Incorporating incident reports in bow-ties with Big Data techniques

Coen van Gulijk, Professor, University of Huddersfield, Queensgate HD13DH Huddersfield.

Miguel Figueres-Esteban, Researcher, University of Huddersfield, Queensgate, HD13DH Huddersfield.

Peter Hughes, Researcher, University of Huddersfield, Queensgate, HD13DH, Huddersfield.

Paul McCullogh, Process Safety & Implementation Consultant, CGE Risk, Vlietweg 17v, 2266 KA, Leidschendam.

The paper shows how text analysis techniques are used to automatically select text-based incident reports to import in commercial BowTie software. Text is analysed with the TFIDF method to create an ontology that facilitates the creation of search queries to match individual reports to threats in a BowTie. The method is illustrated with an example from the railway industry but it is equally applicable in the chemical industry. The approach saves time and allows for the analysis of huge volumes of incident reports. This work demonstrates big-data techniques can add value to chemical safety and paves the way to digital safety management systems.

Toward digital safety management systems

This paper focuses on the development of digitally-enabled dynamic barrier management as a key element of future safety management systems. The methods were developed for the GB railways but are equally applicable in the chemical industry. The methods that were developed are based on the BowTie as a navigation tool for dynamic barrier management. The key technology is the integration of data that is available in existing systems such as numerical risk data or incident databases. This paper explains how the lessons learned in the railways can benefit the chemical industry.

The GB railways are exploring the potential for the unremitting ingress of data systems in their industry. One of the areas that much progress was made is in the area of railway safety where the objective is to create a ‘safety control centre’ where safety experts monitor the safety condition of the railways. This vision was recently published in RSSB’s vision document entitled: ‘The Rail Industry’s Data and Risk strategy.’ This work shows some of the work toward that vision and the potential use for chemical safety (RSSB, 2017).

This paper approaches data-driven BowTies from a specific angle: the analysis of text documents. A fairly straightforward solution is described to interpret text-based incident reports which demonstrates how the BowTie approach could be made much more efficient by digital "Big Data" techniques.

Aim

The aim of the IT transformation of safety management systems is to create safety management support systems that deliver safety efficiently, effectively and rapidly. We have coined the approach as BDRA or Big Data Risk Analysis. Concisely, it is the application of digital "Big Data" techniques for safety analysis and safety management purposes. The volume of data is not actually that big in this paper but the techniques are, in theory, scalable to very large data sets. The aim of a BDRA safety management system is to:

- Extract information from mixed data sources to
- Processes it quickly to infer and present relevant safety management information which
- Combines applications to collectively provide sensible interpretation and
- Uses online interfaces to connect the right people at the right time in order to
- Provide decision support for safety and risk management

This aim is the definition that guides the development of the IT backbone for Safety Management systems.

Enablers for BDRA

Components of BDRA

Prior research performed in the BDRA research program has identified the basic enablers of a BDRA system (Van Gulijk 2015). They are the following:

- Data & data-management
- Ontology & knowledge representation
- Analytics & software
- Visualization & interface

These four enablers have to be integrated so that the basic functions of BDRA can be supported. Figure 1 shows the enablers.

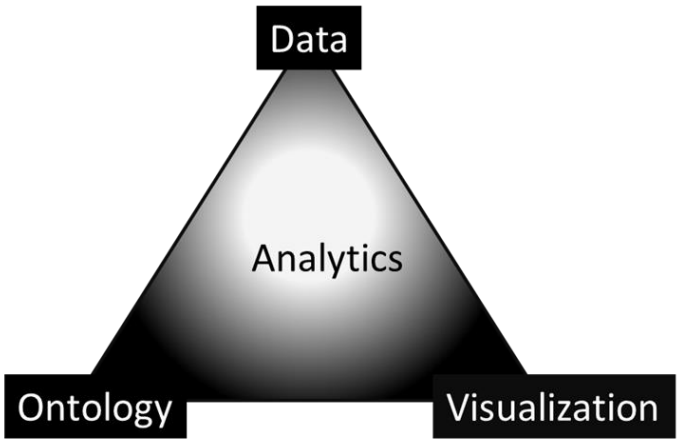

Figure 1: Basic components of BDRA.

