or overriding of the on-board train control (signaling) systems;
  - operation of the audible warning device;
  - operation of door controls (release, closure), if fitted;
  - detection by on-board alarm systems related to the safe operation of the train, if fitted;
  - identity of the cab for which data is being recorded to be checked;
  - Further technical specifications concerning the recording device are set out in the TSI LOC & PAS.

[11] TSI OPE – Reg. 2015/995