Safety Science 168 (2023) 106305

Contents lists available at ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/safety

Knowledge in graphs: investigating the completeness of industrial near miss reports

Francesco Simone [a,*], Silvia Maria Ansaldi [b], Patrizia Agnello [b], Giulio Di Gravio [a], Riccardo Patriarca [a]

[a] *Department of Mechanical and Aerospace Engineering, Sapienza University of Rome Via Eudossiana, 18, 00184 Rome, Italy*
[b] *Department of Technological Innovations and Safety of Plants, Products and Anthropic Settlements, INAIL (Italian National Institute for Insurance Against Accidents at Work), Monteporzio Catone, Rome, Italy*

ARTICLE INFO

Keywords:
Industrial plants
Refineries
Safety management
Safety ontology
Computational safety

ABSTRACT

Learning from near misses has a large potential for improving operations especially in high-risk sectors, such as Seveso industries. A comprehensive analysis of near miss reports requires processing a large volume of data from various sources, which are not standardized and seemingly disconnected from each other. A knowledge graph is here used to provide a comprehensive safety perspective to near miss data. In particular, this paper presents an analysis of a knowledge graph for near miss reports with the objective to measure systematically their completeness based on an integrated multi-criteria decision-making technique. The reports completeness fosters a meta-analysis of available data, highlighting systems' strengths and vulnerabilities, as well as disseminating best practices for industry stakeholders.

1. Introduction

The study of major failures is a valuable aspect of any safety management process. Traditionally, these analyses focus on identifying hazards to be anticipated to prevent or to mitigate possible adverse events (Dekker, 2019). However, as severe industrial accidents become less common, it is increasingly important to consider other data sources. In addition to major accidents, there are numerous minor failures, known as "near misses", that have the potential to cause accidents, which did not lead to catastrophic outcomes. Formally, a near miss refers to an event that could have caused an incident, or a damage to health, or even a fatality, but for some reasons, it did not (Phimister et al., 2003). In many cases, near misses go unreported, and their potential value in improving safety or preventing future incidents is lost. Learning from near misses may be crucial for the identification of potential larger system failures, for the analysis of causal factors, and for testing the effectiveness of safety barriers (Bugalia et al., 2021). Additionally, near misses provide an opportunity to learn about potential adaptations and response strategies to prevent the occurrence of more serious events.

On this premises, near misses reporting becomes a proactive tool for safety management systems: if properly documented, near misses can provide information comparable to that obtained from major accidents (Caspi et al., 2023).

The significance of near misses is particularly relevant in critical industrial settings involving the storage, handling, manufacturing, or usage of hazardous substances. These industries are addressed by the EU Seveso III Directive (EU Council, 2012), i.e., a legislation that identifies industrial sectors with major safety, health, and environmental risks involving dangerous chemicals. The Seveso III Directive applies to various industries, such as oil and gas, refineries, chemical and petrochemical, pharmaceuticals, metal processing, and explosives, and it demands manager and practitioners to pay particular attention to near miss events, with periodic inspection made by competent authorities. It appears clearly how safety practices in Seveso establishment may benefit from tools enabling an integrated analysis of near miss data. However, these data, usually represented by text documents, remains hard to manage as they are represented by linguistic data, not standardized, and apparently disconnected from each other. The main problem in this sense is the lack of a comprehensive near miss management system and the subsequent presence of a large amount knowledge that remains tacit (Pedrosa et al., 2022). To cope with this challenge, a knowledge graph is here proposed as a technology for modelling knowledge in complex systems and integrating diverse data sources (Khan et al., 2022).

[*] Corresponding author.
E-mail address: francesco.simone@uniroma1.it (F. Simone).

https://doi.org/10.1016/j.ssci.2023.106305
Received 31 May 2023; Received in revised form 3 August 2023; Accepted 30 August 2023
Available online 12 September 2023
0925-7535/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Nomenclature

$\mathbb{L}$	set of criteria to be evaluated in the AHP
$\mathbb{L}^*$	subset of criteria
$\mathbb{N}$	set of nodes (vertices) in the graph G
$\mathbb{N}^*$	subset of nodes (vertices)
$\mathbb{R}$	set of relationships (edges) in the graph G
H	subset of nodes to be used for the average completeness calculation
η	completeness metric
$\overline{\eta}$	average completeness
λ_{max}	maximum eigenvalue of the pairwise comparison matrix A (and the grouped pairwise comparison matrix B)
A	pairwise comparison matrix, it has dimension $K \times K$
B	grouped pairwise comparison matrix, it has dimension $K \times K$
CR	consistency ratio related to a pairwise comparison matrix A (or a grouped pairwise comparison matrix B)
D	total number of pairwise comparison matrices (and related respondents) to be grouped in B
G	knowledge graph structure containing vertices and edges
I_n	maximum number of properties of the node N_n
J_m	maximum number of properties of the relationship R_m
K	total number of criteria to be evaluated through the AHP
L_k	generic criterion in $\mathbb{L}$
L_n^N	label of the n-th node (N_n)
L_m^R	label of the m-th relationship (R_m)
M	total number of relationships in the graph structure G
N	total number of nodes in the graph structure G
N_n	generic node in $\mathbb{N}$
$N_m^{'}$	node from which the m-th relationship starts
$N_m^{''}$	node from which the m-th relationship ends
R_m	generic relationship in $\mathbb{R}$
RI_K	estimated average consistency for matrices of dimension $K \times K$
$YoY_{\%}$	year over year percentage change of a time-dependent variable $v_{(t)}$
a_{xy}	generic element of the pairwise comparison matrix A
d	identifier for the d-th pairwise comparison matrix (related to the d-th respondent), it varies from 1 to D
e	vector of K elements all equal to 1
i	identifier for the i-th property of the node N_n, it varies from 0 to I_n
j	identifier for the j-th property of the relationship R_m, it varies from 0 to J_m
k	identifier for the k-th criterion in $\mathbb{L}$, it varies from 1 to K
m	identifier for the m-th relationship in $\mathbb{R}$, it varies from 1 to M
n	identifier for the n-th node in $\mathbb{N}$, it varies from 1 to N
$p_{i,n}^N$	i-th property of the n-th node (N_n)
$p_{j,m}^R$	j-th property of the m-th relationship (R_m)
$v_{(t)}$	value of a generic variable v at time t
w	vector of weights related to the K criteria
w_k	generic element of the vector of weights w (i.e., weight of the k-th criterion in $\mathbb{L}$)
x	row index for the pairwise comparison matrix A and the grouped pairwise comparison matrix B, it varies from 1 to K
y	column index for the pairwise comparison matrix A (and the grouped pairwise comparison matrix B), it varies from 1 to K
z	generic index for the pairwise comparison matrix A (and the grouped pairwise comparison matrix B), it varies from 1 to K

Newman (2010) defines a knowledge graph as a structure of objects that are related in pairs and shares a common knowledge which is represented as a simplified network of nodes and edges. To build a knowledge graph, an ontology is required, as a formal and explicit definition of a shared conceptualization (Studer et al., 1998). An ontology provides a semantic data model that describes the types of objects, their properties, and their relationships, all within a certain scope of analysis, as safety reporting may be (Hughes et al., 2019). By combining these concepts, a knowledge graph can be conceived as a large semantic net that brings together various and diverse sources of information to represent available knowledge.

By engineering a knowledge graph, data from multiple sources can be integrated, in agreement with the structure imposed by the underlying ontology, as proven by a wide range of applications (Abu-Salih, 2021). However, the application of knowledge graphs for safety management is still in its early stages and, by now, limited to the storage of data resulting from previous hazards and risks analyses (Li et al., 2021). For example, Zhu et al. (2021) proposed the usage of a knowledge graph to collect risks related to terrorist attacks on LNG storage tanks. However, risks were identified and analyzed through a Bayesian network and event trees, and the graph only served as a tool to visualize hazardous scenarios and their related risks. Similarly, Peng et al. (2023) built a knowledge graph for collecting operational hazards for utility tunnels. Starting from the definition of a custom ontology, the authors merged the data set containing possible hazard description, with the normative in force. The resulting graph relates hazardous scenarios with prescribed avoidance/control measures.

In this work, we present a method to analyze a knowledge graph containing data from near miss reports collected in Seveso industries. The analysis is grounded on the concept of report completeness as a leading indicator of safety management systems performance. Accordingly, the graph has been weighed by subject matter experts, involved as respondents to questionnaires designed based upon multi-criteria decision-making principles.

The research question of this paper can be stated as: "*To which extent knowledge graphs may support a* safety meta-*analysis of completeness for near miss reports in Seveso industries?*". We leverage on the notion of safety meta-analysis to identify a type of research that uses a systematic approach to statistically combine the findings of many near miss reports into a systemic overview of safety reporting systems. The meta-analysis is meant to analyze the tools (in this paper the safety reports) that are connected to the risk and safety analyses, but it has no direct connection with the amount of risk to be averted or reduced.

The article is structured as follows. Section 1 just introduced the background and motivations for this work. Section 2 describes the methodological foundations and tools used for the analysis. In Section 3 the completeness metric is computed by experts' interviews. We present and discuss the results obtained by different perspectives of meta-analysis in Section 4. Finally, Section 5 provides concluding remarks, as well as limitations and potential future developments of the research in the context of safety management.

2. Materials and methods

In this section we present the methodological fundamentals to guide the analysis of near miss reports by means of the completeness of their informative content. The methodology makes use of the knowledge graph technology to model near miss data and extends it through a weighting mechanism to assign them proper relevance. Accordingly, this section presents a formalization of the knowledge graph structure, and the linked notions to ground the safety meta-analysis via a novel completeness metric.

2.1. Knowledge graph formalization

A graph G can be formally defined as a data structure containing nodes (vertices), and relationships (edges):

$$\mathrm{G}(\mathbb{N}, \mathbb{R}) \tag{1}$$

where $\mathbb{N}$ is the set of nodes N_n with $1 < n < N$, and $\mathbb{R}$ is the set of relationships R_m with $1 < m < M$.