Data & data-management

Data is a key ingredient for safety management. Modern technological systems collect huge amounts of data for several reasons. All that data might not necessarily be designed for safety purposes but safety learning can be extracted from it anyway. Supervisory Control And Data Acquisition (SCADA) systems, being very rich in data, offer potentially useful data for safety purposes. SCADA systems data use Internet communication channels for the control of machinery in chemical plants. Error messages in SCADA provide a very rich source of info that complements the data that is available in safety management systems.

Accident databases are the heart of modern risk management. The chemical industry is particularly proficient when it comes to incident reporting, accident investigation and monitoring safety critical equipment. However, these reports are often text documents that require human interpretation for safety learning. Also, there tend to be several different databases that are rarely integrated to provide a single view of the safety status.

Ontology and Knowledge representation

Ontology is the word that captures philosophy, knowledge management, semantic networks and elements of database design in one word. Ontology forms the basis for systematically capturing and classifying domain knowledge. In computing, it is used to support database design and the design of data-models. In its simplest form an ontology is a list of words that holds the right search keys to query databases. Normally, the search keys are based on a repository of concepts and words that represent the knowledge structure of a specific domain. In this work, it is the primary instrument for organizing data and it supporting the systematic mapping of data instances onto risk sources. In that sense, ontologies are at the heart of data-model creation for risk. This paper will explain ontologies in some more detail below.

Analytics and software

Analytics and data-analytical software architecture are the backbone of any computer-based risk analysis tool. All tools and enablers are software services. In the BDRA program, a cluster computer with 180 nodes is used to support the software services and store large amounts of data. It is the IT backbone in this work. In the case of BDRA, several different software services may have to run in parallel and the results of these tools have to be combined into a higher layer of the software hierarchy. In the end, the software should also enable the use of data through visualization techniques that are accessible through the Internet. For this paper such computing power was not required, it was performed on a standard desktop computer.

Visualization & interface

Visualization is important when dealing with large amounts of data. Usually, visualization is thought of in terms of dealing with results but visual analytics tools also offer analysts tools to manipulate and process data. Visualization techniques are imperative for understanding safety management. Many digital visualization techniques are used today but they are normally not made specifically for safety management purposes. In this work we the use BowTie Server software for the visual interface.

Ontology, a basis for the interface between data and BowTie

The basis for ontology was laid down in ancient Greece and treats the fundamental nature of reality. Aristotle distinguished ten categories of reality including: objects, properties, states and Relations (Ritter 1989). Questions like ‘what is real’ and ‘what can be said to exist’ were the core questions for the ancient Greeks. Philosophical ontology exists as an independent research domain that deals with structures of objects, properties, events, processes and relations in every area of reality. It even comes with its own jargon (Searle 2006, Smith 1998) but the majority of research in ontologies is linked with computer sciences in one way or another.

Computer scientists, use ontologies in a practical manner. It is the vehicle to capture real-life concepts that computers have to deal with. Evermann (2010) show the process of building an ontology in its most rudimentary form. Ontology building involves two basic steps: the conceptualization of the real world (or part of it) by perception, recognition and sharing, and the formalization of the concepts into a specification by explication for use in software. The schema below illustrates this.

Real World to Concept to Ontology

by perception

by explication

by recognition

by formalization

by sharing

Dahlgren (1995) demonstrates that concept-based ontologies are suited for computer systems to describe a "world view". "World view" is based on naive assumptions of 'what there is in the world' and how it should be classified. Naive points to the fact that most conceptualizations are made with relatively shallow common-sense assumptions in a dominant belief system. Guarino (1997) show how naive ontology building works. The brief explanation below is based on based work by Genesereth (1987):

Consider five blocks (numbered a, b, c, d and e) in two separate piles on top of a table. A possible conceptualization can be given by:

<{a, b, c, d, e},{on, above, clear, table}>

Where {a, b, c, d, e} is the universe of discourse: the five objects, namely blocks we are interested in. {on, above, clear, table} is the set of relevant relations where table entails the concept of an object holding the blocks, clear entails the concept of not touching and on and above are relational between objects and concepts. This naive conceptualization enables all possible combinations of blocks in two different stacks on the table without the need to create new rules. So different instances (e.g. instance a,b and c,d and instance a,c,e and d,b; Figure 3) are both allowed with the same conceptualization which reduces the complexity of the description of the "world view". The reduction of complexity is of great value in the computer science domain. It is also sufficient for most uses in safety mangement because humans normally use a shallow layer of knowledge to describe meaning and intentions. Dahlgren (1995) states that nearly 80% of common-sense reasoning is based on the naive approach. The approach forms the basis for incorporating incident reports into bow ties, indeed the basis for incorporating all potential data-sources into bowties.

