Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen

doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P0478-cd

Framework for an Open-Access Dynamic Accident Knowledge Graph Platform for the Critical Infrastructures in Process Industry

Shuo Yang

DISAT, Politecnico di Torino, Italy. E-mail: shuo.yang@polito.it

Micaela Demichela

DISAT, Politecnico di Torino, Italy.

Maria Chiara Leva

Food Science & Environmental Health, Technological University of Dublin, Ireland.

Jie Geng

Department of Government and Regulation, Zhejiang University of Finance & Economics, China.

HuXiao Shi

DISAT, Politecnico di Torino, Italy.

The resilience of Critical Infrastructures (CIs) in the process industry is fundamental to the common good of the environment, economy, and society. Hazard identification is the starting point of risk and resilience management. However, state-of-the-art techniques like HAZOP and FMEA highly depend on the expert group's subjective judgment. Business managers complain about the multiple safety certificates required from different regulators, as the hazard identifications from various groups are inconsistent and do not recognize each other.
There are already many established accident report databases. However, they only provide information and other knowledge about the accident hazards data and are usually unstructured. Thanks to the development of large language models, this research aims to propose a framework to connect and extract evidence-based hazard information from these databases and keep updating the data at season intervals. The extracted data could be structured into a dynamic Knowledge Graph (KG) according to the different functions of CIs and the category of the hazards. An example of KG based on the Major Accident Reporting System (eMARS) dataset is developed as a case study.

Keywords: Critical infrastructure, resilience, hazard identification, knowledge graph.

1. Introduction

Strengthening the resilience of Critical Infrastructures (CIs) in the process industry during disruptions is more challenging than ever. CIs are increasingly complex and interdependent, and they should be viewed in connected scenarios rather than as isolated pieces of equipment. Recent studies (e.g. Twumasi-Boakye, R., & Sobanjo, J. 2019) show a clear need for developing resilience capacities and indexes demonstrated on real-life regional network models. In addition to hazard identification, resilience decision-making requires more intuitive data demonstration for possible deviations.

Hazard identification provides the foundation knowledge for risk and resilience management. However, current popular hazard identification techniques in the process industry, such as HAZOP, are often isolated and rely heavily on the subjective judgment of expert groups. Business managers complain about the numerous safety certifications required from various regulators, as hazard identifications from different groups are inconsistent and do not recognize one another.

The accident dataset could provide evidence for hazards and identify connected scenarios. Regulators increasingly use accident report databases to share previous experiences and lessons learned. These online databases are well-established for unstructured investigation reports or basic category data (e.g., eMARS by the EU, Chemical Safety Board (CSB) Incident Reports by the U.S., and ARIA Database by France). However, they are limited to essential information-sharing functions. Integrating ontology-based knowledge graph (KG) into accident data analysis could offer a transformative approach to understanding and managing process industrial safety concerning CIs.

2. Literature Review

2.1. *Hazard Identification*

Indices are widely used for hazard identification in the process industry. Some focus on substance hazards, like the Dow Chemical fire and explosion index(Company 1967). The safety-weighted hazard index(SWeHI) not only identifies and ranks hazards but also indicates corresponding safety measures(Khan, Husain, and Abbasi 2001). The draw of static checklist and cue word methods is their possibility of limiting creative divergence(Pasman, Rogers, and Mannan 2018).

System Theoretic Process Analysis(STPA) based on an extended modal of accident causation gave hazard identification technique a systematic angle(Leveson and Thomas, 2018). However, this method is quite time-consuming. Pasman, Rogers, and Mannan reviewed the history of hazard identification methodologies and compared them to accident investigation methods. An important suggestion from their work was to fully utilize the potential of IT technologies(Pasman, Rogers, and Mannan 2018). Additionally, KG is a fundamental step in the digitization and computer understanding of real-world conceptual hierarchy systems.

2.2. *Knowledge Graph*

The KG concept was proposed by Google in 2012. The definition of KG G is composed by a series of entities E(in form of nodes), a series of relational links R(in form of edges), and the facts F they represent ($Entity_{head}$, Relation, $Entity_{tail}$)(Zhang et al. 2025). There is a consensus that KG is an efficient knowledge representation and storage tool that supports rational and propagation processes among people or computers.

