Towards Harmonizing Safety Databases: An Assessment of Existing Data Sources

Research has shown that complex sociotechnical accidents, when analyzed carefully, share a conceptual sameness in the way they occur through a combination of technical design flaws, improper operations, and organizational failings. Given the similar nature of major accidents in such systems, interdisciplinary learning from such accidents can potentially benefit all high-risk industries. As a first step towards fostering interdisciplinary learning this paper systematically identifies and reviews publicly-accessible safety databases spanning multiple domains including aviation, process industry, offshore oil & gas. A criteria-based analysis across databases is conducted to identify variabilities in format, coding structure, accessibility, quality, and quantity of information.

INTRODUCTION

Understanding accidents and accident causation remains one of the key research themes in safety research, worldwide. It is generally accepted that accidents in complex sociotechnical systems are characterized by multiple inter-related causes that can reside at various levels of an organization, system design and operation. It is also acknowledged that, while our understanding of accident causation in complex sociotechnical systems remains incomplete, accidents involving these systems will continue to occur (Hollnagel, 2004). This does not necessarily reflect poorly on the safety community; rather, it can be attributed to the ever-evolving and complex nature of sociotechnical accidents. Therefore, there is a need to capture and analyze such complex information to develop a holistic understanding of accidents.

Research has shown that complex sociotechnical accidents, when analyzed carefully, share a conceptual sameness in the way they occur through a combination of technical design flaws, improper operations, and organizational/managerial failings (Saleh, Marais, Cowlagi, & Bakolas, 2010; Cowlagi & Saleh, 2013). This sameness can be found in accidents such as Bhopal (chemical; Shrivastava, 1987), Three Mile island (nuclear; Hopkins, 2001), BP Texas City refinery (oil & gas; Saleh, Haga, Favorò, & Bakolas, 2014), and Piper Alpha (offshore; Paté-Cornell, 1993). Given the similar characteristics of complex sociotechnical systems across different domains (e.g., tight coupling, non-linear causality, cascade effect), as well as the similar nature of major accidents in such systems, interdisciplinary learning from such accidents can potentially benefit all high-risk industries. To integrate learning from accidents, safety researchers have suggested an interdisciplinary approach that synthesizes information and harmonizes links between domains (Margaryan, Littlejohn, & Stanton, 2017).

One approach to accomplishing interdisciplinary learning is by harmonizing several domain-specific databases. In fact, several sophisticated accident analysis models and techniques are traditionally applied retrospectively to the safety data that are contained in these databases (Salmon, Walker, Read, Goode, & Stanton, 2016). However, unfortunately, safety databases often exist in isolation and are not always easily accessible. While organizations such as the National Aeronautics and Space Administration (NASA), National Transportation Safety Board (NTSB) and Occupational Safety and Health Administration (OSHA) host domain-specific, publicly-accessible databases, there exist several databases that are either private, proprietary, or only accessible for a fee. To facilitate harmonization of databases, it is essential to survey databases spanning several sociotechnical domains, and identify those data sources that are readily available.

To facilitate effective learning and dissemination from these databases, it is important to understand the breadth and depth of information contained in these heterogeneous datasets. The nature of information contained in these databases are not necessarily consistent. These databases contain a variety of information including root causes, injury levels, personnel involved, operation environment, and geographic locations, which can be accessed in multiple formats (e.g., textual reports, coded data). The nature and quality of safety data reflected in organizational databases can influence analysis results (i.e., the output of any data analysis is limited by the quality of the underlying dataset). This problem is magnified when attempting to harmonize databases across multiple organizations (Wilke, Majumdar, & Ochieng, 2014). Therefore, prior to integration, it is imperative to understand the nature of the data contained in these repositories.

As a first step towards a novel global safety search engine that will enable cross-disciplinary lexical and semantic inquiries, this paper carries out a systematic review of safety databases. Specifically, the goals of this paper are: (1) To survey publicly-accessible safety databases spanning multiple domains including aviation, process industry, offshore oil, gas, and other relevant industries; and, (2) Conduct a criterion-based analysis across databases to identify variabilities in format, coding structure, accessibility, quality, and quantity of information. The criterion-based analysis will evaluate "*searchability*" (ease of searching and downloading data in analysis-friendly formats) and "*information fidelity*" (availability of dataset features such as summary statistics, investigation reports, detailed coding systems and manuals, and causal factor reports).