Each N_n node is defined as a multi-dimensional object:

$$N_n = \left(L_n^N, p_{i,n}^N\right), \ 0 \le i \le I_n \tag{2}$$

in which L_n^N is the label to be assigned to the n-th node, and $p_{1,n}^N, p_{2,n}^N, \cdots, p_{I,n}^N$ are the properties of the n-th node. As long as a node can have multiple properties or even none, the i index can vary between 0 and I_n.

Similarly, the relationships in $\mathbb{R}$ are defined as:

$$R_m = \left(N_m^{'}, N_m^{''}, L_m^R, p_{j,m}^R\right), \ 0 \le j \le J_m \tag{3}$$

where $N_m^{'} \in \mathbb{N}$ is the node from which the m-th relationship starts, $N_m^{''} \in \mathbb{N}$ is the node to which the m-th relationship ends, L_m^R is the label assigned to the m-th relationship, and $p_{1,m}^R, p_{2,m}^R, \cdots, p_{J,m}^R$ are the properties assigned to m-th relationship. A relationship can have multiple properties or even none, so j ranges between 0 and J_m.

2.1.1. A knowledge graph to model near misses

The knowledge graph used in this research leverages from a previous contribution by Simone et al. (2023), where the reader can find a complete explanation of the knowledge graph structure. This latter originates from the formal definition in Section 2.1, and combine it with the semantic rules of an near misses ontology (Ansaldi et al., 2021). Please note that hereafter, italic capital letters will refer to labels, while italic lower case will refer to properties. Concerning nodes, nine different labels (L_n^N) have been identified. Depending on their label, nodes have different properties ($p_{i,n}^N$). Nodes' labels and resulting properties are briefly summarized below:

- *INDUSTRIAL_SECTOR*: nodes containing data about the industrial sector in which the establishment that redacts the near miss report operates. These nodes have a property *industrial_sector_id* that specifies a unique Seveso-related industrial sector.
- *ESTABLISHMENT*: nodes containing data about the industrial establishment from which the near miss report has been collected. These nodes have three properties: (i) *establishment_id*, that is a unique identifier for a specific industrial establishment; (ii) *location_region*, which contains the region in which the industrial plant operates; (iii) *location_district*, which contains the district within the region in which the plant operates.
- *DOCUMENT*: nodes containing data about a specific near miss report document. Nodes with label *DOCUMENT* have three properties: (i) *document_id*, that is an unique identifier for a specific near miss report and consists in the name of the source pdf file; (ii) *collection_date*, which contains the year in which the near miss report was collected by controlling authorities; (iii) *occurence_date*, which contains the year of occurrence of the near miss that is described in the report.
- *EVENT*: nodes containing data about an event that happened in the near miss. Nodes with label *EVENT* have two properties: (i) *value* which includes the actual word from the pdf file that has been identified to be an event, and (ii) *type*, which specifies the type of event that is described in the report, it can be "Loss", "Failure", "Deterioration", "Major", or "Success".
- *ACTIVITY*: nodes containing data about an activity that was carried out when the near miss happened. Nodes with label *ACTIVITY* have one property, namely, *value* which includes the actual word from the pdf file that has been identified to be an activity.
- *APPARATUS*: nodes containing data about an industrial apparatus that was involved in the near miss. Nodes with label *APPARATUS* have two properties: (i) *value* which contains the actual word from the pdf file that has been identified to be an apparatus, and (ii) *type*, which specifies the type of apparatus that is described in the report, it can be "Equipment", or "Component".
- *SUBSTANCE*: nodes containing data about a specific substance that was involved in the near miss. Nodes with label *SUBSTANCE* have the property *value* which includes the actual word from the pdf file that has been identified to be a substance.
- *PEOPLE*: nodes containing data about the role of a person (or a group of people) who was involved in the near miss occurrence. Nodes with label *PEOPLE* have one property, namely, *value* which includes the actual word from the pdf file that has been identified to refer to human factors.
- *BARRIER*: nodes containing data about a safety barrier (both technical or organizational) that worked (or not) when the near miss happened. Nodes with label *BARRIER* have two properties: (i) *value* which contains the actual word from the pdf file that has been identified to be a barrier, and (ii) *type*, which specifies the type of barrier that is described in the report, it can be "Technical", or "Organizational".

Similarly, seven different labels for relationships (L_m^R) have been identified, no properties have been considered for relationships. Relationships labels can be:

- *BELONGS_TO*: it identifies a connection between an industrial establishment and an industrial sector.
- *FROM*: it maps the relationship between documents and establishment in which they have been redacted/collected.
- *CONTAINS*: it relates all the data (nodes) contained in a document to the corresponding node with label *DOCUMENT*.
- *RELATED_TO*: it is a generic relationship between two nodes contained in a report.
- *PART_OF*: it describes a physical connection between two nodes in the report. Accordingly, it can be used to relate *APPARATUS* and *BARRIER*.
- *INVOLVES*: it relates a node to a node with label *SUBSTANCE*.
- *CAUSES*: it states a causal connection between two nodes, and always points at a node with label *EVENT*.

2.2. Completeness assessment

The completeness of a near miss reports can be seen as an i-th additional property $p_{i,n}^N$ to be assigned to specific nodes, i.e., the ones representing the report itself in the knowledge graph. The knowledge graph structure is able to map a large amount of data taking into account their source, and their content, in principle without apportioning scaled relevance. However, not all data have the same importance for safety reporting. For example: the substance involved may be more relevant than the activity performed when the near miss verified. Based on this assumption, the following paragraphs present a procedure to compute a measure of report completeness. The procedure relies on a multi-criteria decision making technique, namely, the Analytical Hierarchy Process (AHP) to systematically assign weights to different nodes (Saaty, 1990). Then, a metric that depicts the completeness of a collection of nodes (as a near miss report can be) is defined to map the informative content.

2.2.1. Analytical Hierarchy process (AHP)

The AHP is meant to support a decision process in finding the best solution upon a finite set of K criteria $\mathbb{L} = \{L_1, \cdots, L_k, \cdots, L_K\}$, with $1 < k < K$ (please note that the L notation has been used for criteria since the nodes' labels L_n^N will be used as criteria afterwards). Accordingly, experts are asked to assign a score to each criterion, and, subsequently, to alternatives related to that criterion, permitting to choose the ones with higher values. For the set of criteria $\mathbb{L}$, a corresponding vector of K weights $w = \{w_1, \cdots, w_k, \cdots, w_K\}$ must be provided, where w_k is the value which coherently estimates the priority of the criterion L_k by means of a specific goal to be reached. The assignment of weights may be tricky when comparing several criteria at the same time, thus, to facilitate this process, pairwise comparison is used. For a given goal, the pairwise comparison values numerically detail the rate of importance of a criterion over the other. These values are collected in a pairwise comparison matrix A:

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1K} \\ \vdots & \ddots & \vdots \\ a_{K1} & \cdots & a_{KK} \end{pmatrix} \tag{4}$$

that is the $K \times K$ matrix made of the elements a_{xy}, each one depicting the degree of preference of two k-th criteria: the one with $k = x$ (i.e., $L_x \in \mathbb{L}$) over the one with $k = y$ (i.e., $L_y \in \mathbb{L}$). The value of a_{xy} can be seen as an estimation of the ratio of the weights to be assigned at the two pairwise-compared criteria:

$$a_{xy} \approx \frac{w_x}{w_y}; \forall x, y \tag{5}$$

Accordingly, the matrix A from Eq. (4) can be simplified as:

$$A = \begin{pmatrix} 1 & \cdots & a_{1K} \\ \vdots & \ddots & \vdots \\ 1/a_{1K} & \cdots & 1 \end{pmatrix} \tag{6}$$

Because of Eq. (5) holding true: (i) there will be no preference upon a criterion and itself, and (ii) the importance rate of a first criterion over a second one will be the inverse of the importance rate of the second one over the first, coherently. Pairwise comparisons may result in inconsistency for the evaluation of singular weights. Specifically, for a triplet of criteria $L_x, L_y, L_z \in \mathbb{L}$, the following condition should be verified to ensure consistency:

$$a_{zy} = a_{xy} a_{yz}; \quad \forall x, y, z \tag{7}$$

Or, in other words, that the maximum eigenvalue λ_{max} of the pairwise comparison matrix A is equal to the dimension of the matrix itself (i.e., $\lambda_{max} = K$). Accordingly, the consistency ratio (CR) can be evaluated to check the trustworthiness of elements in A:

$$CR = \frac{1}{RI_K} \cdot \frac{\lambda_{max} - K}{K - 1} \tag{8}$$

where RI_K is the estimated average consistency obtainable from a large set of $K \times K$ pairwise matrices. In practical applications, the threshold condition $CR \leq 0.1$ has been commonly accepted to depict consistent judgments (Brunelli, 2015).

The last step consists of aggregating the relative weights coming from the different elements a_{xy}, to find the components of the vector of criteria weights w. Above all, the eigenvector method is simple yet effective, identifying the priority vector w in the principal eigenvector (the one with highest eigenvalue) of A. Thus, given a pairwise matrix A, the vector w results from:

$$\begin{cases} Aw = \lambda_{max} w \\ w^T e = 1 \end{cases} \tag{9}$$

where e is the vector of size K defined as $e = (1, \cdots, 1)^T$.

Please note that the concepts presented above can be extended to any number of respondents (i.e., multiple pairwise comparison matrices) by deriving a grouped pairwise comparison matrix B with elements:

$$b_{xy} = \prod_{d=1}^{D} a_{xy}{}^{d} \tag{10}$$

where $1 < d < D$ identifies the d-th matrix compiled by the d-th decision maker, who is part of the group of D people. Please note that both B and A are $K \times K$ matrices.

Summarizing: one (or a group of D) expert(s) is asked to compare couples of criteria $L_x, L_y \in \mathbb{L}$. Consequently, the responses of each d-th expert fill a pairwise decision matrix A. Multiple matrices can be aggregated through the grouped pairwise comparison matrix B. The condition $CR \leq 0.1$ must be verified to ensure consistency among data, and the vector of weights w can be computed defining priorities among the given criteria. Similarly, the process is repeated for the lower level of the multi-criteria decision problem by evaluating alternatives related to each criterion. Finally, alternatives' priorities are computed averaging over the weights of correspondent criteria. In this paper we assume all alternatives to be equally important in reaching the goal of the problem, this assumption is further detailed in Section 3.

