EUROPEAN UNION AGENCY FOR RAILWAYS

Making the railway system work better for society.

Big data in railways

COMMON OCCURRENCE REPORTING PROGRAMME

Document Type: Technical document

Document ID: ERA-PRG-004-TD-003

Origin: ERA

Activity Based Item: 5.1.2 Activity 1 - Harmonized Approach to Safety (WP2016)

Unit: Safety

Sector: Strategy and Safety Performance

	Elaborated by	*Validated by*	*Approved by*
Name	Antonio D'AGOSTINO	Jennifer ABLITT	Christopher CARR
Position	Project Officer	Head of Sector	Head of Unit
Date	12/10/2016	20/10/2016	20/10/16
Signature			

Document History

Version	*Date*	*Comments*
0.1	01/04/2016	For internal consultation, COR team
0.2	06/05/2016	For internal consultation, Agency
1	12/10/2016	Final

1. Table of Contents

1.1. List of Figures

1.2. List of Tables

2. Definitions and abbreviations

2.1. Standard Terms and Abbreviations

The general terms and abbreviations used in the present document can be found in a standard dictionary. Furthermore, a glossary of railway terms that focuses primarily on safety and interoperability terminology, but also on other areas that the Agency can use in its day-to-day activities as well as in its Workgroups for the development of future publications, is available on the Agency website.

2.2. Specific Terms and Abbreviations

Table 1 : Table of Terms

Term	*Definition*
Agency	The European Railway Agency (ERA) such as established by the Regulation (EC) No 881/2004 of the European Parliament and of the Council of 29 April 2004 establishing a European railway agency, as last amended by Regulation (EC) No 1335/2008.
Hazard	Potential source of harm or adverse health effect on a person or persons.
Information	Data endowed with meaning and purpose. It is interfered from data and deemed useful.
Occurrence	Occurrence means any safety-related event which endangers or which, if not corrected or addressed, could endanger a train or any rolling stock, its passengers, staff or any other person, and includes in particular an accident and incident.
Risk	Means the frequency of occurrence of accidents and incidents resulting in harm (caused by a hazard) and the degree of severity of that harm[1].

Table 2 : Table of Abbreviations

Abbreviation	*Meaning*
EASA	European Aviation Safety Agency
ECM	Entity in charge of maintenance
COR	Common Occurrence Reporting
ERA	European Railway Agency
EVN	European Vehicle Number
IM	Infrastructure Manager
MS	Member State
NSA	National Safety Authority
NIB	National Investigation Body
NOR	National Occurrence Reporting
OTRD	On-Train Recording Device
OR	Occurrence Reporting
RFID	Radio Frequency Identification Device
RSD	Railway Safety Directive
RU	Railway Undertaking
SERA	Single European Railway Area
SMS	Safety Management System
WTMS	Wayside Train Monitoring System

[1] EC Regulation No. 402/2013 - common safety method for risk evaluation and assessment

3. Purpose of the document

This document provides a comprehensive overview on the potentials of big data and possible applications in the railways domain.

The purpose is contribute to the Common Occurrence Reporting Programme providing knowledge and potential implementations on the technology in the railway domain.

4. Scope and objectives

This deliverable is meant to provide:

- An overview on big-data;
- Overview on big-data implementation in the transport industry;
- Pros and cons of big-data;
- Explanation on where big-data fits into the COR project;
- A summary of what the Agency learnt from consultation;
- Proposals for big-data work package.

5. Background

5.1. Monitoring of railway operations as part of safety management

Monitoring is an essential part of any management system. The well-known P-D-C-A principle, common to all the management systems, includes a monitoring process which is often condensed in the "Check" step but it is actually covering the whole cycle. In fact, monitoring needs to be designed during the planning (P), implemented during the implementation of each process (and of the management system as a whole) (D). Its results are analysed during the "Check" step (C) and finally the result of the analysis is used to improve the core business when reasonable and practicable.

The same concept is applicable to railway safety, the management systems are oriented to keep all safety risks under control, therefore there is a focus on a specific aspect of the business: Safety. In fact, the RSD requires Infrastructure Managers, Railway Undertakings and Entities in Charge of Maintenance to control risks arising from their operations. This includes also those generated by contractors and their use. The RSD impose the use of a Management System to control those risks. A monitoring process is also required and a specific Common Safety Method[2] has been drafted by the Agency.

The CSM for Monitoring requires the operators to monitor all the processes of the Management System and the Management System as a whole, this has to be done defining:

- Strategies and plans for monitoring;
- A system to collect data;
- A process to analyse data, turning it into information;
- Use of the information to improve the processes and the management system.

The monitoring framework described in the CSM for Monitoring (to be applied by RUs, IMs and ECMs) shall be proactive in order to give early warnings.

5.2. Information based decision making

The monitoring of safety performance, at all levels, from operational to regulatory, is defined with the aim to continually improving the safety level of the railway system, when reasonably practicable.

[2] EU Regulation No. 1078/2012 - common safety method for monitoring to be applied by railway undertakings, infrastructure managers after receiving a safety certificate or safety authorisation and by entities in charge of maintenance.

According to the CSM for Monitoring, data has to be collected and then analysed. An action plan should be defined whenever the analysis shows that targets are not met or that something in the system is not working according to the specifications.

Decisions are therefore based on information extracted from the data collected during the monitoring process. The higher is the volume and accuracy of the information, the more effective the decision could be.

5.3. Current practice in monitoring

The Agency do not have a clear overview of the strength and quality of monitoring processes within sector companies. Even information on the implementation of the CSM for Monitoring is missing or incomplete[3].

According to the information collected during the development of this paper, with the exception of the UK, it is reasonable to assume that the completeness and complexity of the monitoring system are related to the size of the railway operator.

The UK has a different scheme in place. Its cooperative approach to railway safety management, makes it more affordable for small operators. This is not to say that safety management undertaken by small operators is less effective, because a simpler approach could be justified by simpler operations, but in general the UK approach is a good reference in terms of safety management and consequently of data gathering and sharing across the whole industry in the whole country. One important example is the Safety bulletin issued regularly by RSSB with the intention to inform the industry about safety performance and risk profiles.

In the rest of Europe, new companies are established with the purpose to support data sharing and cooperative safety management, but they are still in an early stage. This means that, in other European member states, data are collected and analysed at a company level and shared with the NSA, only when strictly necessary or to fulfil legal requirements.

In this scenario, the incumbents are trying to define and implement their own data collection systems, which are often part of their digitalisation strategy, with some tentative ways to harmonise their approach. This is more driven by business needs, such as the possibility to better manage traffic, passengers attendance and quality of the service in general. Safety is not explicitly used as purpose for the railway digitalization.

5.4. Data collection and structure

The completeness and complexity of the internal monitoring process of each railway operator is also linked with the collection of data. Data collection can be classified as automatic and manual.