METHODOLOGY

This section details the methodology used in this paper. We begin by presenting the inclusion and exclusion criteria that were used to identify databases for review. Next, we present the classification and criteria-based evaluation of safety databases

that were identified. Finally, we discuss the evaluation metrics that were used to rank the different databases.

Database Selection Criteria

A comprehensive internet-based search of safety databases was conducted. Data sources that were publicly available and reported information in English were included for analysis. Databases that were private, proprietary, or available for a fee were not included. Additionally, databases that were accessible, but no longer actively maintained (or functional) were not included. Databases from the healthcare domain were not included in the current analysis owing to the wide variety of databases such as electronic health records, clinical, and drug safety databases. Future research will account for healthcare databases.

Sociotechnical Domain/Industry Type Classification

First, databases are classified based on the industry domains of the records in the databases. With the information, it is convenient to cluster the databases within same industry domain and further harmonize and compare the databases within same domain or cross different domains. The classification system used in this paper is shown in Table 1.

Table 1: Classification of databases based on industry type

No.	Domain/Industry	Description
1	Chemical	Includes databases that record accidents/incidents in the chemical and process industries
2	Ground Transportation	Includes databases that record ground transportation accidents/incidents, including road, highway, and railway transportation
3	Aviation	Includes databases that record accidents/incidents in the aviation industry
4	Marine	Includes databases that record accidents/incidents in the marine industry
5	Nuclear	Includes databases that record accidents/incidents in the nuclear industry
7	Other	Includes databases from several other domains that are not mentioned above. These include databases on occupational safety and health, offshore oil and gas, hydrogen-related incidents and accidents

Databases classified under Other included domains such as occupational safety and offshore oil & gas. Some databases listed under this category can span multiple domains. For example, the Hydrogen Incident and Accident Database collects the records incidents, which can occur during chemical processing, transportation, or during commercial use. Another example is the Database of Radiological Incidents and Related Events, which compiles information relating to incidents involving radiation. Such events can occur at multiple sites such as nuclear industries and medical facilities.

Database Evaluation Criteria

To compare and rank the merits of the safety databases, it was necessary to develop evaluation criteria. While some research has evaluated and ranked specific aviation databases (Wilke et al., 2014), to the best of the authors' knowledge, there exists no published criteria for evaluation of sociotechnical system safety databases. Therefore, in this paper we propose two overarching criteria—*searchability* and *information fidelity*—to assess the various databases.

Searchability refers to the ease of *downloading* and *searching* information from databases. Databases store information that maybe downloadable in multiple file formats. The formats include file extensions such as pdf, Microsoft Excel and comma separated variable (.xlsx, .csv), database files (.dbf and .mdb), text-based files (.txt, .rtf, and .docx), and PC-axis format (.px).

To leverage the information contained in the databases, it is essential that the data source facilitate ease of *searching* for relevant data. Some databases have *search interfaces* that allow the user to query specific information through dropdown menus and keyword searches. Other databases provide a *list of incidents with links to reports*, where specific information for each incident can be gathered by usually clicking on a link to an associated investigation report. In some cases, however, databases are not user-friendly as they only provide a *list of different incidents on their webpages*, making it challenging to search for information.

The *information fidelity* criterion was used to assess the level of information contained in the databases. This criterion, in turn, has five sub-criteria: (i) summary statistics; (ii) summary of events; (iii) coded information; (iv) final investigation reports; and, (v) causal factors. Some databases only provide *summary statistics* of the accidents recorded. Such statistics, while providing information for users to study potential patterns and trends of the incidents, do not offer insights into accident causation mechanisms. Databases that provided *summary of events* allow users to quickly understand the different events that transpired to results in an accident. The lengths of the summaries can vary from a few sentences to paragraphs based on several factors including investigation depth and information available at the time the summary was prepared. *Coded information* facilitates comprehensive analysis of the contents of the databases to carry out multi-year analyses of accidents to understand root causes and event chains (e.g., Rao & Marais, 2018). The *final investigation reports* supplement the information from coded data and summaries by capturing detailed accounts from investigators, eye witnesses, and survivors, thereby facilitating a holistic understanding of an accident. Finally, some databases report *causal factors* for accidents. Since the causes that are reported in the database are highly dependent on investigation method and depth, this sub-criterion is generalized and the causal factors could be a direct cause, intermediate causes, root causes, or contributing factors. A database satisfies this sub-criterion as long as the causal

factors of the accidents are explicitly stated in the database or indicated separately in the accident narratives.