2.2.2. Completeness metric

The vector of weights w is used to define a completeness metric for reports. A near miss report is seen as a collection of elements, each one with its corresponding weight. Specifically, the different labels that can be assigned to a node represent the criteria to be evaluated in terms of their contribution to make a report complete. In this perspective, the set of possible criteria within the whole graph becomes:

$$\mathbb{L} = \{L_k = L_n^N \, \forall n = 1, \cdots, N \, : L_{n'}^N \neq L_{n''}^N \, , \, n', n'' = 1, \cdots, N\} \tag{11}$$

that is the set including only the unique values among all the n-th labels assigned to the N nodes in $\mathbb{N}$. Being $\mathbb{L}$ the set of criteria, each element L_k will be uniquely related to one exact element in w ($L_k \leftrightarrow w_k$). Please note that both w and $\mathbb{L}$ will count the same number of elements, i.e., K. Similarly, one can define the unique values in a subset $\mathbb{N}^* \subseteq \mathbb{N}$ as $\mathbb{L}^*$. This latter includes all the different labels of nodes contained in $\mathbb{N}^*$, please note also that $\mathbb{L}^* \subseteq \mathbb{L}$. Accordingly, the completeness metric referred to a subset of nodes $\mathbb{N}^*$ is defined as:

$$\eta_{\mathbb{N}^*} = \sum_{L_k \in \mathbb{L}^*} w_k \tag{12}$$

The metric is built by considering the recognition of at least one node with label L_k to be a sufficient condition to increase the metric value of the corresponding w_k. Accordingly, the completeness metric ranges from 0 to 1: it equals 0 if no nodes are present in $\mathbb{N}^*$, and it equals 1, instead, when all the labels in $\mathbb{L}$ have been recognized a least once in the subgraph of $\mathbb{N}^*$ nodes.

3. Computing the completeness metric

In this section the completeness metric assessment for near miss reports is instantiated on a dataset of almost 4,000 near miss reports, collected from more than 250 Seveso industrial establishments operating in the 26 Seveso industrial sectors. The resulting knowledge graph counts more than 45,000 nodes and 75,000 relationships. These dimensions motivate the need for a computational process to ground the proposed safety meta-analysis.

To frame the AHP decision problem, a goal, a criterion, and the alternatives must be selected. The goal is to maximize the completeness of a near miss report (as a proxy measure of the usefulness that the near miss document acquires for safety analyses); the criteria are the different types of information a compiler can insert; and the alternatives are the terms the compiler can use to represent that information. Accordingly,

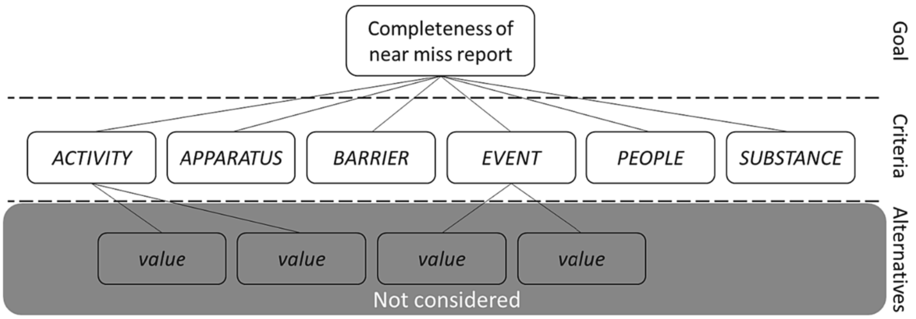

Fig. 1. Schematization of the AHP decision problem. The lower level in grey (i.e., occurrences of the *value* property) has not been considered in the analysis.

Table 1
Data about the experts involved in the AHP questionnaire.

ID	Affiliation	Description
d_1	Supervisory authority institute	Seveso inspector with more than 25 years of experience in the field of industrial health and safety
d_2	Supervisory authority institute	Seveso inspector with more than 10 years of experience in the field of environmental protection
d_3	Academia	Researcher with more than 10 years' experience in the field of safety and resilience
d_4	National research center	Researcher with more than 25 years of experience in the field of industrial health and safety
d_5	Petroleum additives manufacturing plant	Health and Safety department director with more than 20 years of experience in the field of chemical industries
d_6	LNG treatment and storage plant	Health and Safety department director with more than 10 years of experience in the field of safety and environmental protection

Table 2
Saaty's scale, as adopted by the authors.

Verbal description	Numerical value
Indifference. The two possible answers have the same contribution on report completeness.	1
Moderate preference. An answer is moderately more important than the other to complete the report.	3
Strong preference. An answer is more important than the other to complete the report.	5
Very strong or demonstrated preference. An answer is definitely more important than the other to complete the report.	7
Extreme preference. The answer is the most important to ensure report completeness.	9

Table 3
Pseudo-code to embed the completeness metric in the knowledge graph.

```
for $n' := 1$ to $N$ :                                   Cycle over source nodes
  if $L^N_{n'} = DOCUMENT$ then:                         Select node with label DOCUMENT
    $p^N_{4,n'} = 0$;                                    Initialize completeness
                                                         property value
    for $k := 1$ to $K$ :                                Cycle over labels (AHP criteria)
      counter $= k$;                                     Initialize counter variable
      for $n'' := 1$ to $N$ :                            Cycle over target nodes
        if ($L^N_{n''} = L_k$ and $\exists R_m : N'_m =  Select node connected to
  N_{n'}$,                                               DOCUMENT and
          $N''_m = N''_{n'}$, $L^R_m = CONTAINS$           verify no target with same
  and                                                    label has
           counter $= k$) then:                            been already visited
          $p^N_{4,n'} = p^N_{4,n'} + w_k$;               Update completeness property
                                                         value
          counter $=$ counter $+ 1$;                     Update counter variable
          $n'' = n'' + 1$;                               Continue cycling over target
                                                         nodes
        else:                                            Logic to deal with not selected
                                                         targets
          $p^N_{4,n'} = p^N_{4,n'}$;                     Not update completeness property
                                                         value
          $n'' = n'' + 1$;                               Continue cycling over target
                                                         nodes
      end for
      $k = k + 1$;                                       Continue cycling over labels
    end for
    $n' = n' + 1$;                                       Continue cycling over source
                                                         nodes
  else:
    $n' = n' + 1$;                                       Continue cycling over source
                                                         nodes
end for
return G with updated                                    Return graph with updated
  properties                                             properties
```

being each term modeled as a node in the knowledge graph, nodes labels are the possible criteria, while all the distinct occurrences of the *value* property represent the possible alternatives. In this work, the alternatives level has been neglected, since more than 2000 alternatives are present in the knowledge graph, making it operationally impossible to evaluate them through interviews with experts.

Each instance of *value* property have been considered to be equally important for its respective criterion. This latter assumption reflects the alternatives (i.e., the values) being dependent from the specific industrial process, and would inhibit the possibilities of sharing knowledge among Seveso industrial sectors. For instance, consider the *SUBSTANCE* label, which possesses varying *value* properties across different industries (e.g., Liquified Petroleum Gas, Ammonia). Nonetheless, each of these properties holds equal significance, as they contribute to classifying industries under the Seveso directive.

On these premises, the ontology model described in Section 2.1.1 gives the total number of criteria K that is equal to 9. However, in terms of report completeness, it is worthy to consider only the labels regarding the report content (i.e., *EVENT*, *ACTIVITY*, *APPARATUS*, *SUBSTANCE*, *PEOPLE*, *BARRIER*) imposing $K = 6$. Fig. 1 summarizes the structure of the AHP decision problem.

A group of $D = 6$ experts was asked to individually compile the pairwise comparison matrix through a questionnaire. Some descriptive user data related to the six respondents are summarized in Table 1.

Since Eq. (6) holds true, the resulting questionnaire counts 15 questions which are meant to quantify the importance of one criterion over the other. The Saaty scale have been used to perform the quantitative comparison with multiple choice verbal judgement, cf. Table 2.

The Saaty scale enabled the completion of six different A matrices. Following Eq. (10), these latter have been grouped in an overall pairwise comparison matrix B. The resulting matrix is presented below:

$$B = \begin{pmatrix} 1.00 & 1.38 & 0.86 & 0.36 & 1.63 & 0.84 \\ 0.72 & 1.00 & 1.18 & 0.44 & 2.18 & 1.25 \\ 1.16 & 0.84 & 1.00 & 0.51 & 1.07 & 0.68 \\ 2.81 & 2.29 & 1.97 & 1.00 & 5.52 & 1.90 \\ 0.61 & 0.46 & 0.93 & 0.18 & 1.00 & 0.36 \\ 1.18 & 0.80 & 1.48 & 0.53 & 2.81 & 1.00 \end{pmatrix} \tag{13}$$

Table 4
Seveso industrial macro-sectors definition. The right column includes all the 26 industrial sectors as defined by the Seveso legislation.