It 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 manually 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 to be considered complementary. Automatic systems allow to detect issues which are not easily detectable by humans. For instance, the actual axle load of a freight wagon could be calculated by humans but it will require the use of a specific balance and then a reporting procedure. A WTMS makes its measurement and reporting much easier and reliable.

On the other hand, humans are still necessary to detect and report new risks or unexpected occurrences.

Due to the current technology limitation, which cannot match the intelligence and flexibility of human beings, it is not possible to replace manual reporting made by humans with automatic reporting systems.

[3] Assertion to be verified

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;
- 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.

5.4.1. Costs

According to the information collected for drafting this paper, the current trend is to implement automatic reporting systems, also in replacement of manual systems, when possible.

This is mainly justified by:

- 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.

To date, the Agency cannot demonstrate the validity of the second point. Automatic systems need to be designed, 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. Technical systems are able to detect what they are designed and programmed for. A system designed to report on overloading will not be able to report fire in rolling stock. Therefore, if in three years' time the company will need more data on occurrences involving fire, it will have to implement a different system.

Human beings are more flexible but also limited in what they can measure.

An impact assessment is needed to provide more information on the cost-effectiveness of reporting systems.

5.5. The role of big-data in monitoring process and in occurrence reporting

Data collection and analytics have changed substantially in the last 10 years. New opportunities are the result of technological progress applied to other industries such as sales, healthcare, road transport and aviation. The big-data technology helps improving data collection and analytics with more sophisticated tool for data collection, analysis and visualization but also through the possibility to reduce the human intervention in the reporting systems.

The Agency believes there is room for improving the detection, reporting and analysis of occurrences. This is why, in the scope of the COR project, the Agency has decided to investigate the potential for big-data in the railway industry with the main objective to improve the safety level of the Single European Railway Area and the efficiency of the occurrence reporting by reducing manual reporting.

6. Basics on Big-data

6.1. Main concept

Big-data is the new frontier for collecting and analysing data and for turning it into usable information. Big-data is the evolution of past data analytics techniques and it is a consequence of the increased computational power, the dramatic reduction of price of storage devices and the increased potential for collecting data due to the technological progress. For instance a 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.

In addition to the sheer computational power, the difference is also in electric power consumption and variety of data. For example, a smartphone can provide raw data on position, temperature, pressure, movements, etc. For instance, from this data it is possible to infer the habits of the smartphone owner.

It has been calculated that in 2014 the number of smartphone users[4] 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[5]. These figures provide an idea of the amount of data generated, which can be used by big-data to infer information on humans and technical systems.

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 the big-data implementations to handle, at high speed, a big volume of data, coming from various sources.

Those 3 elements are summarised using the 3 Vs approach: Volume, Velocity and Variety:

- Volume is the size of the data sets: the magnitude order is from Terabyte to Petabyte;
- 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 quickly input data (the speed to which data arrive) or to provide real-time information as output (the speed to which a meaning is extracted).

For the purpose of this document, 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 the end-users.

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

6.2. From data analytics to big-data

Big-data is the latest evolution of data analytics methodologies. The step change is significant in terms of speed and possibility to deal with big volumes of data, but the main difference is in the working principle. In fact, while the traditional techniques were based on descriptive statistics, Big-data uses inferential statistics. The difference is significant because "*while the descriptive statistics aim to summarize a sample, the inferential one uses the data to learn about the population that the sample of data is thought to represent*"[6] (Figure 1).

This is the main change that allowed big-data to be a flexible tool, able to handle structured and unstructured data, to learn from "experience" detecting patterns, relationships and dependencies and to predict outcomes and behaviours.

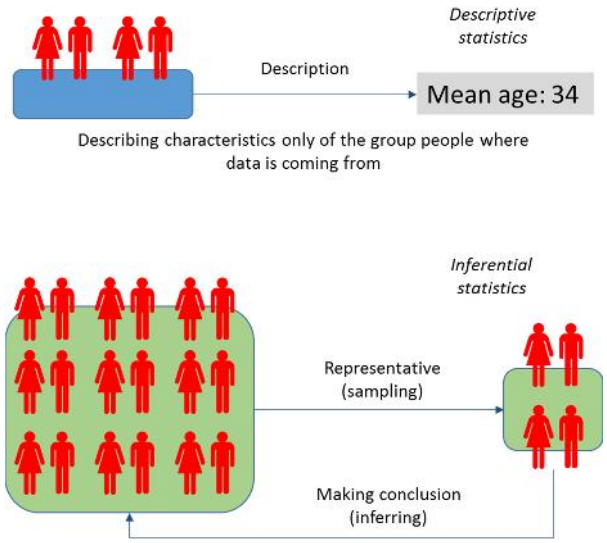

Figure 1 – Descriptive and inferential statistics

[6] Wikipedia: https://en.wikipedia.org/wiki/Descriptive_statistics;

6.3. Big-data for data collection and analytics

Applications of big-data techniques can be very different, depending on the objectives. The stock market is a domain where big-data has been used for a long time and it can be used as a valuable example.

Decisions made by traders are supported by specific algorithms. Those algorithms (models) function on "live" data (real-time streaming data) and they have been built using historical data (static data), describing the market for the past years. Streaming data, become static once they are used and stored.

This is a case of big-data implementation which includes real-time analytics (data feed the model and turn into information) but also machine learning because a big volume of historical data has been used to build the model which turns data into information for the final user (trader in this case).