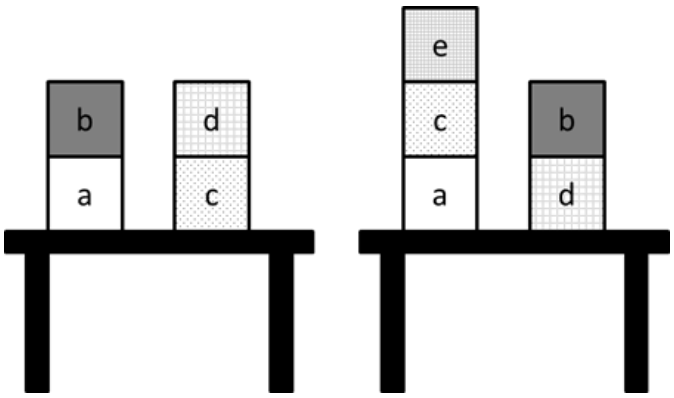

Figure 2: one naive conceptualization, 2 real world instances.

Conceptualization of BowTies

The bowtie model is the choice in the BDRA program for interfacing with data sources and management processes that are relevant for safety management in the GB railways. The approach is based on commercially available CGE-Risk software BowTie Server. It provides a framework for understanding that is shared amongst many chemical safety experts and software to present results. The ontological definitions are based on teaching materials that are delivered alongside the commercial software to ensure consistency with the software. Figure 1 shows a UML-representation of the BowTie ontology model and its semantics in order to populate bowties (Figueres-Esteban 2017).

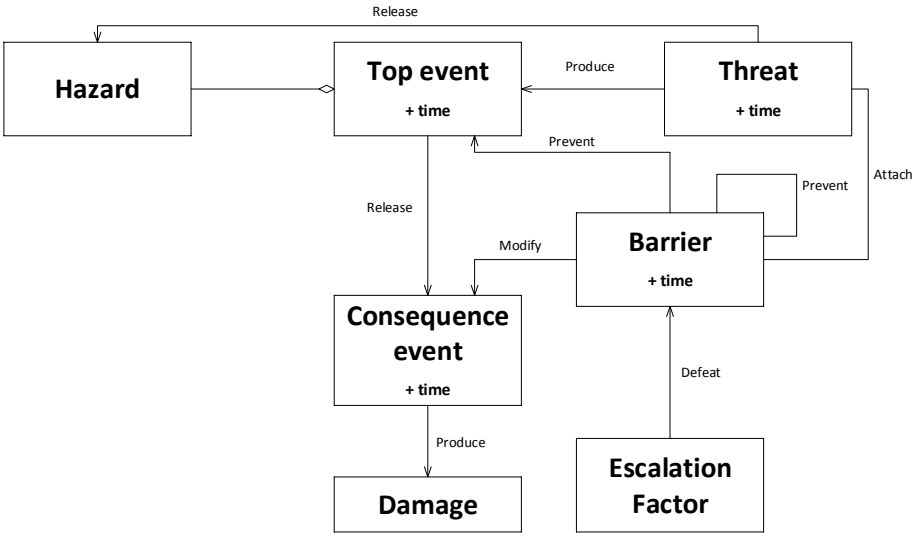

Figure 3. UML diagram that represents the bowtie model.

Table 1. Definition of key elements of a bowtie model.

Concept	Ontology definition
Hazard	OBJECT or ACTIVITY which has the potential to cause HARM. It is PART OF the TOP EVENT and is RELEASED by a THREAT
Top event	SCENARIO or UNDESIRED STATE which is PRODUCED in a point of TIME by a THREAT. It is PREVENTED by BARRIERS and RELEASES CONSEQUENCE EVENTS.
Threat	A possible CAUSE that produce TOP EVENT where the HAZARD is RELEASED. BARRIERS are ATTACHED to specific CAUSES.
Barrier	A MEANS of PREVENTING a TOP EVENT or MODIFYING the CONSEQUENCE EVENTS in order to REDUCE the DAMAGE. It PREVENTS other BARRIERS in a point of TIME and is ATTACHED to a specific THREAT.
Escalation factor	A CONDITION that DEFEATS a BARRIER.
Consequence event	A potential EVENT RELEASED by a TOP EVENT, which directly PRODUCE DAMAGE. This EVENT is MODIFIED by BARRIERS.
Damage	HARM PRODUCED in a CONSEQUENCE EVENT.