ID	Description	Sectors included
MS1	The main hazards associated with these industrial activities are not directly connected with the Seveso legislation. Their main business is not Seveso-related	Mining activities Metal processing Ferrous metalworking Non-ferrous metalworking Chemical/electrolytic metal treatment Pottery manufacturing Others (not directly specified)
MS2	All the activities related to these industrial sectors involve oil and petroleum	Petrochemical and oil refineries Energy production, supply, and distribution (fossil fuel power stations) Storage of fuels (no LPG and LNG) Fuels wholesale and retail storage and distribution (no LPG)
MS3	All the activities related to these industrial sectors involve explosives	Production, destruction, and storage of explosives Production and storage of pyrotechnical goods
MS4	All these industrial sectors work with gases	Production, bottling, and distribution of LPG Storage of LPG Storage and distribution of LNG
MS5	All these industrial sectors work with chemicals or related products	Production and storage of pesticides and biocides Production and storage of fertilizers Pharma manufacturing Chemical plants Production of basic organic chemicals Plastics and rubber manufacturing Chemicals manufacturing (not directly specified)
MS6	All the activities related to these industrial sectors are connected to waste management	Storage, treatment, and disposal of waste Collection, supply, and treatment of water and wastewater

where the rows (and the column) are respectively referred to: L_1 = *ACTIVITY*, L_2 = *APPARATUS*, L_3 = *BARRIER*, L_4 = *EVENT*, L_5 = *PEOPLE*, L_6 = *SUBSTANCE*. The matrix B in Eq. (13) is checked for consistency through Eq. (8), which becomes:

$$CR = \frac{1}{1.25}\cdot\frac{6.13-6}{6-1} \cong 0.02 \tag{14}$$

depicting a consistent result since $CR < 0.1$. The vector of weights w is finally computed through Eq. (9), obtaining:

$$w = \begin{pmatrix} 0.41 \\ 0.45 \\ 0.38 \\ 1.00 \\ 0.23 \\ 0.50 \end{pmatrix} \rightarrow \begin{pmatrix} 0.14 \\ 0.15 \\ 0.13 \\ 0.34 \\ 0.08 \\ 0.17 \end{pmatrix} \tag{15}$$

where, again, rows are respectively referred to: L_1 = *ACTIVITY*, L_2 = *APPARATUS*, L_3 = *BARRIER*, L_4 = *EVENT*, L_5 = *PEOPLE*, L_6 = *SUBSTANCE*. Please note that a normalization has been performed on w elements to ensure the completeness metric η to range between 0 and 1.

4. Analysis and discussion

Once the weights for each report element have been computed, the completeness metric for each report must be integrated in the knowledge graph. Accordingly, the pseudo-code in Table 3 formalizes the calculation of the completeness metric, which is added as an additional property $p^N_{4,n}$ (namely *completeness*) to nodes with label *DOCUMENT*. Where the `counter` variable has been used to ensure labels to be counted only once in the metric calculation, in accordance with Eq. (12).

The knowledge graph model enables the analysis of changes in the completeness metric in relation to different dimensions. Report completeness can be investigated by means of: (i) occurrences of the *value* property nodes can assume, (ii) industrial sector of provenience (i.e., connections with nodes with label *INDUSTRIAL_SECTOR*), (iii) establishment of provenience (i.e., connections with nodes with label *ESTABLISHMENT*), (iv) date of collection (i.e., *collection_date* property in nodes with label *DOCUMENT*), (v) date of redaction (i.e., *occurence_date* property in nodes with label *DOCUMENT*), (vi) location (i.e., *location_region* and *location-district* properties in nodes with label *ESTABLISHMENT*), or any combination of the above.

The assumptions for the calculation are:

- Some affinities between Seveso-related industrial sectors exist, by means of the reasons they must comply with the Seveso legislation. Accordingly, the outcome of the analysis is obtained considering 6 macro-sectors. Details about each macro-sector definition are reported in Table 4. This stratification is presented to reflect the assumption that the macro-sectors are assumed to be homogeneous in terms of reporting peculiarities.
- Data related to the establishments have been anonymized to guarantee companies' privacy.
- The date of collection and the date of redaction of the near miss reports only consider the corresponding year. This choice was forced since ca. 70% of available reports do not contain data about the month and the day of collection/redaction.

Aggregated measures of the completeness metric η have been used, too. Specifically, the average value of η with respect to a set of reports is calculated under different query conditions. For example, the completeness capacity of an industrial establishment can be seen as the average value of completeness of all the reports it submitted. In general, the value of average completeness can be defined as:

$$\bar{\eta} = \frac{\sum_{N_n \in \mathrm{H}} p^N_{4,n}}{|\mathrm{H}|} \tag{16}$$

where H is the set of all the nodes with label *DOCUMENT* obtained because of queries applied to the graph database, |H| is the cardinality of H (i.e., number of resulting nodes), and $p^N_{4,n}$ is the value of η assigned to a node with label *DOCUMENT*. The definition of queries follows the ontological explorative analysis presented by Simone et al., (2023).

Based upon the six dimensions defined above, in the following paragraphs we show four different dimensions to guide the meta-analysis of near miss reports completeness:

(i) the first one will investigate the completeness of reports by considering the different occurrences of the *value* property (i.e., reports' terms) each label can assume
(ii) the second one will account the average completeness of reports grouped by establishment and industrial macro-sector
(iii) the third one will consider the date of collection and the industrial macro-sector
(iv) the fourth one will investigate the date of redaction and the industrial macro-sector.

Each dimension will be grounded on a distinct definition of the H set

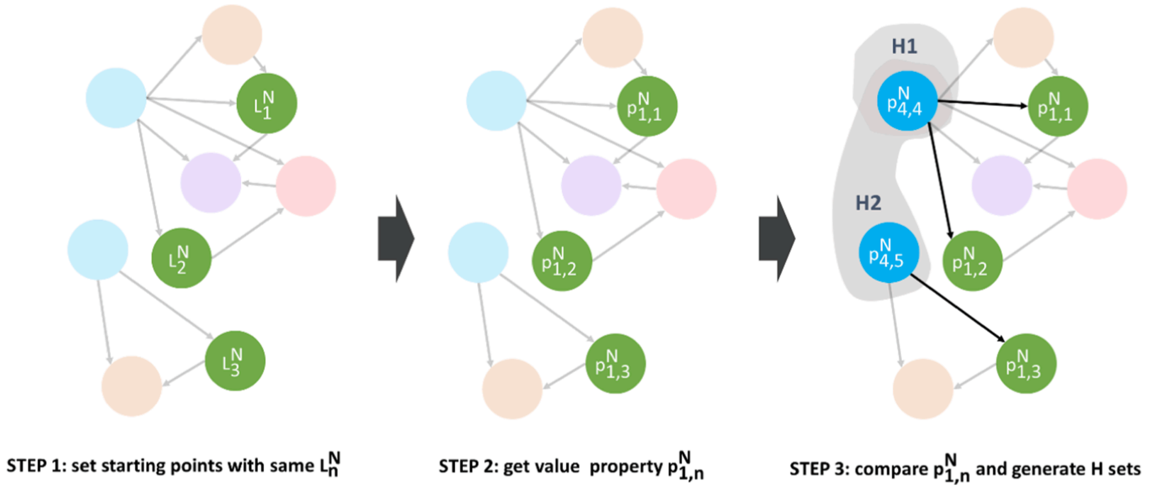
L^N_1
L^N_2
L^N_3
$p^N_{1,1}$
$p^N_{1,2}$
$p^N_{1,3}$
H1
H2
$p^N_{4,4}$
$p^N_{4,5}$
STEP 1: set starting points with same L^N_n
STEP 2: get value property $p^N_{1,n}$
STEP 3: compare $p^N_{1,n}$ and generate H sets

Fig. 2. Schematization of the ontological explorative analysis to investigate completeness based upon the *value* property. Light blue nodes identify *DOCUMENT* labels, green nodes identify *BARRIER* labels, the other colors are referred to the other labels (*ACTIVITY*, *APPARATUS*, *EVENT*, *PEOPLE*, *SUBSTANCE*), grey areas identify the H sets. Solid black arrows represent existing graph relationships. Nodes' numbering in the figure is purely exemplary. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

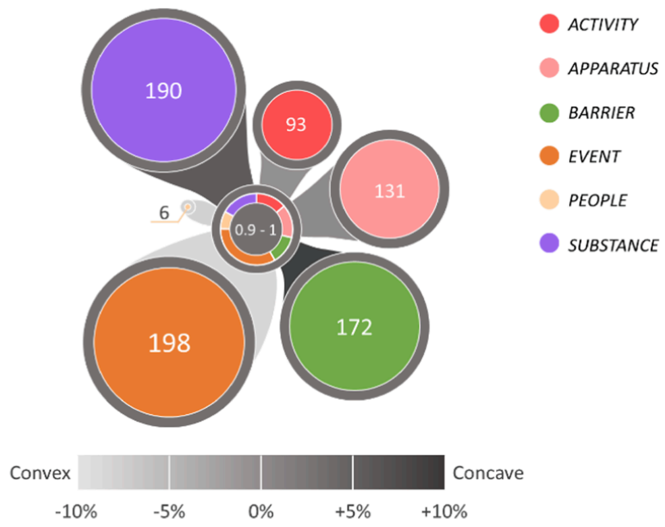

Fig. 3. Flower plot showing numerosity of different terms (i.e., *value* instances) for each label, referred to an average completeness between 0.9 and 1.

Table 5
Nodes' labels numerosity by means of different occurrences of the value property with respect to average completeness ranging between 0.9 and 1. The central column includes the labels' weights as obtained from the AHP questionnaires. The right column reports the difference between the two.

	Weight by % numerosity of terms	Weight by AHP results	Variation on weight
ACTIVITY	11.77%	14.00%	−2.23
APPARATUS	16.58%	15.00%	+1.58
BARRIER	21.77%	13.00%	+8.77
EVENT	25.06%	34.00%	−8.94
PEOPLE	0.76%	8.00%	−7.24
SUBSTANCE	24.05%	17.00%	+7.05

obtained following an ontological explorative path analysis on the graph (Simone et al., 2023). The spatial dimension (i.e., establishment location) has not been detailed in this paper to avoid privacy concerns. However, there could be the possibility to define a proper H set grouping nodes with label *ESTABLISHMENT* by means of their *location_region* or *location_district* properties.

4.1. Semantical overview

The first ontological path is meant to explore how the completeness of reports is influenced by the *value* property. This analysis is proposed to partially overcome the operational limitation which has prevented the possibility of evaluating the *value* property as an alternative in the AHP questionnaire. Accordingly, the different instances of the *value* property are extracted from the graph and the average completeness of reports containing them is evaluated.

An example for the ontological explorative analysis to be made is graphically summarized in Fig. 2. A specific label L^N_n is set and nodes with that label are set as starting points. Green nodes (i.e., nodes with label *BARRIER*) are used as an example. A second step extracts the property $p^N_{1,n}$ (i.e., *value*) from the selected nodes. This property represents the unique key for the analysis. The last step involves the navigation of the relationships with label *CONTAINS* to move to the corresponding nodes with label *DOCUMENT* (i.e., light blue nodes in Fig. 2). This step permits to relate each instance of the *value* property to the corresponding *completeness* property of *DOCUMENT* nodes by extracting $p^N_{4,n}$. The H sets, and their cardinalities |H|, are defined by comparing values of $p^N_{1,n}$: *DOCUMENT* nodes that have been navigated from nodes with same $p^N_{1,n}$ belong to the same H set. Through Eq. (16), a value of $\bar{\eta}$ can be assigned to each instance of the *value* property averaging over the matching *completeness* values. The resulting extracted data comprehends: all the *value* occurrences, their corresponding labels, and each *value* resulting average completeness.