Database Scoring and Ranking

A simple unweighted sum of searchability and information fidelity criteria was used to compute the score (S_i) for each database. This approach uses the "equal weights approach", which assumes that all criteria are of equal importance (Wang, Jing, Zhang, & Zhao, 2009). After the scores for each of the databases are computed, the values are normalized by the difference between the maximum (S_{max}) and minimum (S_{min}) scores, such that the normalized value is between 0 and 1. The expression to compute the normalized score is as shown in Eq. 1.

$$S_{i,normalized} = \frac{S_i - S_{min}}{S_{max} - S_{min}} \quad (1)$$

Based on the normalized scores, databases were assigned ranks. For example, if the National Transportation Safety Board (NTSB) aviation database had a normalized score of 1, then it would be ranked first.

RESULTS

Using the inclusion and exclusion criteria mentioned in the previous section, 29 databases were identified for review. Table 2 presents the different databases, organizations that maintain them, geographical distribution, evaluation criteria and scores, and ranking. It is important to note that while some databases were exclusive to a domain (e.g., NTSB aviation database), others reported accidents across multiple domains (e.g., Australian Transport Safety Databases)

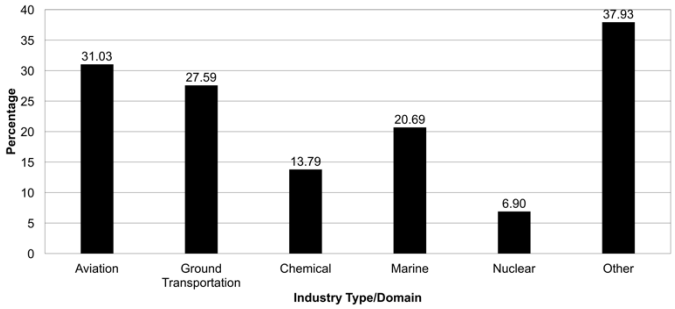

Figure 1. Proportion of databases reviewed that belong to each industry type/domain.

Figure 1 presents the distribution of databases across the different industry types. Note that percentages do not necessarily need to sum to a 100% as a single database can report accidents in more than one domain. Databases relating to aviation safety had the highest presence, accounting for 31.03% (9/29) of the databases reviewed. These include the NTSB aviation accident and incident database, FAA accident and incident data systems, and NASA's aviation safety reporting system. Ground transportation databases had the second highest presence with 27.59% (8/29) of reviewed databases. Databases included in the "Other" category (37.93%; 11/29) include the OSHA database, hydrogen incident and accident database, and the BSEE offshore incident database.

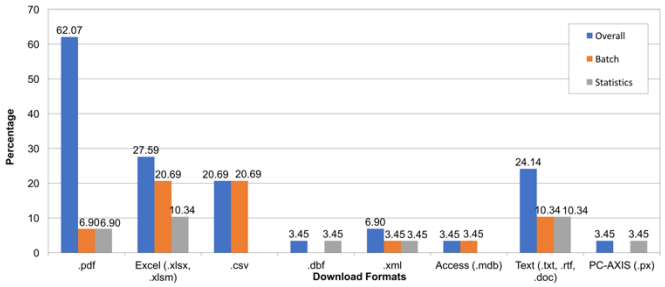

Figure 2. Proportion of download formats for the databases reviewed.

Across most databases (62.07%; 18/29), information could be downloaded in the pdf format, as illustrated in Figure 2. In 27.59% (8/29) of the databases reviewed, data could be downloaded in the form of Excel sheets, while 24.14% (7/29) of databases provided text-based information. Only a small proportion of databases (20.69%) permitted batch-downloads of data in the form of Excel sheets or csv files, while an even smaller proportion (10.34%) allowed batch downloads of text-based data.