Early research on KG of accidents and hazards mainly focuses on the construction and transportation industry (Fang et al., 2020, Liu et al., 2021). Peng et al. collected data from structured inspection reports and unstructured text sources to build a KG regarding utility tunnels(Peng et al.2023). Additionally, Huo et al. employed a data-driven approach to create a KG concerning subway construction accidents(Huo et al.2024). Zhang et al. incorporated real-time information into knowledge graphs for dynamic hazard analysis and ensured their work was updated based on changes in site conditions. They also developed metrics to evaluate system hazards at micro, meso, and macro levels(Zhang et al.2025). Hong et al. proposed an intelligent ontology to evaluate accident risks on construction sites utilizing a Natural Language Processing (NLP)-based framework(Hong et al. 2024).

The above-mentioned research validated the feasibility of applying KG in hazard identification and accident analysis. However, only a few works were given to the guild on KG regarding accidents in the process industry. Mao et al. developed KG on the delayed coking process(Mao et al. 2020). Wang et al. built KG of the indirect coal liquefaction process for the industry design stage(Wang et al. 2022). Xue et al. proposed a KG model for process accidents in Chinese(Xue et al. 2025). These results mostly focused on some special scenarios. Therefore, this research aims to propose a more transformative KG model framework for CIs in the process industry.

3. Methodology

As shown in Fig.1., the holistic framework includes four steps. Step 1 is building the general ontology for accidents in the process industry. This step is mainly based on the "4Ws" narrative framework proposed in our early work(Yang and Demichela 2023).

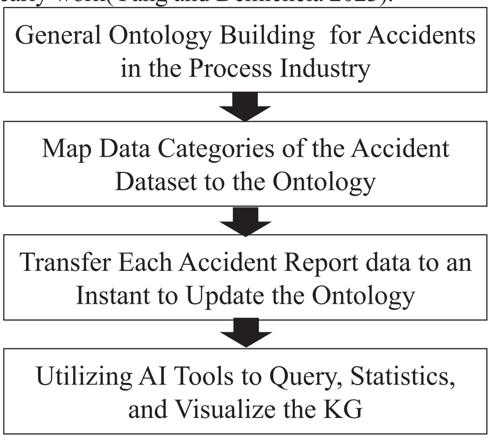

Fig. 1. The holistic framework.

Step 2 maps the accident dataset categories to the ontology classes and properties. This step needs to carefully check the categories of the target accident dataset, considering the different data types. Step 3 transfers each accident report data to an instant to update the ontology. This step needs to utilize data mining tools to clean and process data. Integration tools are also required to incorporate the instants into the ontology. Step 4 is mainly about the KG building and demonstrating. The case study part will explain the framework step by step.

4. Case Study and Results

The eMARS dataset[a] was chosen for the case study because it includes multiple process industries and is open-access.

4.1. *Ontology building for accidents in the process industry*

This step could use the owlready2 package in Python or utilize the ontology design tool Protégé developed by Stanford University (Musen 2015). This research chooses the Protégé for its better ability at visualization. The whole ontology is centered on the concept of an accident, as shown in Fig.2.

The proposed Accident Ontology(A_Onto1.0) has 37 classes under the default class thing, as shown in Fig.3., including Event, Accident, Time, DateTime, Duration, Location, ProcessEquipment, StorageEquipment, TransferEquipment, TransportEquipment, People, Operator, Contractor, FactoryStaff, Cause, HumanFactor, OrganizationFactor, PlantEquipmentFacotor, Occurrence, Fire, Release, Explosion, Consequence, EnvironmentPollution, HumanHealthHarm, Fatality, Injury, Substance, HazardSubstance, Hazard, Explosive, Flammable, Oxidizing, Pyrophoric, Toxic, and SelfReactive.

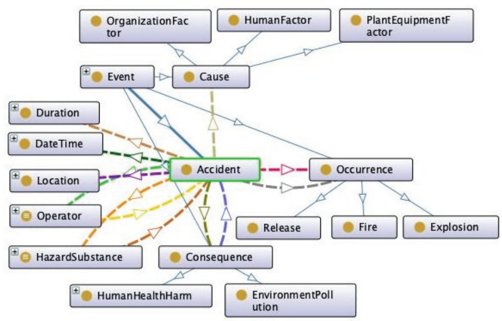

Fig. 2. The main ontology graph focuses on the Accident class.

Fig.4. shows the data properties list. The Accident class has four data properties: eid, title, and type. Duration class has duration_in_days, DateTime class has start_date. EnvironmentPollution, Explosion, Fire, Release, HumanFactor, OrgnationFacor, PlantEquipmentFactor, ProcessEquipment, StorageEquipmen, TransferEuipment, and TransportEquipment all have _type data properties.