Results obtained from the knowledge graph query are summarized in the flower plot in Fig. 3, showing only instances of the *value* property in the range $0.9 < \bar{\eta} \leq 1$. The proposed flower plot can be interpreted as it follows:

- the petals represent the six different labels.
- the size of each petal is proportional to the numerosity of the instances of the *value* property satisfying the condition on $\bar{\eta}$.
- the center of the flower plot reports a pie-chart containing the weight of each label as obtained from the AHP questionnaires. Convex petals (lighter ones) show a reduction in the importance from expert judgement by means of actual numerosity spotted in data, concave petals (darker ones) show an increase, instead.

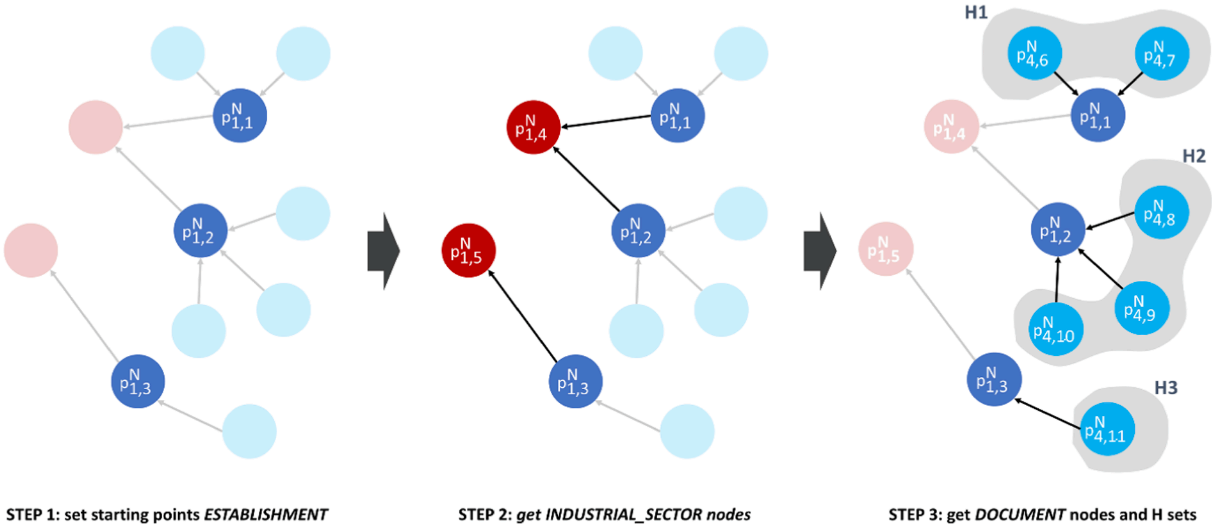
H1
H2
H3
p N 1,1
p N 1,2
p N 1,3
p N 1,4
p N 1,5
p N 4,6
p N 4,7
p N 4,8
p N 4,9
p N 4,10
p N 4,11
STEP 1: set starting points ESTABLISHMENT
STEP 2: get INDUSTRIAL_SECTOR nodes
STEP 3: get DOCUMENT nodes and H sets

Fig. 4. Schematization of the ontological explorative analysis to investigate completeness of industrial establishments. Blue nodes identify *ESTABLISHMENT* labels, red nodes identify *INDUSTRIAL_SECTOR* labels, light blue nodes identify *DOCUMENT* labels, grey areas identify the H sets. Solid black arrows represent existing graph relationships. Nodes' numbering in the figure is purely exemplary. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

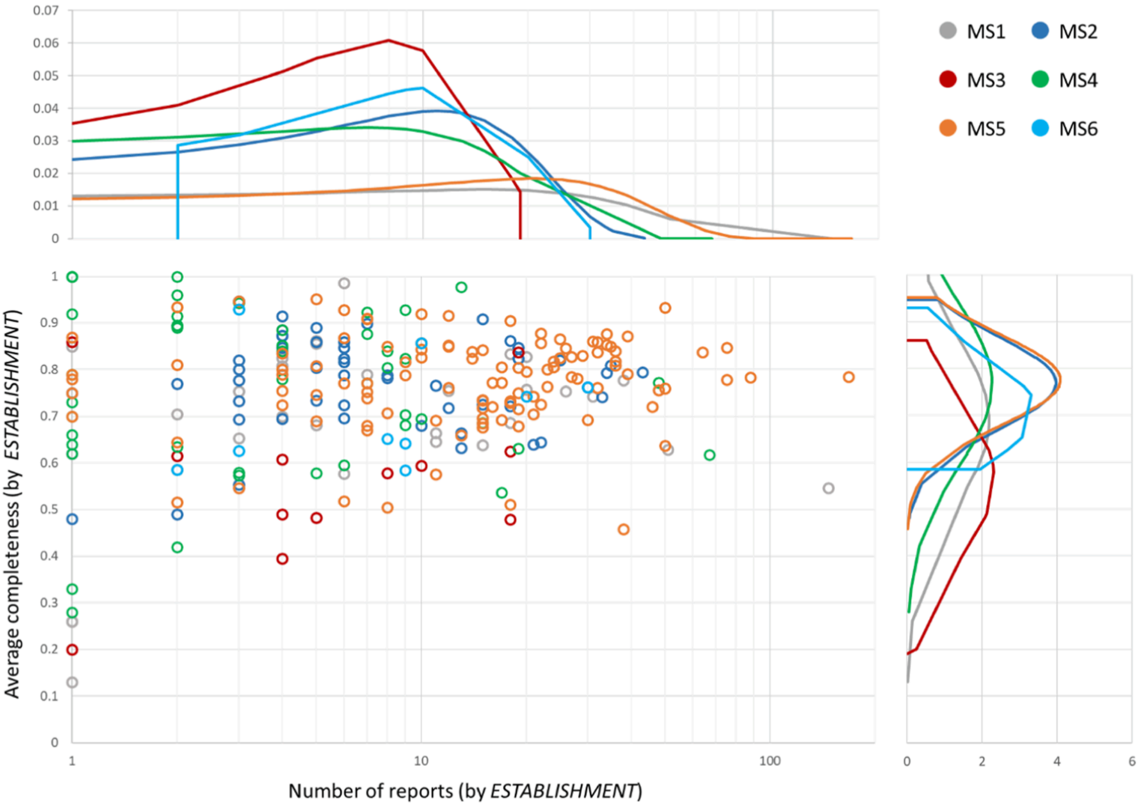

Fig. 5. Industrial establishment analysis per macro-sector by means of average completeness and number of reports submitted (on a logarithmic scale).

Numerical values of the resulting investigation are reported in Table 5. It is clear how the weight of some labels as obtained from their frequency is not matching the AHP results. The *EVENT* label remains the most important but it registers a significant reduction (−9%), being almost equaled by the *SUBSTANCE* label (less than 1% difference). An impressive result is the one related to the *BARRIER* label, which shows an increase of almost 9% in weight. Accordingly, they may play a central role in telling what went well (or wrong) avoiding the near miss to evolve into a major adverse event.

It is worth noticing that the tagging algorithm may have paid a role in these results, since its performance in classifying terms may have been not completely uniform (Ansaldi et al., 2021). However, this result shows how practitioners actually stress the need to give information about the events which occurred, the substances involved, and the barriers which worked/failed. At a deeper level of analysis, the terms contained in the value property can be investigated, too, pointing out the topics in which the system showed a more punctual storytelling. Analogous plots and analyses can be made for different ranges of

Table 6
Number of establishments per macro-sector.

	MS1	MS2	MS3	MS4	MS5	MS6
Number of establishments	32	49	12	47	111	9
Number of reports	483	538	94	328	2286	94
Average number of reports per establishment	15.09	10.98	7.83	6.98	20.59	10.44

completeness to spot (e.g.,) which are the actual terms (or even topics) contained in uncomplete reports. These latter can be seen as a symptom of the challenges emerging from the management of such narratives, or can point at possible weaknesses in reporting systems.

4.2. Establishments overview

The second path is set to assess the ability of each industrial establishment to complete a near miss report. An example for the ontological explorative analysis to be made is graphically summarized in Fig. 4. The nodes with label *ESTABLISHMENT* (blue nodes in Fig. 4) are set as starting points and they are identified by the property *establishment_id* (i.e., $p^N_{1,n}$ in nodes with label *ESTABLISHMENT*). Such property is never repeating over different industrial establishments and represents the unique key for the analysis. From the *ESTABLISHMENT* nodes, the relationship towards the *INDUSTRIAL_SECTOR* nodes are explored. For these latter, the *industrial_sector_id* property (i.e., $p^N_{1,n}$ in nodes with label *INDUSTRIAL_SECTOR*) is initially stored. This step permits to relate each industrial establishment to the industrial sector it belongs to, enabling a parallel analysis of industrial sectors, too. The third and last step navigates the relationships with label *FROM* to match all the nodes with label *DOCUMENT* related to each industrial establishment.

This results in the definition of the H sets, and their corresponding cardinalities |H|. Thus, through Eq. (16), a value of $\bar{\eta}$ can be assigned to each industrial establishment averaging over the matching *completeness* values (i.e., $p^N_{4,n}$ in nodes with label *DOCUMENT*). The resulting extracted data comprehend: all the *establishment_id*, the establishments corresponding *industrial_sector_id*, and each establishment corresponding H set.