As mentioned in the preceding section, to enable learning, it is essential that the data source facilitate searching of relevant information. Of the 29 databases reviewed, it is encouraging to note that 76% (22/29) had search interfaces. Unfortunately, only 17% (5/29) provided incidents with associated pdf reports. Additionally, 7% (2/29) of the data sources required a manual search of their web page to extract accident information.

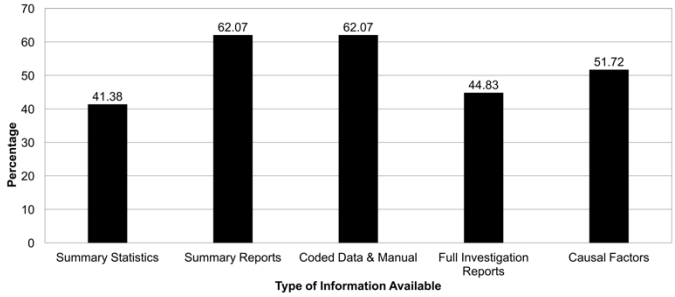

Figure 3. Proportion of nature of information contained in the databases reviewed.

Figure 3 presents the distribution of the five sub-criteria under *information fidelity*. Similar to the previously reported results, the sum of the percentages can exceed a 100% as each database can contain multiple types of information. It is encouraging to note that 62.07% (18/29) databases offer summary reports and data in coded format. However, less than half of the databases provide summary statistics (41.36%; 12/29) and full investigation reports (44.83%; 13/29). A little over half the databases (51.72%; 15/29) explicitly state the causal factors in accidents, as shown in Figure 3.

Table 2. Criteria-based scoring and ranking of safety databases reviewed

Database	Organization	Geographical Distribution	Searchability Score	Information Fidelity Score	Normalized Score	Rank
NTSB Aviation Accident Database	National Transportation Safety Board	Primarily U.S.	5	4	1.00	1
Australian Transport Safety Databases	Australian Transportation Safety Board	Australia	4	4	0.89	2
BSEE Offshore Incident Database	Bureau of Safety and Environmental Enforcement	U.S. Outer Continental Shelf	5	3	0.89	2
Major Accident Reporting System (eMARS)	European Commission	Europe	4	3	0.78	3
European Railway Accident Information Database - European Railway Association	European Railway Association	Europe	4	3	0.78	3
Aviation Safety Reporting System	National Aeronautics and Space Administration	Primarily U.S. and Canada	4	2	0.67	4
Aviation Safety Network	Aviation Safety Network	Worldwide	2	4	0.67	4
Fatality Analysis Reporting System	National Highway Traffic Safety Administration	U.S.	4	2	0.67	4
The International Road Traffic and Accident Database Road Safety Database	The International Traffic Safety and Analysis Group	Worldwide	5	1	0.67	4
Railroad Accident/Incident Reporting System	Federal Railroad Administration	U.S.	2	3	0.56	5
Licensee Event Reports	US Nuclear Regulatory Commission	U.S.	3	1	0.44	6
European Marine Casualty Information Platform	European Maritime Safety Agency	Europe	2	2	0.44	6
Aviation Accidents.net	Privately compiled	Worldwide	2	2	0.44	6
Incident and Accident Database	Japan Transport Safety Board	Japan	2	2	0.44	6
Incident and Accident Database	Accident Investigation Board Norway	Norway	2	2	0.44	6
PHMSA Hazmat Incident Database	Pipeline and Hazardous Materials Safety Administration	U.S.	2	2	0.44	6
FAA Accident and Incident Data Systems	Federal Aviation Administration	Primarily U.S.	2	2	0.44	6
Marine Accident Investigation Branch	Marine Accident Investigation Branch	Primarily U.K.	2	2	0.44	6
Database of Radiological Incidents & Related Events	Privately compiled	Worldwide	1	2	0.33	7
Relational Information System for Chemical Accidents Database	Japan Science and Technology Organization	Worldwide	1	2	0.33	8
Chemical Accidents Investigation System	US Chemical Safety Board	U.S.	2	1	0.33	8
Accidents reports by mode	National Transportation Safety Board	Primarily U.S.	2	1	0.33	8
Fatality and Catastrophe Investigation Summaries	Occupational Safety and Health Administration	U.S.	1	2	0.33	8
Incident / Accident Data from Gas Distribution, Gas Gathering, Gas Transmission, and Hazardous Liquids and LNG Operators	Pipeline and Hazardous Materials Safety Administration	U.S.	1	2	0.33	8
UK Hydrocarbon Releases System	Health and Safety Executive	U.K.	2	1	0.33	8
IADC Incidents Statistics Program	International Association of Drilling Contractors	Worldwide	2	1	0.33	8
Database of nuclear and radiological incidents	Laka Foundation	Primarily Europe	1	1	0.22	9
Mariners Alerting and Reporting Scheme (MARS)	The Nautical Institute	Worldwide	1	1	0.22	9
H2Tools	Pacific Northwest National Laboratory and US Department of Energy	Unknown	1	1	0.22	9