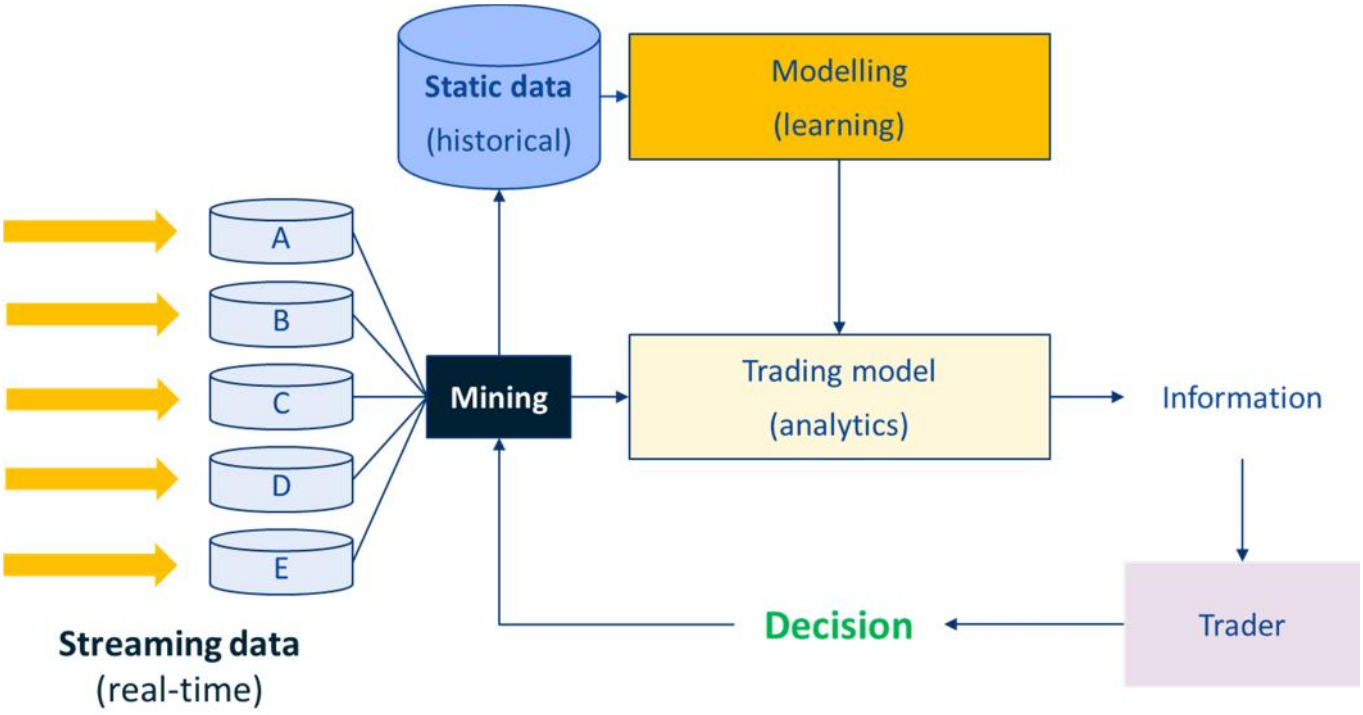

Figure 2 – Big-data workflow

6.4. Application of big-data in the public sector and industries other than transport

Big-data technology has multiple and different applications.

The public sector is using big-data, among others, to detect tax-frauds, forecast the diffusion of diseases and of meteorological disasters[7]. Cities such as New York[8], London[9], Rome[10] or Paris[11] are monitored using big-data providing information to the citizens and to app developers (open data policy) and research projects are developed to detect genomic patterns which can lead to rare diseases, this increases the chance to find quickly the most effective cure.

The private sector is using big-data mainly for customers profiling, for example, a famous grocery shop discovered a relationship between buying diapers and 6-packs of beer. Other applications are in the domain of product design and supply chains. Product suggestions in on-line stores, such Amazon, are also based on big-data technology.

Big-data was also used for managing electoral campaigns, it is the case of the re-election campaign of Barack Obama in 2012[12].

In this scenario data is always an asset. Being able to predict behaviours and wishes of customers and voters makes their personal information highly valuable, this is how companies provide "free" access to their services supported by income streams: selling user's data.

6.5. Limitation of big-data

Big-data has a huge potential in terms of information extraction from diverse data sets, which may be inconsistent and not structured but it is not a tool which can predict future events without any data related to it.

Common to all data projects, managers will still need to define a clear objective for collecting data (data modelling). Organisations will still have to invest in data reporting and data management processes to ensure that "data is valid and of high quality"**Error! Bookmark not defined.**. Big-data cannot compensate for a lack of strategic objectives and poor data quality.

Summarising, big-data is a tool to extract information from existing data. It is necessary to keep in mind that the quality of the information extracted will be as good as the quality of the data, which is a necessary condition for big-data applications.

Again, another myth which needs to be dismantled is the capability of big-data to make decisions. Big-data can provide information, based on data which cannot picture completely scenarios like human beings can do. Data is a partial description of reality and therefore is it limited in its extension and completeness.

Last but not least, big-data requires human beings for being effective. "It finds statistical relationships between variables, but it requires a skilled data scientist to determine if the relationship is independent of the data or merely an artefact of the data"[13]. This means that the implementation of big-data need to be supported by a proper team of data analysts and business experts. Sometimes those resources are available

[7] "The quiet revolution of numerical weather prediction" by P. Bauer, A.J. Thorpe, G. Brunet in Nature, 3 September 2015;
[8] NYC Opendata: https://nycopendata.socrata.com/
[9] London Datastore: http://data.london.gov.uk/
[10] Dati Roma Capitale: http://dati.comune.roma.it/cms/it/homepage.page
[11] Paris open data: http://opendata.paris.fr/explore/?refine.theme=Administration&sort=modified
[12] http://www.infoworld.com/article/2613587/big-data/the-real-story-of-how-big-data-analytics-helped-obama-win.html;
[13] "Big-data analytics" – Kim H. Pries, Robert Dunningan

in the organisations struggling to improve the analytics but a cultural change is needed to ensure their availability for safety analysis[Error! Bookmark not defined.].

7. Big-data in the transport industry

The transport industry has implemented big-data essentially to monitor and improve the quality of service and maintenance of assets. Only a few apply this technology in the domain of transport safety.

The most known big-data implementation is in private road transport. Private car drivers can be supported by GPS applications able to provide real-time information on traffic, accidents and disruptions due to maintenance works.

Applications such as Google Maps, Apple Maps or Waze rely on the position of the users (provided by the smartphone) to calculate average speed and detect traffic jams. Moreover, Waze is a good example on 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 data, combined with those provided automatically by the smartphones, is used to improve the quality of the information on traffic and itineraries.

7.1. The case of management of public transport on light rail

In public transport the use of big-data is more oriented to manage the quality of the service. For example, one company managing public transport on light rail and buses in a European capital. The company relies on the data captured by contact-less readers, based on RFID technology, to detect the position of the passengers. Following the tap-in and tap-out, the company is able to measure the number of passengers on their train routes and their distribution in time. One use of the application is to provide information to passengers concerning rush hours and the travel habits of the citizens.

For buses, passengers do not tap-out when they get off the bus, therefore the company does not get direct information from the sensors on passengers exit points. This is a typical case when big-data can help with inferring information from other data. To do so, the company, supported by university researchers, developed an algorithm able to infer the exit point of the passengers. This algorithm is using the passenger position and the position of the bus (provided by specific sensors installed on-board). Knowing the exit points helps the company in monitoring passenger numbers and journey, supporting planning of connections and minimising walking time.

7.2. A case of big-data implementation in railway transport

A Swedish train operator recently came into the spotlight because of a new algorithm able to forecast delays. According to the press[14], the 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.

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

8. Big-data in railway safety

As mentioned in the previous sections, most of the big-data applications are in the on-line sales domain, where customers behavior can be effectively described through a massive data collection done through their product search history, electronic wish-lists, social media and others.

In the safety domain, the potential sources of data are different.

Safety related data can be collected with automatic systems but also through ad-hoc monitoring activities, which can include human observation, audits, manual reporting, etc. This diversity of data sources has to be considered and requires care in order to ensure that data is correctly prepared to be analysed.

The Agency is not aware of specific applications of big-data technology to railway safety management. A stakeholders consultation did not reveal any real big-data approach to mine information to be used for operational railway safety management, for preventing accident and incidents, for risk profiling and for occurrence analysis.