Data obtained from the knowledge graph is used to characterize each industrial establishment with its ability to compile reports in a complete way. The result is shown in the scatter plot in Fig. 5. An industrial establishment is identified by a point in the plot by means of its average completeness $\bar{\eta}$, and the number of reports this average was computed on (i.e., |H|), which is plotted on a logarithmic scale. A total of 260 industrial establishments have been analyzed, the number of establishments belonging to each macro sector is summarized in Table 6, along with the total number of reports submitted by the macro-sector, and subsequent average number of reports submitted by a single establishment. The color code of the points relates them to the corresponding macro-sector (cf. Table 4). The additional plots that are positioned above and left to the scatter plot, show the normalized frequency of values the points assume, grouped by macro-sectors. It is clear how some macro-sectors are made up of establishments that have a better ability in ensuring a complete storytelling of near misses. Specifically, MS2 and MS5 are shown to be the best two macro-sectors by means of the average completeness their establishments can guarantee.MS5 also has the higher scores by means of reports submitted. This result may depict a strong sensibility of MS5 in managing near misses, but it is also clearly related to the higher number of establishments belonging to this macro-sector. A particular case is the one of MS6, which can guarantee a quite constant completeness value, that is majorly condensed between 0.6 and 0.8, with a scarce presence of establishment. This result suggests a proved tendency in the sector in avoiding some specific type of details in the near miss reports. On the other hand, MS1 and MS4 show a wide distribution of the average completeness, stressing the presence of internal differences between actors in the same sector. This result can be seen as an inner characteristic of MS1 since it has been built by grouping different industries that falls under the Seveso legislation for some side processes they perform. Nevertheless, it is also an indicator that some establishments of both MS1 and MS4 may benefit from lessons learned from other actors in their field to improve their ability in managing near misses. The worst performance is given by MS3 both by means of number of reports and average completeness. If the first result can be partially explained by the scarce presence of industries in this macro-sector, the second one is depicting a clear overall low completeness of MS3 reports, that may be improved through the few establishments in MS3 with positive scores. These latter may then be used as leading examples to improve the macro-sector performance in managing near misses.

4.3. Temporal overview

The next analysis is meant to investigate how the ability of industrial sectors in reporting near misses changed over time. This temporal analysis can be twofold: (i) using the date in which the near miss report was collected by the competent authority (i.e., $p^N_{2,n}$ in nodes with label *DOCUMENT, namely, collection_date* property), or (ii) using the date in which the near miss event happened (i.e., $p^N_{3,n}$ in nodes with label

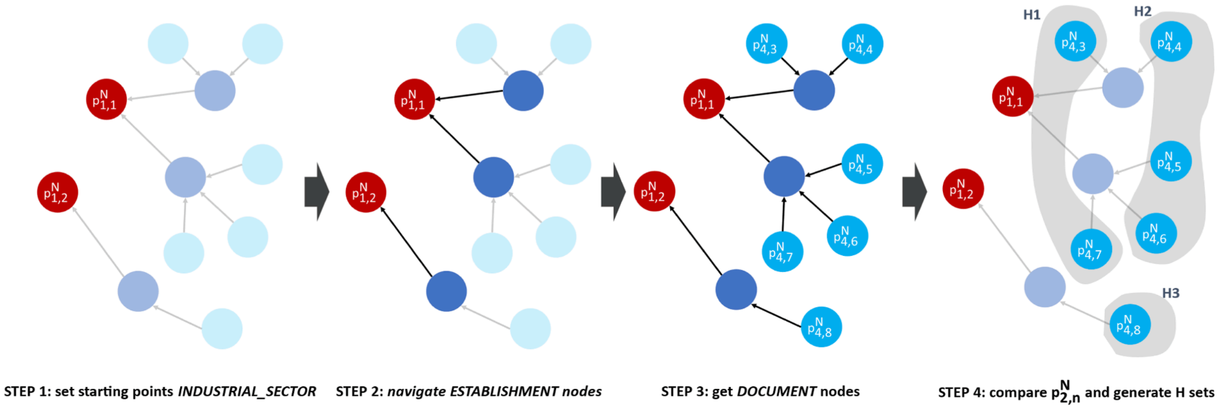
$p^N_{1,1}$
$p^N_{1,2}$
$p^N_{4,3}$
$p^N_{4,4}$
$p^N_{4,5}$
$p^N_{4,6}$
$p^N_{4,7}$
$p^N_{4,8}$
H1
H2
H3
STEP 1: set starting points *INDUSTRIAL_SECTOR*
STEP 2: *navigate ESTABLISHMENT nodes*
STEP 3: get *DOCUMENT* nodes
STEP 4: compare $p^N_{2,n}$ and generate H sets

Fig. 6. Schematization of the ontological explorative analysis to investigate completeness by collection and occurrence year. Red nodes identify *INDUSTRIAL_SECTOR* labels, blue nodes identify *ESTABLISHMENT* labels, light blue nodes identify *DOCUMENT* labels, grey areas identify the H sets. Solid black arrows represent existing graph relationships. Nodes' numbering in the figure is purely exemplary. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

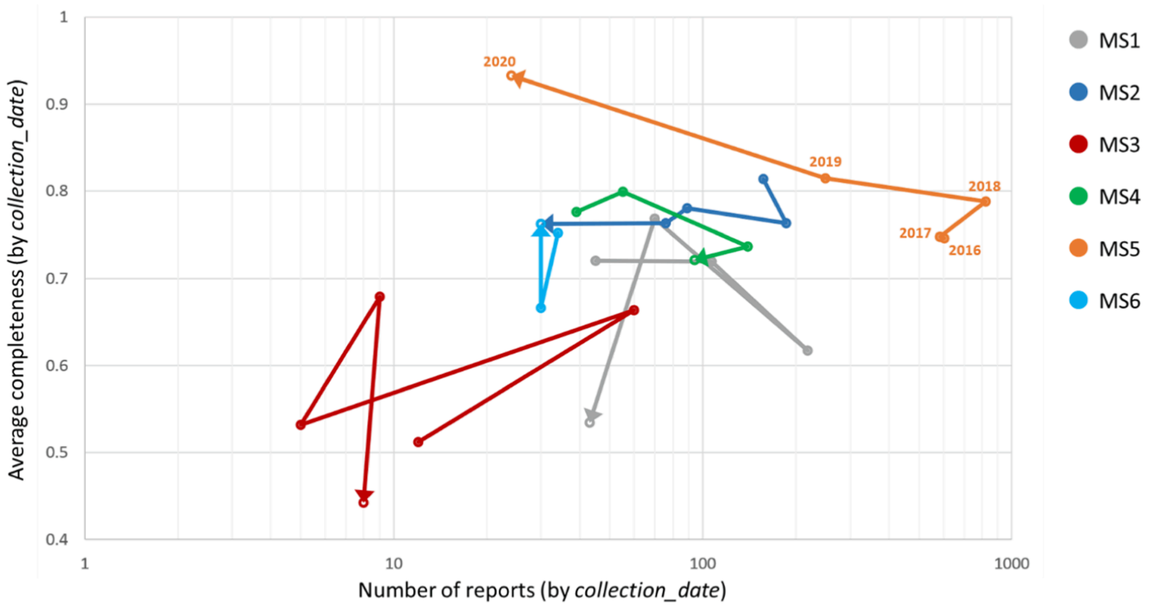

Fig. 7. Temporal analysis by *collection_date* by means of average completeness and number of reports submitted (on a logarithmic scale).

Table 7
Average number of reports submitted in a year by an industrial establishment per macro-sector. This table specializes Table 5 providing an average yearly estimate.

	MS1	MS2	MS3	MS4	MS5	MS6
Reports submitted by an establishment in a year	1.34	0.61	0.67	2.00	0.22	3.33

DOCUMENT, namely, occurrence_date property). The corresponding ontological explorative analysis is graphically summarized in Fig. 6. The nodes with label *INDUSTRIAL_SECTOR* (red nodes in Fig. 6) is set as starting points and they are identified by the property *industrial_sector_id* (i.e., $p^N_{1,n}$ in nodes with label *INDUSTRIAL_SECTOR*). The analysis then moves from the *INDUSTRIAL_SECTOR* nodes towards the *ESTABLISHMENT* nodes (blue nodes in Fig. 6). This is an intermediate step, and no property is stored at this stage. The third step navigate from the nodes with label *ESTABLISHMENT* to the nodes with label *DOCUMENT* (light blue nodes in Fig. 6), matching the documents related to each industrial sector and extracting their *completeness* property (i.e., $p^N_{4,n}$ in nodes with label *DOCUMENT*). At this point a comparison between *DOCUMENT* nodes referring to the same *INDUSTRIAL_SECTOR* is made by means of their *collection_date* or *occurrence_date* property. Nodes with matching values are grouped in the same H set. Please note that each *INDUSTRIAL_SECTOR* can be related to more than one H set. Thus, the unique key for the analysis is the combination of *INDUSTRIAL_SECTOR* and a specific value the *collection*_date/occurrence_*date* property can assume.

The definition of the H sets, and their corresponding cardinalities $|\text{H}|$ permits the calculation of $\bar{\eta}$ through Eq. (16). The resulting extracted data comprehend: all the unique combinations of *industrial_sector_id* and *collection*_date/occurrence_*date*, and their corresponding H sets. Extracted data enable two temporal analyses of reports completeness and numerosity of the industrial macro-sectors by means of *collection_date* and *occurrence_date*, respectively. Fig. 7 reports the temporal behaviour of the macro sectors with respect to *collection_date*. The arrows in the plot depict the variation of reports completeness and numerosity year by year, starting from 2016 to 2020. Results show an overall tendency of all macro sectors in diminishing the number of reports to be submitted to the competent authority. To explain this fact, it is important to point out that the implementation of the Seveso III directive has improved the inspections' scheduling guaranteeing that all upper-tier establishments

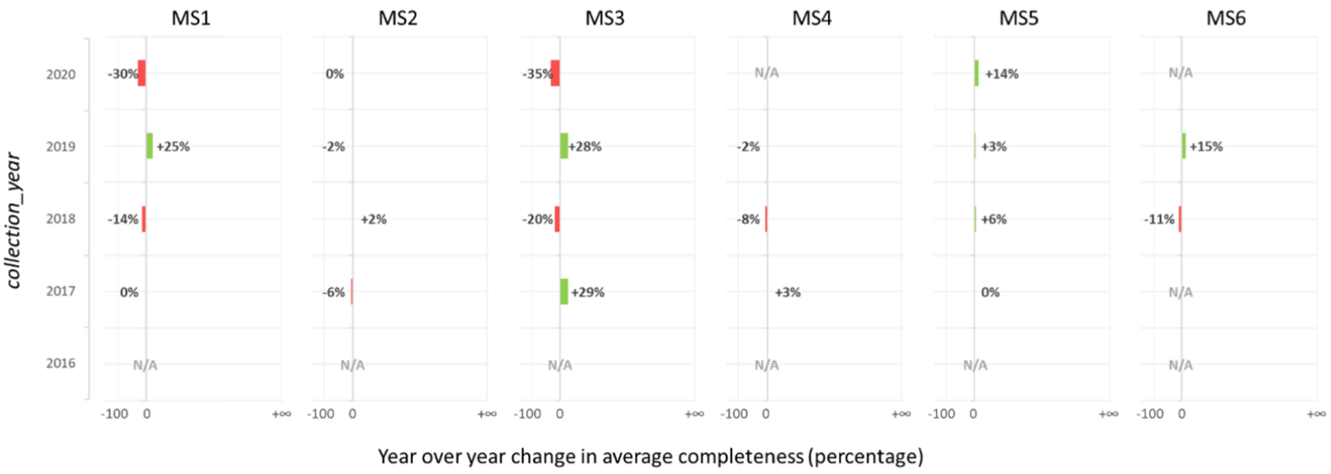

Fig. 8. Temporal analysis by *collection_date*. Year over year percentage change of the average completeness by macro-sectors.