After reviewing the databases, corresponding *searchability* and *information fidelity* scores were computed, normalized, and ranks were assigned (Table 2). The NTSB aviation database, BSEE offshore database, and international road traffic databases had the highest *searchability* scores (5), while databases maintained by the Australian Transportation Safety Board (ATSB), Aviation Safety Network, and NTSB had the highest *information fidelity* scores (4). After normalizing the scores and sorting them in descending order, the NTSB aviation database emerged as top-ranked database based on the review criteria. There was two-way tie for the second and third ranked databases, as shown in Table 2.

DISCUSSION AND CONCLUSION

This paper conducted multi-criteria review of 29 publicly available safety databases spanning eight different sociotechnical domains. The databases were scored using an unweighted arithmetic sum, which was normalized for subsequent ranking.

The findings from this paper show that aviation industry's attention to its safety record has resulted in superior data collection and warehousing, relative to other sociotechnical domains. The high *searchability* and *information fidelity* scores for the NTSB and ATSB aviation accident databases are testament to this fact. These aviation databases provide an array of download formats and information types that can facilitate analysis ranging from a detailed case study on one accident to multi-year trend and root cause analyses of several thousand accidents. Citing the breadth and depth of information in aviation reporting, researchers in other sociotechnical domains (e.g., healthcare) have called for the adoption of safety measures and metrics used in the aviation domain (Kapur, Parand, Soukup, Reader, & Sevdalis, 2015).

The review also shows that databases varied in scope and the level of information provided. The diversity in scope and depth could be attributed, in part, to the theoretical interests or practical requirements of recording organizations. Our findings suggest that databases maintained by federally mandated, primary organizations (e.g., NTSB, ATSB) tend to have a higher ranking relative to those that are maintained by secondary, private entities—a finding echoed by Wilke et al. (2014). Some of the likely reasons for this systematic data collection, storage, and dissemination include the utilization of pre-defined, structured accident investigation and modeling techniques coupled with investigative and computing resources. In contrast, the secondary entities generally aggregate data from primary sources.

While there has been limited research on harmonizing safety databases, lessons can be learned from the healthcare domain. Initial research on integrating healthcare databases has shown promise (Trifirò, 2014). Avillach and colleagues (2012) described the procedure of harmonizing Electronic Health Record data from eight European databases to extract information relating to high-risk medical events. This iterative harmonization process used a code-based algorithm, which facilitated the homogeneous identification of adverse events from different databases. Burnstead & Furlan (2013) argued that unifying drug safety and clinical databases were essential to maintain data consistency and provide a platform for continuous access to data, thereby facilitating pharmacovigilance.

The findings from this review facilitate the initial integration of safety databases. As a first step towards creating a lexical and semantic database, the authors propose to integrate aviation and offshore oil & gas safety databases. This effort will leverage previous work that conceptualizes a data network for harmonizing databases (Shetty, Avnet, & Sasangohar, 2018). Specifically, the data network will enable work with structured and unstructured data by integrating multiple relational and NoSQL databases, so users can perform integrated queries across previously disparate sources. Further, a common industry language will be elicited and applied to avoid integrating incomparable datasets.