Therefore, the Agency is investigating the potential for big data in the safety domain.

8.1. From traditional monitoring to big-data

Traditional safety monitoring is based on collection of indicators defined in order to represent specific events.

Those indicators are collected and analysed in order to have information on safety performance and on achievement of safety targets.

In order to ensure a common understanding of occurrences, indicators are defined to precisely identify the event to be reported.

In this context, an organization can manage safety degradations proactively and receive early warnings by modelling the occurrence that has to be avoided and identifying and monitoring precursors to those occurrences. If the accident (risk) is correctly modelled, avoiding precursors means avoiding the accident. The use of leading indicators is therefore necessary to feed the model and to get a risk profile related to the specific accident.

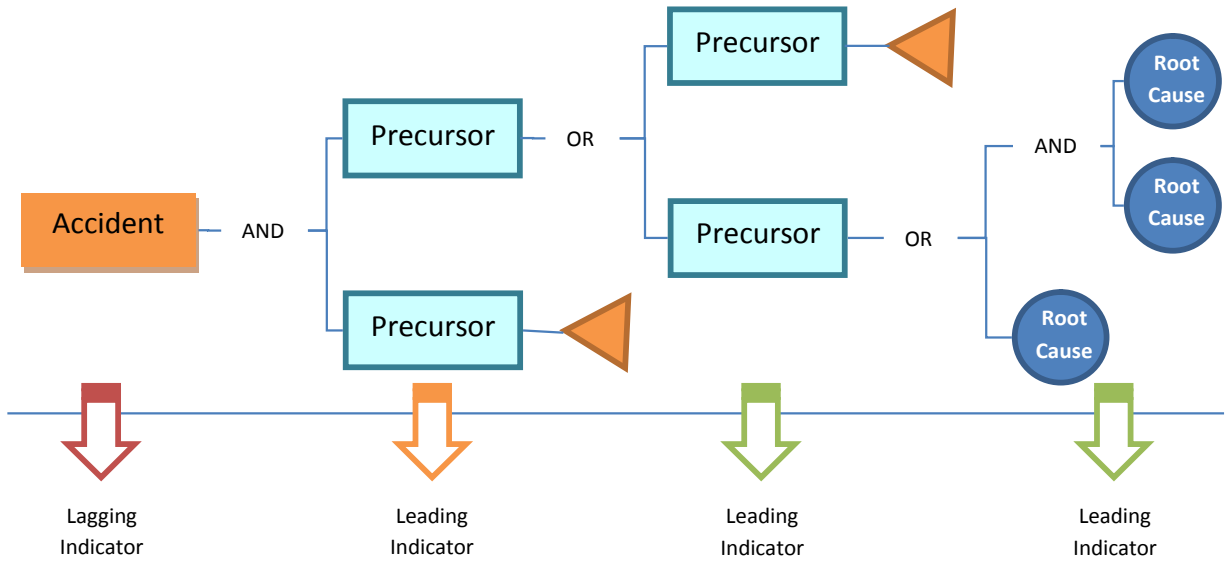

Figure 3 – An example of FTA and precursors

Building a risk model requires:

- Historical data, to identify all the causes of past accidents;
- Technical expertise, to identify all the possible causes of future accidents;
- Time and resources for analyzing all the historical data and to identify the hazards which can cause the accident;
- Time and resources to maintain the model, adapting it to the changes applied to the railway system and taking into account new hazards.

Because the data analysis is mainly made by humans, the data sets has to be "High information density", for two reasons:

- Patterns and correlations have to be simple enough to be understood by human beings, which are naturally limited in spotting complex relationships hidden in a high volume of data;
- The volume of data has to be limited by the need to reduce the time for the analysis.

Typical "high information density" datasets are safety reports or whatever dataset where the cause-effect relation is rather explicit. These exclude normal operational data from the potential data set and leave only safety-related indicators (accidents and their precursors) in the list of valuable data to be used in the monitoring process. For the same reason, providing real-time information to manage safety is rather complicated.

Big-data helps a lot in this field. Thanks to its capability to elaborate quickly big volumes of "low density information" data, coming from various sources, and its ability to infer information from it, big-data could speed up the data analysis and consequently the modelling of risks. Moreover, using machine learning, it could be possible to enable a process of self-improvement of the risk models.

All this will still require the necessary human resources but the system will be more tolerant to the structure and the content of data sets. Being able to use information that was not previously used, big-data could finally reveal hidden information that could not be extracted in the past. Big-data could also facilitate the data collection through open text used in manual reporting and real-time safety management.

8.2. An overview of the railway system and data

8.2.1. The railway system

According to the Interoperability Directive, the railway system can be categorised following two different principles. The first one, based on a functional approach, divides the railway system into sub-*systems,* such as:

1. Infrastructure;
2. trackside control-command and signalling;
3. on-board control-command and signalling;
4. energy;
5. rolling stock;
6. operation and traffic management;
7. maintenance;

Sub-systems can be classified in Fixed, Mobile and Organisational elements:

1. Fixed elements: the network, which is made of lines, stations, terminals, and all kinds of fixed equipment needed to ensure safe and continuous operation of the system; and
2. Mobile elements: vehicles travelling on that network;
3. Organisational elements: sub-systems, dealing with the functioning of the fixed and mobile elements.

Assuming that the safety performance of the railway system is depending on the safety performance of the sub-systems and of the way they are managed (SMS and operations), it is possible to assume that all data coming from sub-systems and from the management systems can support safety management.

Table 3: Railway sub-systems classification

Sub-system	*Belongs to*
Infrastructure	Fixed
trackside control-command and signalling	Fixed
on-board control-command and signalling	Mobile
Energy	Fixed
rolling stock	Mobile
operation and traffic management	Organisational
Maintenance	Organisational
telematics applications for passenger and freight services	Organisational

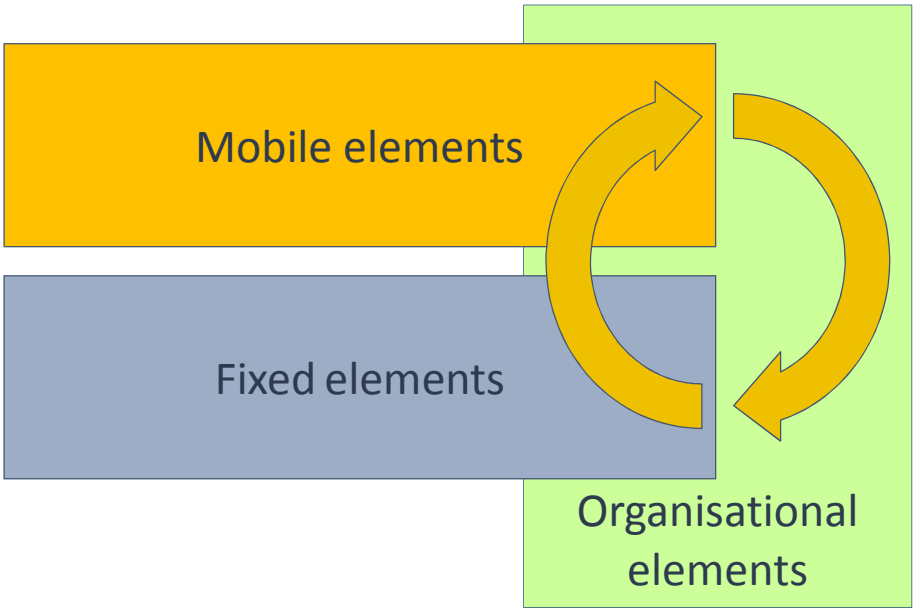

Figure 4 – Elements of the railway system

To operate, those elements are exchanging data generated as normal part of operations. This data could be used for safety purpose.