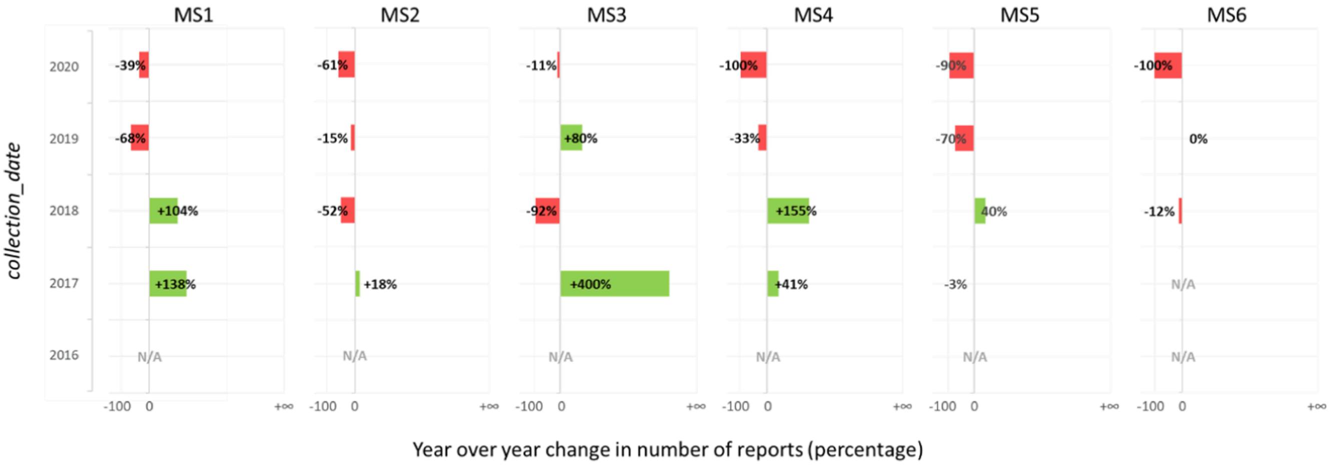

Fig. 9. Temporal analysis by *collection_date*. Year over year percentage change of the number of reports collected by macro-sectors.

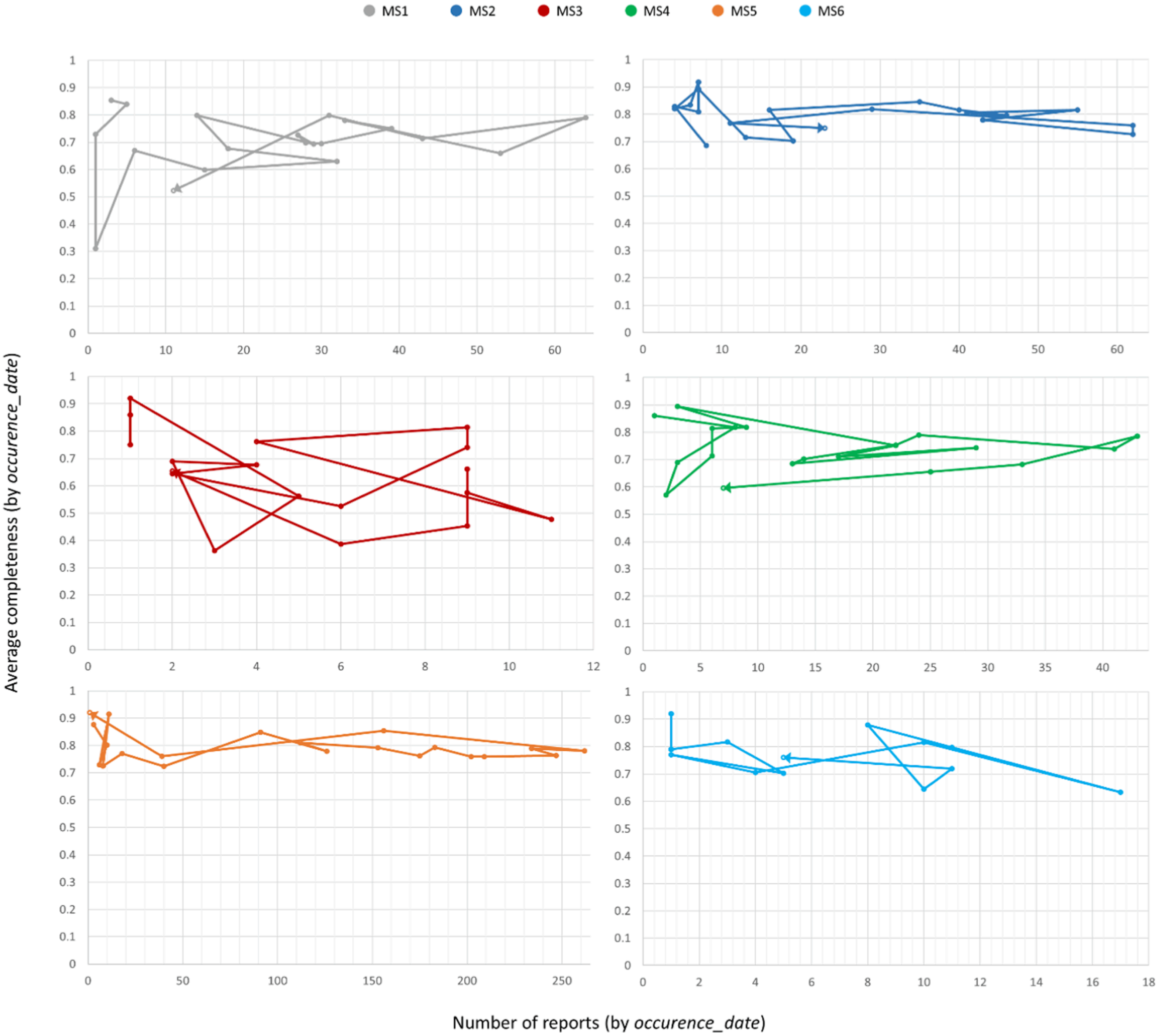

Fig. 10. Temporal analysis by *occurence_date* by means of average completeness and number of reports submitted. While the vertical axes are left unchanged, one should pay attention to the scale on the horizontal axis for comparison purposes.

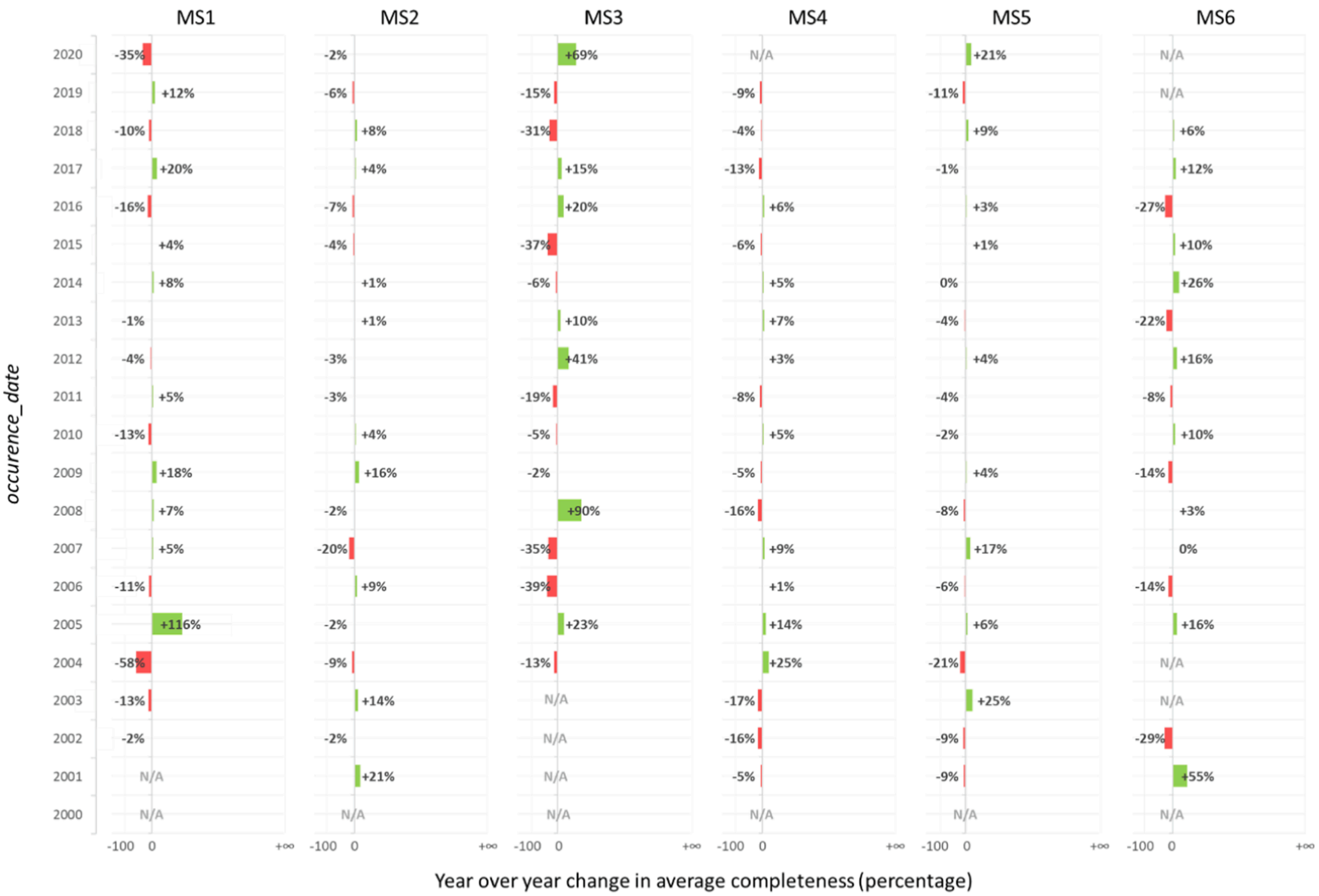

Fig. 11. Temporal analysis by *occurence_date*. Year over year percentage change of the average completeness by macro-sectors.