Ontology for text analysis: case study for public transport alighting

Text analysis is one of the linking pins between the BowTie and legacy data from incident databases. Dedicated text analysis queries translate concepts from the bowtie into queries that can be programmed in software to query text-based incident reports. This section uses a case study to demonstrate how queries can be used to create ontologies and extract safety risks from text. The case study focuses on the following to event: passenger injuries whilst boarding, travelling on, or alighting from trains. Five different threats are queried. In this case study, we do not drill down to individual safety barriers.

In undertaking this work, the following procedure was followed:

1. the textual descriptions of the incidents were imported into a custom-built NoSQL database;
2. text analysis techniques were used to identify terms in the text that appeared to be significant within the corpus of incident reports;
3. an ontology of key terms was created within the database to represent relationships between the concepts described by the identified terms and the operational transport network;

4. the ontology was used to structure and perform queries on the source records; and
5. the results were reviewed.

These stages are described in the sections below:

Import source data into a custom-built database

The source records were obtained as comma-separated values (.csv) incident reports. This data was imported into a NoSQL graph database that structures data in accordance with the structures used in graphs, i.e. *nodes* and *edges*.

Data relating to an individual incident was imported as a single node in the database; the database contained thousands of records. An automatic process was used to create a new node for each sentence in the text. Subsequently, each sentence was broken into individual words: punctuation marks were separated from words by inserting spaces, for example a space was inserted between a word a full-stop that followed it. Each word was then converted to lower-case text and added as a further node in the graph. This process was performed in accordance with the method described in Lyon (2015). During this process the frequency of occurrence of each word is stored in the word node. Since words represent concepts which are the basic constituents for ontologies, the process to establish meaning from text is performed by analysing the occurrence, frequency, and colocation of words or groups of words.

Apply text analysis techniques to identify key terms

Key terms that were used in the text were identified by using the *term frequency inverse document frequency* (TFIDF) method of word ranking (introduced by Spärck Jones, 1972). The TFIDF method provides a score to each word based on the frequency of occurrence of the word, but reduces the ranking for words that are used in many sentences. In this way the method attempts to identify key terms from the text whilst ignoring commonly occurring words that provide little semantic meaning, so called *stop words*. For example in English the following words are usually considered to be stop words: *the*, *of*, *a*, *to*. A modification of the TFIDF method was used to also identify bigrams in the text that appeared to be key terms.

Establish ontology of terms

Identified terms in the database were linked to an ontology structure. An ontology node was created for each concept that was represented by the identified terms. For example within the source text there are a number of terms that refer to the concept of a *train*; each of these term nodes were linked to the single ontology node manually. During this process terms that have equivalent meaning are linked to a common ontology node. A two-level ontology was used in this analysis.

Perform queries

The database was used to perform queries to identify records that contained information relevant to the threats; queries were structured in accordance with concepts that occurred in the ontology. For example, to identify records where old people were injured on trains, the ontology items for *old people* and *trains* were used as the basis for starting the queries. From the relevant ontology items, terms, words, sentences, and eventually records were identified. Note that these queries represent threats in a BowTie, that, without intervention, could immediately lead to injury.

In this case study, queries were performed for a limited number of threats relating to boarding, travelling on, or alighting from trains, being the following:

Query 1: passengers injured whilst alighting vehicles

Query 2: passengers inured whilst falling down stairs

Query 3: passengers injured whilst boarding vehicles

Query 4: passengers injured by closing doors

Query 5: passengers struck by falling bags

Review by human specialists

Safety experts reviewed the results from each query. The reviewer deciding whether an identified record correctly described the event relating to the query assessed accuracy. For example, the reviewer surveyed records obtained from query 1 to determine whether each record described an event where a person was inured whilst alighting vehicles.

Results

The two-level ontology identified seven core concepts within the text that appear to be related to the queries. A further 47 concepts were identified that appeared to be subordinate to these core concepts. Table 2 shows the list of core concepts and subordinate concepts that were identified in relation to the five queries to identify threats. Table 3 shows the number of records that were returned.