8.2.2. Possible available data

To date, the Agency is not aware of the complete data set generated by operations. The following list is an attempt to map the data and is built considering data needed by the operators to manage their business.

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[15] 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[16]:
  - 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;
- Pantograph and catenary monitoring.

[15] TSI Infrastructure – Reg.1299/2014
[16] TSI OPE – Reg. 2015/995

- 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[17]:
  - 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.

The agency is not aware of the volume of data available in the industry; the Agency intends to commission a study on the application of this technology that includes this aspect, because volume and accessibility of data is an essential condition for big-data implementation.

8.3. A new approach to data collection

The new data analytics approach helped several industries in understanding the importance and value of data. Data is more and more considered a strategic asset able to guide business decisions so data analytics is gaining importance within the organisations.

The first consequence of this change is that monitoring is more and more integrated. The idea of monitoring to fulfil separate legal obligations or business purposes is obsolete, cohesive monitoring is essential to understand the reality and to manage business in a more effective way. A centralised data management system is a competitive advantage in the modern world.

Being able to extract safety information from pure operational data, like a timetable, or being able to accurately quantify the impact on the network availability of a safety occurrence, is driving railway operators to have comprehensive monitoring systems able to guide decisions of traffic managers, maintenance engineers or safety managers. The industry is then moving from having the safety department monitoring safety occurrences and the maintenance department monitoring technical failure of rolling stock to a unique

[17] TSI OPE – Reg. 2015/995

data set (not a unique database) defined with data modelling techniques which are able to show interrelations between operational events and safety occurrences.

For instance, there has been a growth in the use of techniques such as enterprise architecture, unified modelling language, functional maps, Better Information Management and the development of standards to support data and information management.

8.4. Real time safety monitoring with big-data

The monitoring of specific occurrences in real time is already a practice in some countries. For instance in Switzerland or in Finland there is a 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.

Other examples are the data collection and communication platforms created by rolling stock manufacturers, able to ensure real-time communication with the train. This system can be used for several purposes, from optimising traffic management to ensuring real-time rolling stock monitoring, including passengers information on delays, speed, etc.

These two examples, which are not exhaustive, can give an idea of the potential of real-time safety monitoring and the current interest from industry.

What the Agency was not able to find out is how these tools are integrated between them and how far the integration can go. IMs collect data for their own purposes and so do the RUs, but sharing data and a consequent shared data analysis is far from reality at EU level.

Big-data can be a big step in this direction. With this technology, it is possible to handle data with structured differently, provided by sensors placed on the infrastructure (e.g. WTMS), traffic management systems and rolling stock to provide real-time information on technical systems but also on the organizations running trains and managing the infrastructure.

Real-time monitoring of organizations is another important point. Monitoring the implementation of internal processes in real-time can be quite difficult. 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 monitoring operations.

This is exactly what universities and researches are focussing on: the use of big-data for safety management and risk assessment .

This is a sophisticated approach to safety management using big data, the main novelties proposed by the research team are:

- 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.

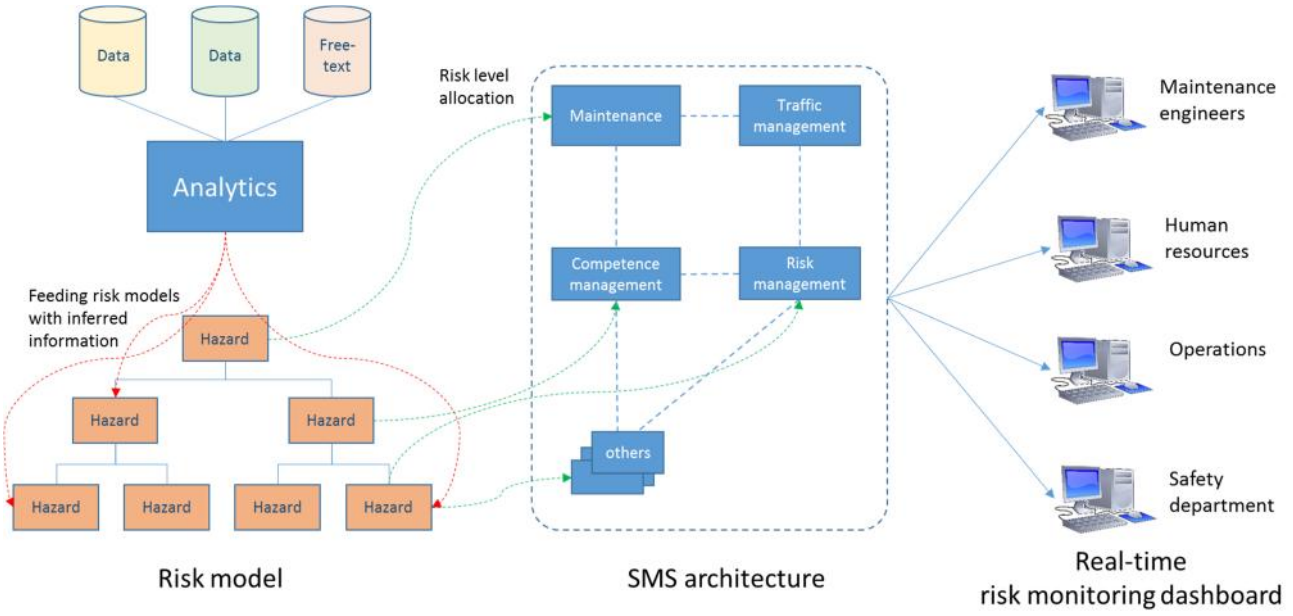

Figure 5 – Big data for Risk Assessment

8.5. Machine learning and risk models

The Agency has identified the prevention of catastrophic accidents as a specific objective for the safety management data element of the Agency's Common Occurrence Reporting project. Catastrophic accidents are rare and often attributed to a range of interrelated causes, making them difficult to model. In this case, the use of several data sets generated by different operators and NSAs for a more complete safety analysis is essential.