In future research, enhancements to the scoring system will be investigated. Specifically, the current scoring system is unweighted, which might not necessarily be a valid assumption.

REFERENCES

Avillach, P., Coloma, P. M., Gini, R., Schuemie, M., Mougin, F., Dufour, J. C., ... & Molokhia, M. (2012). Harmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project. *Journal of the American Medical Informatics Association*, 20(1), 184-192.

Burnstead, B., & Furlan, G. (2013). Unifying drug safety and clinical databases. *Current drug safety*, 8(1), 56-62.

Cowlagi, R. V., & Saleh, J. H. (2015). Coordinability and consistency: Application of systems theory to accident causation and prevention. *Journal of Loss Prevention in the Process Industries*, *33*, 200–212. doi: 10.1016/j.jlp.2014.12.004

Hollnagel, E. (2004). *Barriers and accident prevention*. Aldershot, UK: Ashgate.

Hopkins, A. (2001). Was three mile island a 'normal accident'?. *Journal of contingencies and crisis management*, 9(2), 65-72. doi: 10.1111/1468-5973.00155

Kapur, N., Parand, A., Soukup, T., Reader, T., & Sevdalis, N. (2015). Aviation and healthcare: a comparative review with implications for patient safety. JRSM open, 7(1), 1–10. doi: 10.1177/2054270415616548

Margaryan, A., Littlejohn, A., & Stanton, N. A. (2017). Research and development agenda for Learning from Incidents. *Safety science*, 99, 5-13.

Paté-Cornell, M. E. (1993). Learning from the piper alpha accident: A postmortem analysis of technical and organizational factors. *Risk Analysis*, 13(2), 215-232. doi: 10.1111/j.1539-6924.1993.tb01071.x

Rao, A. H., & Marais, K. (2018). High risk occurrence chains in helicopter accidents. Reliability Engineering & System Safety, 170, 83-98. Doi: 10.1016/j.ress.2017.10.014

Saleh, J. H., Marais, K. B., Bakolas, E., & Cowlagi, R. V. (2010). Highlights from the literature on accident causation and system safety: Review of major ideas, recent contributions, and challenges. *Reliability Engineering & System Safety*, 95(11), 1105–1116. doi:10.1016/j.ress.2010.07.004

Saleh, J. H., Haga, R. A., Favarò, F. M., & Bakolas, E. (2014). Texas City refinery accident: Case study in breakdown of defense-in-depth and violation of the safety–diagnosability principle in design. *Engineering Failure Analysis*, 36, 121-133. doi: 10.1016/j.engfailanal.2013.09.014

Salmon, P. M., Read, G. J., Walker, G. H., Goode, N., Grant, E., Dallat, C., ... & Stanton, N. A. (2018). STAMP goes EAST: integrating systems ergonomics methods for the analysis of railway level crossing safety management. *Safety science*, 110, 31-46. doi: 10.1016/j.ssci.2018.02.014

Shetty S., Avnet M.S., Sasangohar F. (2018). System Safety Data Network: Architecture and Blueprint. In: Madni A., Boehm B., Ghanem R., Erwin D., Wheaton M. (eds) *Disciplinary Convergence in Systems Engineering Research*. Switzerland: Springer, Cham

Shrivastava, P. (1987). Preventing industrial crises: the challenges of Bhopal. *International Journal of Mass Emergencies and Disasters*, 5(3), 199-221.

Trifirò, G., Coloma, P. M., Rijnbeek, P. R., Romio, S., Mosseveld, B., Weibel, D., ... & Sturkenboom, M. (2014). Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? *Journal of internal medicine*, 275(6), 551-561. doi: 10.1111/joim.12159

Wang, J. J., Jing, Y. Y., Zhang, C. F., & Zhao, J. H. (2009). Review on multi-criteria decision analysis aid in sustainable energy decision-making. Renewable and sustainable energy reviews, 13(9), 2263-2278. doi: 10.1016/j.rser.2009.06.021

Wilke, S., Majumdar, A., & Ochieng, W. Y. (2014). A framework for assessing the quality of aviation safety databases. *Safety Science*, 63, 133-145. doi: 10.1016/j.ssci.2013.11.005