Table 2: Concepts identified from the text

Core concept	Subordinate concepts
actions	hit, closing, medical, injure, get_out, fall, enter, rush
body_parts	foot, head
direction	direction, in_between, in_front
other	bags, alcohol, drugs, stairs, footboard, customer_information_system, ticket, door
person	doctor, self, customer, person, driver, passenger, months_old, years_old, child, baby, young, old, female, male
places	line, station, pavement, hospital, ground, platform
vehicle	carriage, vehicle, ambulance, tram, train, bus

Table 3: Number of records returned for each query.

Query			
	Records	**Occurrence in database**	**accuracy**
1. Alighting	167	12%	100%
2. Falling down stairs	14	1.0%	95,5%
3. Boarding	238	17%	100%
4. Closing doors	173	13%	100%
5. Falling bags	11	0.80%	75%

Table 3 demonstrates that about 45% of the incident reports relate to boarding and alighting incidents on public transport vehicles. This is a fairly high fraction of the incidents, which is not unexpected in the public transport sector. The overall accuracy in identifying scenarios over 98%.

The findings were imported into commercial CGE Risk software: BowTie Server. This shows the results on expansion level 2 (this investigation does not identify BARRIERS).

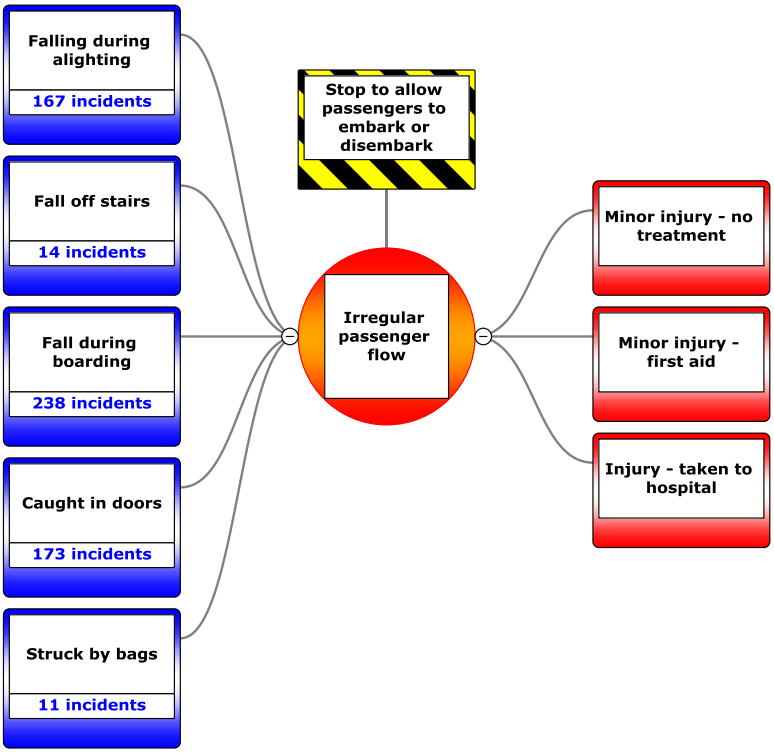

Figure 4. Results of data queries represented in BowTie Server diagram.

Discussion

This paper sheds light on what data integration can do for safety management in the future. It demonstrates a method that facilitates the automated classification of large volumes of text-based reports and the linking of the findings to a BowTie. In this paper, only one data source is used but other sources of data could be used: other incident registration systems, maintenance records, or near-miss reports. This is not limited to text-based methods; numerical data could also be used, for instance from SCADA systems.

In this case study thousands of records were used. If we assume that an experienced safety expert would take just a minute to read each record; the workload would be close to 24 hours to review every record. This is now done in a matter of minutes. Apart from the fact that this approach speeds up the analysis, the computer is also more consistent when it comes to interpretation. In fact, it would be perfectly consistent because it uses the exact same rules for detection every time. Obviously, the added value of this approach increases as more documents are gathered and additional types written documents are added.

The accuracy of detection of these threats is higher than 98%. Even if that is a relatively high accuracy it still means that some records are not interpreted accurately. This shortcoming is often due to vagueness in the text in the record but mostly it is due to records containing multiple events. Those reports might still have to be reviewed by experienced safety experts but then only 1 to 2% of records have to be read.