Building and maintaining a risk model requires a lot of data, time and resources. The use of big-data, supported by a team of data scientists and railway specialists can support this process and, because of the capability to use different types of data sources as well as identifying new patterns and correlations in the data, could also help in identifying additional and unknown hazards, as well as providing evidence to contradict "safety myths".

Building a railway risk model using big-data is an option that the agency can explore. To do so, it is necessary to check the availability of historical data in the industry. Operational data, asset data and safety records are also necessary for building a risk model using big-data. Although new to railways, other sectors, including European and US aviation, are investigating and implementing this technology.

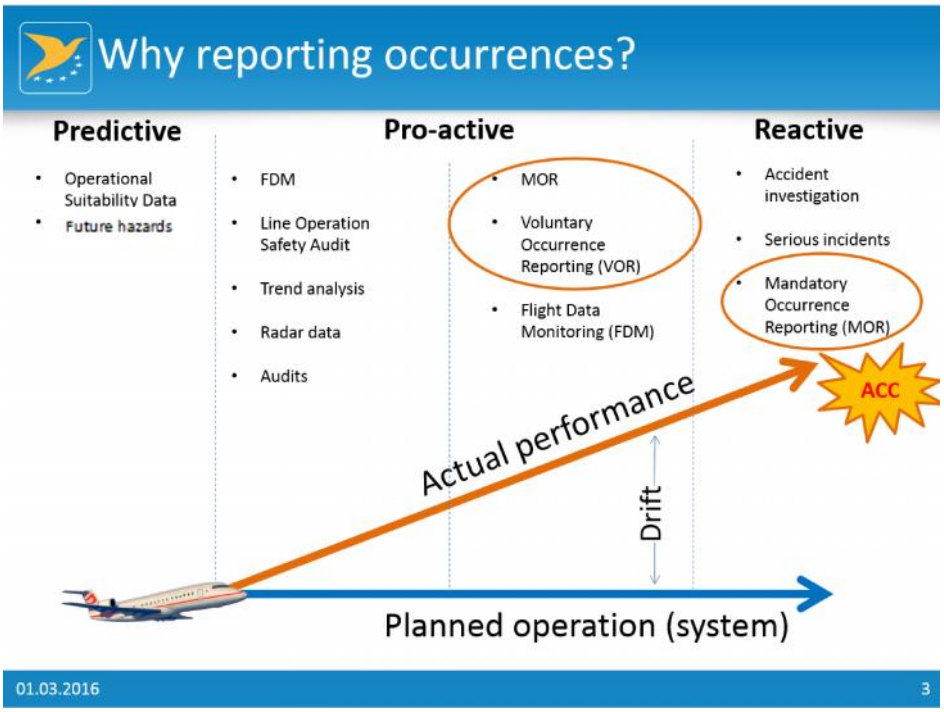

Figure 6 – Slide on Occurrence reporting from EASA

With reference to the figure above, EASA is trying to apply a big-data application used in the US. This includes a machine learning approach to analyse all the past data and to identify new event patterns and correlations that may predict the degradation of safety performance. Still, referring to the figure above, it is important to underline the difference between a pro-active and a predictive approach. The first one is essentially monitoring precursors which means when the event occurs, there is already a degradation of the safety performance.

The predictive approach is based on building knowledge. Being able to predict an occurrence from normal operational data would enable the operators to intervene before any safety degradation.

8.6. Big-data improves usability of safety information

Usability of safety information can be a critical point in the decision-making chain. The correct level of information has to be provided to the decision-makers in order to ensure efficiency and effectiveness of the continual improvement of the SMS.

The top management needs high-level information focused on supporting a budget allocation based on priorities to control risks. Operational staff needs more detailed and on-time information in order to ensure safe and continuous operations.

Big-data can help in this field because of its capability to feed visual analytics. The same data can be presented is several ways; charts, maps, functional diagrams, etc. according to the need of the users.

8.7. What big-data could help to achieve, the example of Human Factors.

Big-data handles more data, at higher speed and from different/varied sources. This is summarised in the 3Vs concept:

- Volume;
- Variety;
- Velocity.

Bigger data volumes mean the possibility to collect more or to use data collected but not analysed because of lack of capacity. **Variety** means more data because several data sources, with different/no structure can be combined. **Velocity** means capability to collect, analyse and visualise more data/information in the same frame of reference (time).

The Volume factor is extremely important in those branches of safety analysis where the description of the surrounding conditions is crucial to understand behaviours and causes. Assuming the optimal level of data quality, the more data is collected (and analysed) the more the surrounding conditions are described.

Monitoring human performance is a good example. Human performance is not easy to document using data recorded by the On Train Recording Device (OTRD), which is too limited to describe the complexity of human performance and the relevant influencing factors. Some of these factors are predictable, some may not be known/identified as of importance in the context.

For instance, it would be interesting to measure the influence of the temperature or the noise of the driving cab on the level of attention of the train driver (given the other influencing factors). Through an understanding of the effects of noise and temperature on human performance and the ability to describe the attention of the driver, it could be possible to detect changes in the drivers behaviour related to changes in the cab. Using this model it could be found that noisy locomotives generate delays, because of the lower traction effort requested by the driver in order to minimise noise. The data to undertake this analysis could come from the OTRD but will have to be complemented by other sources describing as much as possible all the influencing factors. This analysis would then require a massive amount of data to produce accurate results (Volume).

Of course, the delay may be generated by other factors like timetable, following a slower train, gradients, weather conditions affecting track adhesion, etc. but big-data will provide all this contextual information to allow a more focused analysis.

Another potential application is the possibility to measure the level of competence of train drivers by observing their behaviour. The existing on-board recording device may provide only a partial view. To stay in the field of automatic reporting systems, it would be necessary to record the conversations of the train driver, monitoring eyes movements with specific monitoring devices and much more. Research on the application of this technology is being led by automotive manufacturers to monitor and support car driver behaviour, to prevent drivers sleeping, for example. It has to be stated that human observation could also be necessary but it would introduce an element of manual reporting in the system, increasing dramatically the complexity of the system. This is an example of the importance of using different and varied data sources (Variety).

The Velocity factor is crucial in this scenario because all the data collected can be analysed quickly and sent to the traffic control centre of the organisations, which could act on time to predict and prevent occurrences.

9. The Agency and big-data, a 2 steps approach

9.1. Big-data and the COR Project

The main objective of the COR project is to preserve or improve, when reasonable, the safety level of the Single European Railway Area through a European occurrence reporting, which should:

1. give early warnings of any deviation from the expected outcome, or assurance that the expected outcome is achieved as planned;
2. give information about unwanted outcomes;
3. support decision making at both regulatory and operational level, by all the relevant actors.

The objectives will be achieved by:

1. Building awareness and support for safety information sharing at a European level;
2. Gathering and disseminating intelligence on state of the art methods;
3. Setting out clearly the cost, benefits and requirements (including legislative, resource and competence, and cultural);
4. Selecting and proposing well supported methods and plans;
5. Describing a long term plan for evolution of risk profiling built on better data.