Fig.5. shows the object properties list. These are the relationships between classes. The HazardSubstance class connects to the Hazard class with the has_hazard object property. The Accident class connects to the Occurrence class with the has_occurrence

[a]https://emars.jrc.ec.europa.eu/en/emars/accident/search,accessed at 2025.01.14

object property. The Accident class connects to the HazardSubstance class with the involve object property. The Accident class connects to the Operator class with the is_performed_by object property. The Accident class connects to the Consequence class with the lead object property. The Accident class connects to the Cause class with the is_triggered_by object property. And others are inverse object properties.

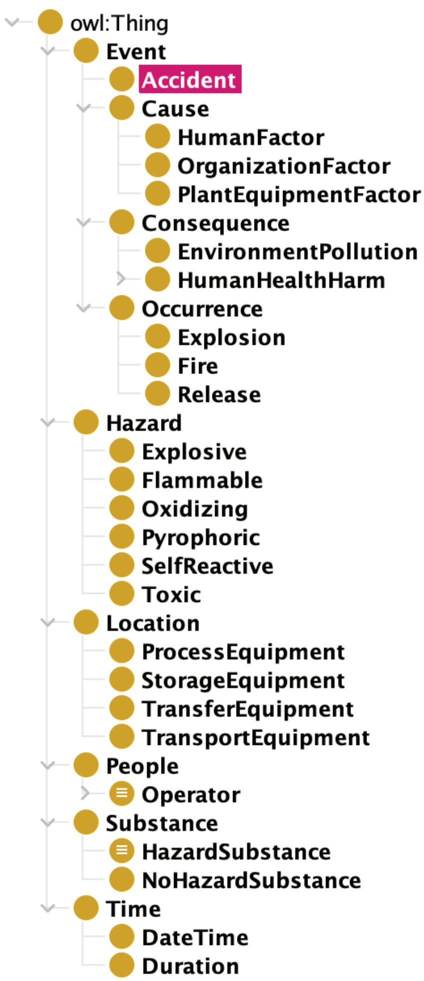

Fig. 3. The tree views of the ontology classes.

4.2. *Mapping the Accident Dataset Categories to the Ontology*

Table 1 illustrates the corresponding mapping between eMARS categories and the A_Onto1.0.

4.3. *Transfering each accident report data to an instant to update the ontology*

As involving an extensive amount of data processing, in this step, the Pandas and Owlready2 packages in Python are chosen to perform the tasks. Owlready2 is employed to add the instants into the build-up ontology and update it to A_onto2.0 with the eMARS data.

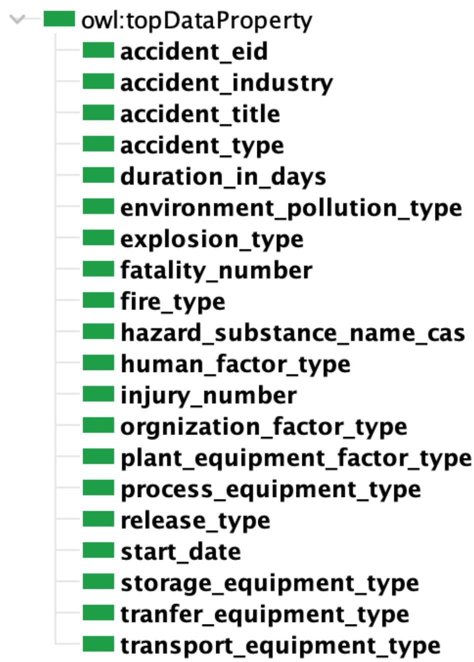

Fig. 4. The tree views of the data properties.

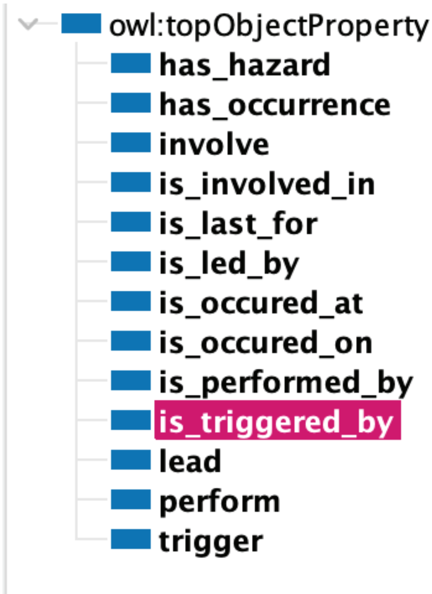

Fig. 5. The tree views of the object properties.