The ontology learning approach based on text techniques such as TFIDF method is very flexible in the sense that it yields good results in a relatively short period of time but it requires a safety expert to train the query engine. This training would have to be done for different language bodies because people that write maintenance reports use a different vocabulary than passengers. So for different sources of text, different linguistic query terms could be developed to extract the same information. Despite that, the approach also allows for changing language use over time. From time to time, the accuracy of detection has to be checked against human interpretation and the query terms can be adjusted if required.

On the contrary, the concepts of the BowTie are inflexible. Once defined they should stay the same. That makes the definition of BowTie concepts a key activity. Table 1 shows the definitions as they were used in this work. They were derived in alignment with the BowTieXP THREAT, TOP EVENT and BARRIER.

Note that this investigation focuses on identifying THREATS. It does not query for TOP EVENT, BARRER or CONSEQUENCE. Detecting those requires the addition of queries that have to be developed, evaluated and tested in the same way as the THREATS. However, in quite a few cases it is impossible to detect barriers because there seems to be no natural tendency to describe them in this incident report format. CONSEQUENCES can be extracted more easily since these records are centred around damage to people. In this investigation, that analysis was omitted as it focussed on causal factors rather than consequences.

The method presented here is flexible, fast and reliable which makes the approach a sensible direction for further research. The capstone, however, is the integration with existent BowTie software that many people are already used to. The integration and adding different data-sources are research topics for the near future.

Conclusion

A BowTie is an efficient way to survey the state of safety controls in a chemical plant but it tends to be laborious to populate it with relevant data. This work paves the way to automatic population of the BowTie. The work focussed on interpreting text analyses to populate BowTies. Despite the limited use case shown here, just 5 threats for a single transport BowTie it is clear that the method offers a consistent, and efficient means of classification that is transferrable to the Chemical Industry. The results can easily be incorporated in existing BowTie software so that experienced safety workers will be able to drill down to relevant events very quickly.

The automated safety control monitoring system can be based on various sources. That makes the approach flexible, scalable and it potentially integrates different data sources. The method has great potential for the Chemical Industry, which is drifting deeper into the IoT-supported industry.

Acknowledgements

RSSB is gratefully acknowledged for co-sponsoring this work. The Swiss FOT is gratefully acknowledged for sharing their incident database.

References

Dahlgren, K., 1995, A linguistic ontology, *Int. J Human-Computer Studies*, 43: 809 – 818.

Evermann, J. and Fang, J., 2010, Evaluating ontologies: Towards a cognitive measure of quality. *Information Systems*, 35: 391–403.

Figueres-Esteban, M. and Van Gulijk, C., 2017, Initial BowTie ontology description, IRR Report 166, Huddersfield, UK.

Genesereth, M.R. and Nilsson, N.J., 1987, Logical foundation of artificial intelligence, Morgan Kaufman, Los Altos CA.

Guarino, N., 1997, Understanding, building and using ontologies, *Int. J. Human-Computer Studies*, 46: 293 – 310.

Lyon, W., 2015, *Natural Language Processing With Neo4j - Mining Paradigmatic Word Associations.* Retrieved from http://www.lyonwj.com/2015/06/16/nlp-with-neo4j/. Retrieved 05 May 2017.

RSSB, 2017, The Rail Industry's Data and Risk Strategy, Railway Safety and Standards Board, London.

Ritter & Kohonen 1989 : Ritter, H. & Kohonen, T., 1989. Self-organizing semantic maps. *Biological Cybernetics*, 61, pp.241–254.

Searle, J.R., 2006. Social Ontology: Some Basic Principles. *Revista de sociologia, Papers,* 80: 51–71.

Smith, B., 1998, Basic concepts of formal ontology, In *Formal Ontology in Information Systems*, (Guarino ed.): 19-28, IOS Press, Amsterdam.

Spärck Jones, K., 1972, A Statistical Interpretation of Term Specificity and Its Application in Retrieval, *Journal of Documentation*, 28: 11–21.

Van Gulijk, C., Hughes, P. and Figueres-Esteban, M., 2015, Big Data Risk Analysis for rail safety?, in: proceedings of the 25th European safety and reliability conference, ESREL 2015, Zurich, Switzerland, 7-10 September 2015.