Specific objectives for each of the specific information sharing purposes have been developed.

For the safety alert, the specific objective is to support the fulfilment the legislative obligations in the revised Railway Safety Directive.
For the Safety Management Data sharing, the specific objectives are:

- Supporting convergence of Member State safety performance across all significant and non-significant accident categories, to achieve current EU average;
- Improved understanding and management of the risks of significant and catastrophic accidents in all Member States.

In other words, the COR project essentially helps in defining new reporting schemes and possibly new approaches to data analytics. This is to improve the level of information on the railway system.

9.1.1. Why big-data

The idea of exploring the potential for big-data was not initially considered in the COR project, which, at the beginning, was mainly dealing with manual reporting of single occurrences to feed risk models.

Later on, given the results of the initial research and the informal consultation of the industry, it became apparent that there are a multiplicity of approaches to monitoring and reporting of occurrences across Europe.

In reality, those reporting systems are based on different data models, different technology with a different level of automation. Given this diversity, big-data can help the industry and the regulators simplifying the collection of data.

Another factor which led to the decision to investigate the use of big-data in the COR project is the data volume and the analytics. One of the objective of the COR project is to provide a wider and more accessible data set to improve risk management in the railways industry. This can be achieved with a phased approach:

1. Collection and sharing of historical data;
2. Learning from experience, where historical data is analysed and the results shared with the industry.

Big-data may substantially contribute to achieve these objectives. It can of course manage big volumes of data (structured and unstructured) and it can help to extract information from varied sources including open texts, using recognition of natural language. Therefore, it may simplify the analytics and the related learning process, which could be supported by the use of machine learning technology.

The vision is to acquire data directly from the operators automatically without manual or double reporting, inferring the information when is not directly collected. This makes the process more efficient.

The Agency is aware of the difficulties related to big data, the presence of data itself is not a sufficient condition to build the systems. There is also the need to build the necessary culture and trust for data sharing. Similar issues were faced by other transport mode, for instance in the aviation industry, EASA has built a Public-Private Partnership for managing operators data regulated by a strong data policy. When an operator decides to quit the partnership all its data are deleted from the centralised database.

9.2. Proposal for big-data

The architecture of the "Safety Management Data sharing" work package is summarised in the project plan[18] and it includes a step dealing with the definition of the occurrence classification and taxonomy. This item is mainly defining the list of indicators to be reported by the operators. Those indicators will deal with occurrences but also causes, consequences and general attributes like time, location, weather conditions, etc.

Big-data could be initially used to feed indicators foreseen in the reporting scheme (Figure 7). This would help the railway operators, which will grant access to their databases, to make their reporting process more efficient and effective because it eliminates the need of data input or delivery. It could also be possible to try to reduce the manual reporting (of operational occurrence) trying to infer information from data reported automatically with technical systems.

Finally yet importantly, the data could be collected using the TAF/TAP TSI and stored in a European database.

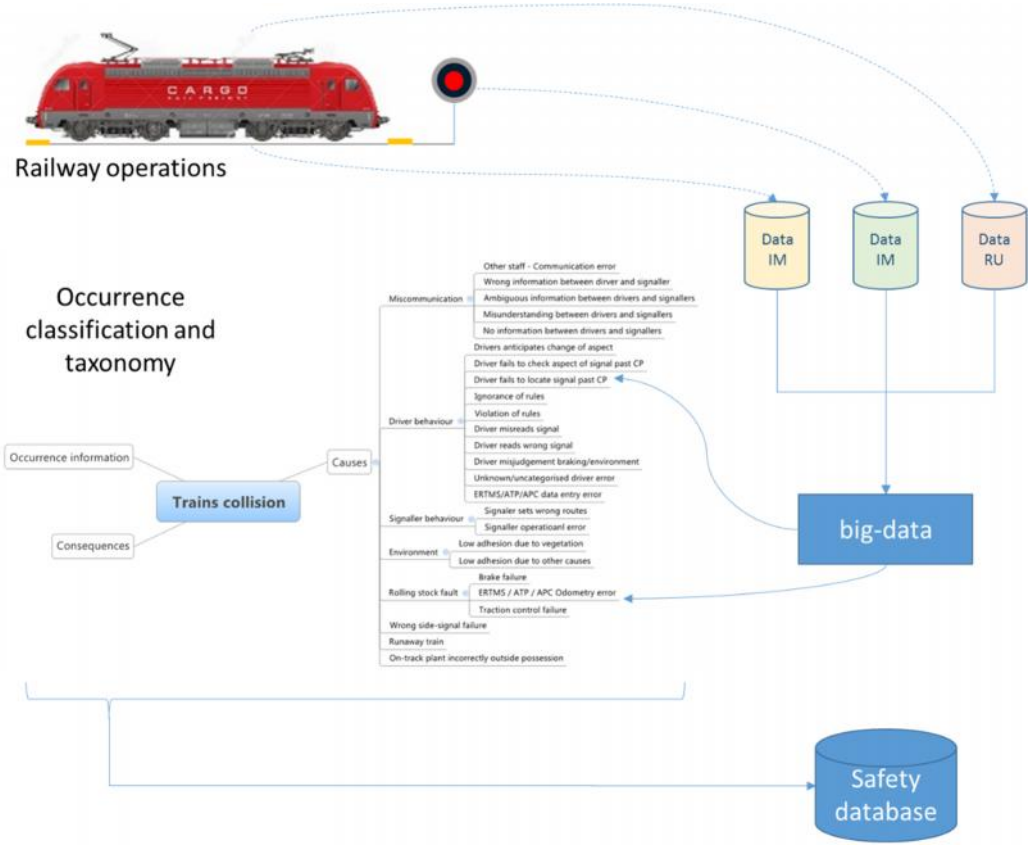

Figure 7 – Big-data for occurrence reporting

[18] ERA-PRG-004 – Project plan - Link

A further development of the use of big-data could be the creation of risk models. Following the definition of a database (or a set of them) at EU level, The Agency could develop a first risk model to be further developed using machine learning implementation of big-data. This system could run in parallel with the first proposal and could provide an evolution relying on machine learning techniques. This version may also manage to take as input the manual reporting done by operators using IT tools and free text.