Table 1. eMARS Categories map to A_Onto1.0.

eMARS Categories	A_Onto1.0	
Accident ID	A_id	
Accident Title	A_title	
Event type	A_type	
Industry Type	A_industry	
Contractors	A	is_performed_by
StartDate	A	is_occured_on
Release Major Occurrences	A	trigger
Fire Major Occurrences	A	trigger
Explosion Major Occurrences	A	trigger
Storage major occurrence	A	is_occured_at_equipment
Process major occurrence	A	is_occured_at_equipment
Transfer major occurrence	A	is_occured_at_equipment
Transport major occurrence	A	is_occured_at_equipment
Toxic	H	has_hazard
Explosive	H	has_hazard
Flammable	H	has_hazard
Oxidizing	H	has_hazard
Self reaction	H	has_hazard
Pyrophoric	H	has_hazard
Substance	A	involve
Plant/Equipment causative factor type	A	is_caused_by
human causative factor type	A	is_caused_by
Organizational causative factor type	A	is_caused_by
Human off + on site fatalities	A	lead
Human off +on site injuries	A	lead
Environment on +off site quantity	A	lead

*A represent Accident class, H represent HazardSubstance class

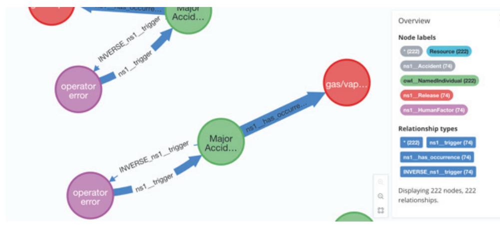

Fig. 6. The queried results for in Neo4j

4.4. *KG building and demonstrating*

The KG building tool Neo4j Desktop demonstrates the KG of A_Onto2.0. Fig.7. shows a small part of the KG, the different colors represent different types of classes, and the piece of critical infrastructure(CI) is highlighted with pink. and perform query function to find the HumanFactor is “operator error” and has release occurrence, results are shown in Fig.6.

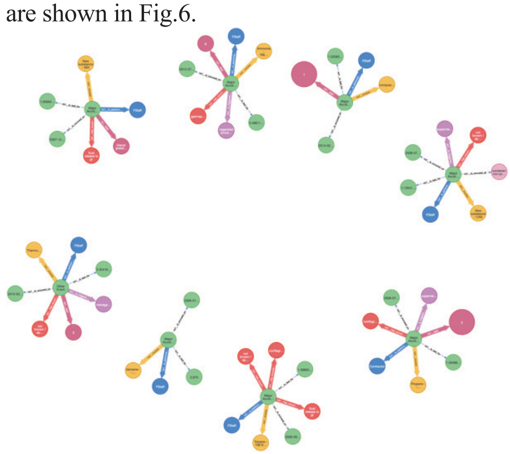

Fig. 7. The small part of KG for A_Onto2.0 in Neo4j

Table 2. shows the query results of the most frequent properties. The “uri” is the total amount of properties.

Therefore, it is in the first place. The frequency of accident_eid is the instant count. Followed by "involve_hazard_substance" and "plant_equipment_factor_type". Based on this frequency, the KG relationship strength could be updated with the corresponding coefficient calculated by the frequency percentage.

Table 2. Most Frequent properties in A_onto2.0.

Property	Frequency
uri	11058
accident_eid	1230
involve_hazard_substance	993
plant_equipment_factor_type	633
release_type	625
orgnization_factor_type	516
injury_number	373
fire_type	339
human_factor_type	277
fatality_number	239
explosion_type	236
process_equipment_type	234
storage_equipment_type	158
tranfer_equipment_type	95

5. Discussions

This research is only a starting point for the ontology and knowledge graph design for hazard information from accident reports. There are some limitations of this work, such as the ontology needing more validation from different data sources and formats. With data from eMARS, we could calculate the frequency of the leading cause factor and the probability of the event having different consequences. However, eMARS is not designed to provide statistical analysis, which needs to focus on a unique industry area or fix scenario. Therefore, the framework could be applied within a company range to do quantitative analysis with internal operational record data. The critical structure of the proposed ontology of accidents could transfer to guide the target information contents. Accident reports in PDF format in a storytelling way could combine the ontology with the large language model or natural language process technical to build the knowledge graph.

6. Conclusions

This research proposed a framework to build ontology-based KG for accidents in process industry-related hazard scenarios based on accidents. And the eMARS dataset is selected to perform a case study. With the help of a visualization dynamic database tool Neo4j, the KG for A_onto2.0 was built with 1230 instants.

By structuring data into interconnected entities and relationships, KGs provide a holistic view of complex accident scenarios, facilitating advanced reasoning and inference to uncover hidden patterns and cascading effects. This method enhances data integration across diverse sources, delivering a unified framework for hazard identification and risk assessment. Furthermore, KGs enable dynamic querying and visualization, thereby improving decision-making processes and supporting proactive safety management. Their scalability and adaptability make KGs a powerful tool for tackling evolving industrial safety challenges.

References

Company, Dow Chemical. 1967. *Dow's Process Safety Guide*. American Institute of Chemical Engineers.

Hong, Eunbin, SeungYeon Lee, Hayoung Kim, JeongEun Park, Myoung Bae Seo, and June-Seong Yi. 2024. 'Graph-Based Intelligent Accident Hazard Ontology Using Natural Language Processing for Tracking, Prediction, and Learning'. *Automation in Construction* 168 (December):105800. https://doi.org/10.1016/j.autcon.2024.105800.

Huo, Xiaosen, Yuan Yin, Liudan Jiao, and Yu Zhang. 2024. 'A Data-Driven and Knowledge Graph-Based Analysis of the Risk Hazard Coupling Mechanism in Subway Construction Accidents'. *Reliability Engineering & System Safety* 250 (October):110254. https://doi.org/10.1016/j.ress.2024.110254.

Khan, F.I., T. Husain, and S.A. Abbasi. 2001. 'Safety Weighted Hazard Index (SWeHI)'. *Process Safety and Environmental Protection* 79 (2): 65–80. https://doi.org/10.1205/09575820151095157.

Leveson, Nancy. n.d. 'This Handbook Is Intended for Those Interested in Using STPA on Real

Systems. It Is Not Meant to Introduce the Theoretical Foundation, Which Is Described Elsewhere. Here Our Goal Is to Provide Direction for Those Starting out with STPA on a Real Project or to Supplement Other Materials in a Class Teaching STPA.'

Liu, Jintao, Felix Schmid, Keping Li, and Wei Zheng. 2021. 'A Knowledge Graph-Based Approach for Exploring Railway Operational Accidents'. *Reliability Engineering & System Safety* 207 (March):107352. https://doi.org/10.1016/j.ress.2020.107352.

Mao, Shuai, Yunmeng Zhao, Jinhe Chen, Bing Wang, and Yang Tang. 2020. 'Development of Process Safety Knowledge Graph: A Case Study on Delayed Coking Process'. *Computers & Chemical Engineering* 143 (December):107094. https://doi.org/10.1016/j.compchemeng.2020.107094.

Pasman, Hans J., William J. Rogers, and M. Sam Mannan. 2018. 'How Can We Improve Process Hazard Identification? What Can Accident Investigation Methods Contribute and What Other Recent Developments? A Brief Historical Survey and a Sketch of How to Advance'. *Journal of Loss Prevention in the Process Industries* 55 (September):80–106. https://doi.org/10.1016/j.jlp.2018.05.018.

Peng, Fang-Le, Yong-Kang Qiao, and Chao Yang. 2023. 'Building a Knowledge Graph for Operational Hazard Management of Utility Tunnels'. *Expert Systems with Applications* 223 (August):119901. https://doi.org/10.1016/j.eswa.2023.119901.

Wang, Zhenhua, Beike Zhang, and Dong Gao. 2022. 'A Novel Knowledge Graph Development for Industry Design: A Case Study on Indirect Coal Liquefaction Process'. *COMPUTERS IN INDUSTRY* 139 (August):103647. https://doi.org/10.1016/j.compind.2022.103647.

Xue, Nannan, Wei Zhang, Huayu Zhong, Wenbin Liao, and Tingsheng Zhao. 2025. 'How to Realize the Knowledge Reuse and Sharing from Accident Reports? A Knowledge-Driven Modeling Method Combining Ontology and Deep Learning'. *Journal of Loss Prevention in the Process Industries* 94 (April):105525. https://doi.org/10.1016/j.jlp.2024.105525.

Yang, Shuo, and Micaela Demichela. 2023. 'A Data-Driven Narratives Skeleton Pattern Recognition from Accident Reports Dataset for Human-and-Organizational-Factors Analysis'. *Journal of Loss Prevention in th Process Industries* 83 (July):105047. https://doi.org/10.1016/j.jlp.2023.105047.

Zhang, Juntong, Xin Ruan, Han Si, and Xiangyu Wang. 2025. 'Dynamic Hazard Analysis c Construction Sites Using Knowledge Graphs Integrated with Real-Time Information'. *Automation in Construction* 170 (February):105938. https://doi.org/10.1016/j.autcon.2024.1059 8.