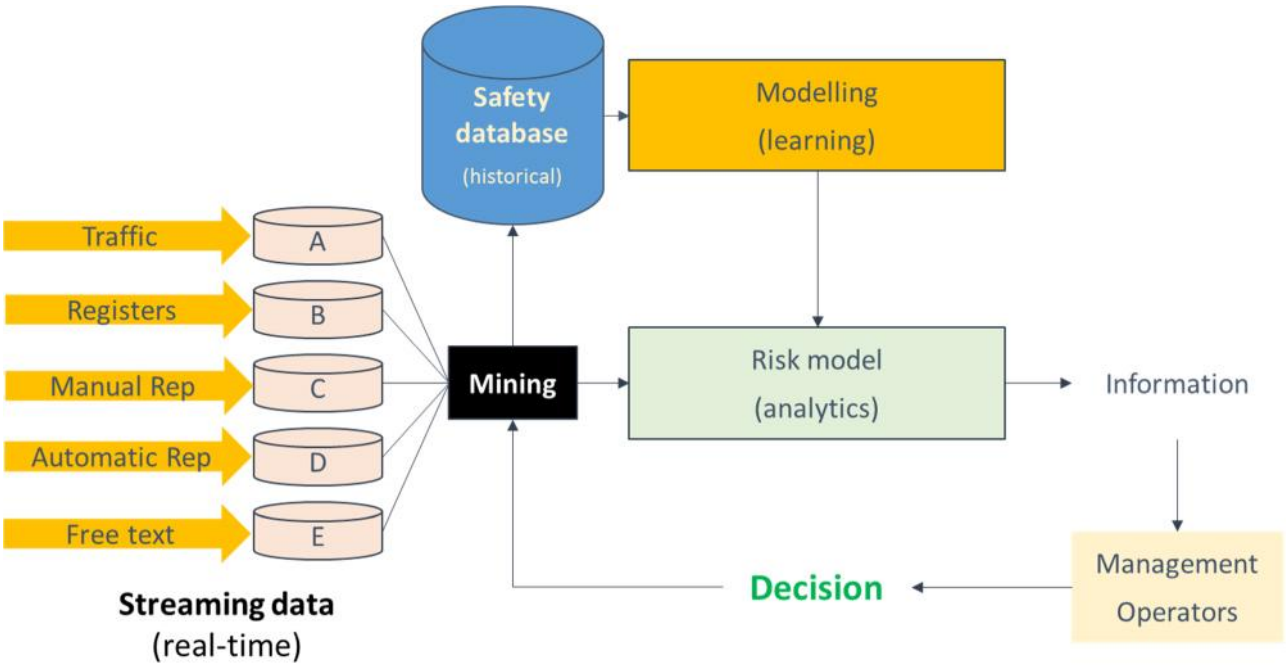

Figure 8 – Big-data and risk modelling

The final vision is a railway system entirely connected (Figure 9) where data is shared and analysed real time and used for operational safety management.

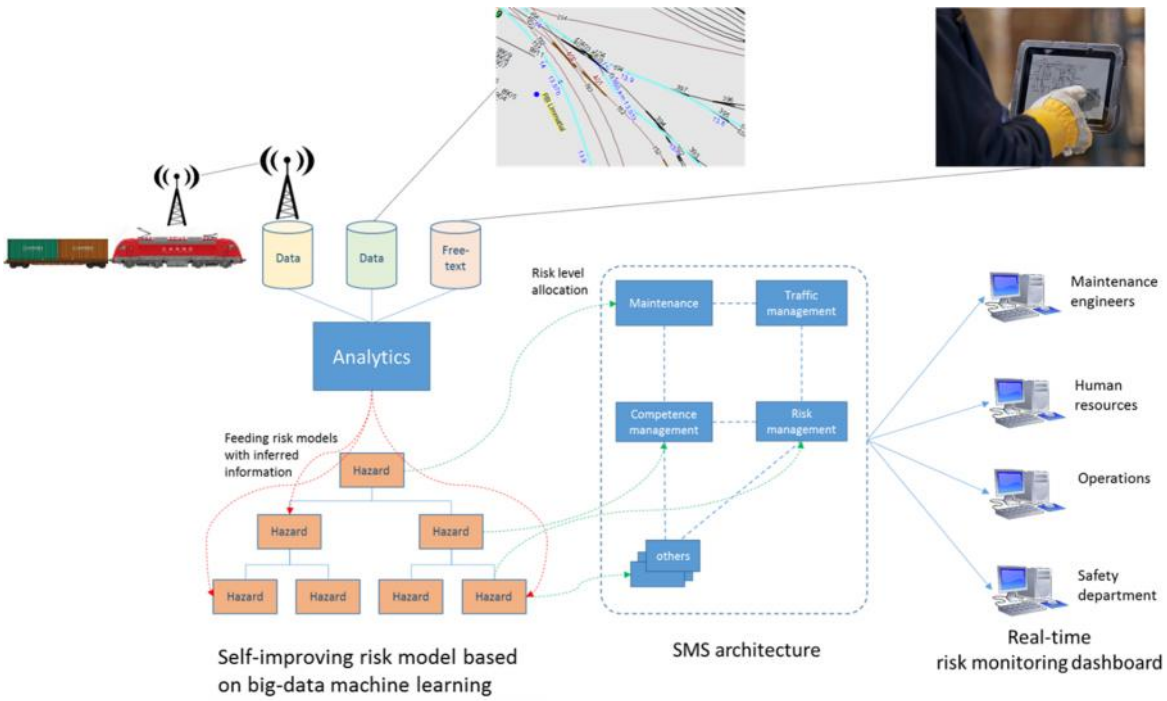

Figure 9 – Vision of big-data in the railway industry

Further details on phasing will be provided in the specific paper to be drafted by November 2016.

Big-data is a rather new idea in railway safety management and, at the Agency level, there is a need to get a better understanding of the situation of the industry concerning data management, including analytics, and availability to share data and information.

Once the scenario is clearer and there is an estimation of the amount of data, its type and possibility to share it, then a proof of concept will be necessary to prove the applicability of big-data techniques and the possible advantages offered by this technology.

9.3. Feasibility study

To explore the potential of big/data in railways, the Agency will conduct a feasibility study in 2017.

9.4. Proof of concept

Depending on the results of the feasibility study, the agency may commission a proof of concept (not yet resourced). This will be a trial implementation of big-data techniques applied to a selected range of operators. Because of the big-data pilot built by EASA, there is the possibility to use the EASA IT infrastructure to run the proof of concept.

10. Conclusions

Because of the potential efficiencies and improved results, it is worthwhile to investigate the potential for the application of big-data techniques to manage railway safety.

Even with its specific aspects, the railway system is made of interactions of technical systems and human beings. Moreover, a long history in investigating accidents and rule making can help in defining functional models and occurrence models that can be strengthened by the use of big data techniques.

The biggest element that can impair the implementation of big-data is the lack and availability of data.

To date, the Agency does not have a clear view on the current state of data collection and analytics and the difference between operators in terms of operations, size, budget, etc. may result is a very varied approach in monitoring. It is then necessary to acquire some knowledge on the topic to verify the possibility of big-data implementation.

It is important to clarify that big-data has also limitations and some myths:

1. Big-data will still require data scientist and railway experts, to validate models generate by machine learning applications;
2. Big-data will not convert manual reporting systems in automatic ones;
3. Big-data cannot create information without meaningful data.