would be involved by three years. Before 2015, some establishments had not been inspected for a while; thus, they have accumulated the near miss reports over time. Accordingly, the diminishing number of reports may be an indicator of a stabilization in the number of near miss events submitted, cf. the number for each macro-sector to be lower than 100 reports per year. This result should be evaluated with respect to the number of analysed establishments belonging to each macro sector. Accordingly, Table 7 summarizes the average number of near misses that an industry in each macro-sector submits in a year. MS1, MS4, and MS6 are the only sectors made up of establishments submitting more than one report per year. Accordingly, to the analysis of industrial establishment in Section 4.2, the first one (i.e., MS1) has a distribution of frequency of number of reports that tend to higher numbers (cf. Fig. 5), depicting a particular attention of industries on the theme of near misses. Nevertheless, the yearly need for submitting report is slightly higher than 1, that, accordingly, depicts few near misses happening and overall, a safe performance of the sector. On the contrary, MS4 and MS6 have the distributions of frequencies of number of reports that tend to lower numbers (cf. Fig. 5), stressing the possibility to improve safety in operations of such sectors.

Concerning the value of average completeness, the best performance is the one of MS5, which shows a continuous improvement in completeness. MS2, MS4, and MS6 keep an almost constant performance overall, while MS1 and MS3 present a decrease. Such result is coherent with the analysis in Section 4.2 which already highlighted the criticalities linked to MS3.

For clarity purposes, an additional visualization of the temporal analysis is proposed. The year over year percentage change ($YoY_{\%}$) in number of reports and completeness is calculated for each macro sector as:

$$YoY_{\%} = \frac{v_{(t)} - v_{(t-1)}}{v_{(t-1)}} \cdot 100 \tag{17}$$

where $v_{(t)}$ is the value of the analyzed variable in a specific year of collection t. Please note that $YoY_{\%}$ ranges between -100% to $+\infty$. $YoY_{\%}$ tends to 0 if no reports are collected in $v_{(t)}$, or the average completeness in $v_{(t)}$ is strongly minor than the average completeness in $v_{(t-1)}$; $YoY_{\%}$ tends to its upper limit instead, if the reports collected in $v_{(t)}$ are much more than the reports collected in $v_{(t-1)}$, or the average completeness in $v_{(t)}$ is strongly major than the average completeness in $v_{(t-1)}$.

Fig. 8 and Fig. 9 reports the $YoY_{\%}$ of completeness and number of reports respectively. It is clear how the number of reports (cf. Fig. 9) is affected by a higher year over year change, and it has a diminishing trend for most recent years, as discussed previously. Overall, by considering the average $YoY_{\%}$ over the whole time period considered, MS1, MS3, and MS4 are the only macro-sectors reporting positive deviations. Nevertheless, MS3 results to be the more critical by means of the variance of its $YoY_{\%}$. Concerning the average completeness change, all macro-sectors have negligible variances. MS1, MS2, and MS4 have, on average, a negative deviation over the five years considered, with MS1 being the most critical, with an average loss of more than 5% on completeness. Numerical values referred to the analyses in Fig. 8 and Fig. 9 are reported in Appendix A.

An analogous analysis is conducted on the property *occurrence_date*. Please note that the ontological exploration from Fig. 6 is applied equivalently with the only difference that H sets have been built by grouping over the $p^N_{3,n}$ property of nodes with label *DOCUMENT*. Fig. 10 reports the temporal behaviour of the macro sectors with respect to *occurence_date*. The arrows in the plot depict the variation of reports completeness and numerosity year by year, starting from 2000 to 2020. Results confirms MS1 and MS3 to be the more variable in terms of average completeness, with MS4 following.

Additionally, the year over year percentage change (cf. Eq. (17) in number of reports and completeness is calculated for each macro sector, and it is presented in Fig. 11 and Fig. 12. Precise numerical values are reported in Appendix B.

It is possible to notice how all macro-sectors show an increasing and then decreasing trend in the number of reports if analyzed by the *occurrence_date* field. This distinctive phenomenon may be caused by different aspects that emerge from an analysis or meta-analysis perspective. Firstly, this result might be partially unreliable due to the incompleteness of the dataset.

One should note that the near miss reports inserted in the graph are the ones collected during the Seveso inspection procedures. Accordingly, Table 8 contains the rate of completion of the number of expected Seveso inspections that are contained in the database. The years 2019 and 2020 (this latter being influenced by the COVID-19 outbreak, too) suffer from increasingly missing data, which have an impact on the subsequent analyses.

Another possible explanation can be related to the absence of a prescription in submitting reports until 2015, resulting in a temporary amplification of the industry concern in reporting near misses. This aspect may be a symptom of potential under-reporting by industries that may perceive limited benefit in reporting. Moreover, further explanation may refer to the fact that Seveso industries are becoming increasingly virtuous in managing industrial safety, reducing the occurrence of adverse events over recent years. In line with the purpose of this work, the analysis is not delving into the operational causes for this behavior, rather suggesting a methodological support for highlighting underlying behaviors as emerged from safety reporting.

5. Conclusion

This paper examines the importance of near misses and highlights how they can be utilized as a key indicator to improve industrial safety in Seveso establishments. On this basis, the information load of near miss reports must be maximized to help organizations identify and mitigate risks before they escalate towards more serious epilogues. The paper presents a meta-analytic safety perspective that updates the knowledge graph of near miss reports (Simone et al., 2023) adding weights to graph nodes. These latter permit to calculate a completeness metric that assesses the informative content of each report. This meta-analytic metric suggests hotspots to be further investigated, but does not directly connect its findings to traditional safety and risk analyses. Future works may investigate this relation with further attention.

Answering the research question of this paper, the knowledge graph enables different meta-analysis declinations permitting to: (i) analyze the terms that have been used in the reports and connect them with reports' completeness, (ii) characterize industrial establishments and their corresponding industrial sector through the completeness of reports they submitted, and (iii) have an understanding of the reports' completeness over time.

Nevertheless, the presented results suffer from some limitations, that are majorly related to: (i) the partial incompleteness of the dataset that have been used, (ii) the ontology that has been tailored on near miss data, remaining to some extent limited to this specific domain, and (iii) the computational costs that are in our case negligible, but they may represent a limitation in larger-scale investigations. Accordingly, some open questions are left to be answered. For example, future works may address the problem of the later decreasing in number of near miss occurrences by interviewing industries and defining whether – if any – the measures they implemented were truly effective. The prioritization of areas to be investigated, the identification of potentially critical or best-in-class sectors or specific enterprises could complement the usage of this completeness metric.

Nonetheless, this approach can be further extended. In particular, the currently defined metric only considers the presence or not of a particular type of data, without considering its actual value, and its topological properties. A more accurate completeness metric that overcomes this limitation can be computed by embedding mechanisms, e.g., the nodes' page rank (Zhang et al., 2022). A more precise analysis should also detail weights for the *value* property, which was shown to be a parameter of interest by means of reports' completeness. Specifically, Section 4.1 highlights the importance of the terms classified through the *BARRIER* label, and it is consistent with several studies detailing the importance of different safety barriers (Casson Moreno et al., 2022; Misuri et al., 2021). Further developments may integrate the meta-analysis with more sophisticated weight to be assigned to *value* properties for specific labels of interest.

The presented meta-analysis is, at this stage, purely descriptive (i.e., how things have been done?). However, these outcomes open the path to the development of a prescriptive tool (i.e., how things should be done?). This latter may support both safety managers from industries in writing highly informative reports, and regulatory authority inspectors in spotting criticalities and suggesting areas of improvement. Starting from this basis, graph embedding may be used to enable the usage of machine learning algorithms to perform edges completion or classification of missing nodes (Wang et al., 2017).

Viewing near misses as opportunities for improvement can make organizations build a culture of continuous improvement, trust, and accountability that fosters a safer and more resilient working environment. As such, the meta-analysis proposed in this research is expected to support industries in encouraging safety reporting and providing a robust basis for critical reflections.

CRediT authorship contribution statement

Francesco Simone: Conceptualization, Writing – original draft, Writing – review & editing, Visualization, Investigation, Validation, Formal analysis, Methodology, Software. **Silvia Maria Ansaldi:** Data curation, Writing – review & editing, Investigation, Validation, Formal

Table 8
Rate of completion of Seveso inspection per year and per macro-sector as contained in the analysed database.

	2016	2017	2018	2019	**2020**
MS1	100.00%	100.00%	100.00%	100.00%	8.33%
MS2	100.00%	100.00%	100.00%	82.50%	13.51%
MS3	100.00%	100.00%	100.00%	66.67%	15.38%
MS4	100.00%	100.00%	100.00%	85.71%	11.11%
MS5	100.00%	100.00%	100.00%	92.50%	21.57%
MS6	100.00%	100.00%	100.00%	100.00%	0.00%

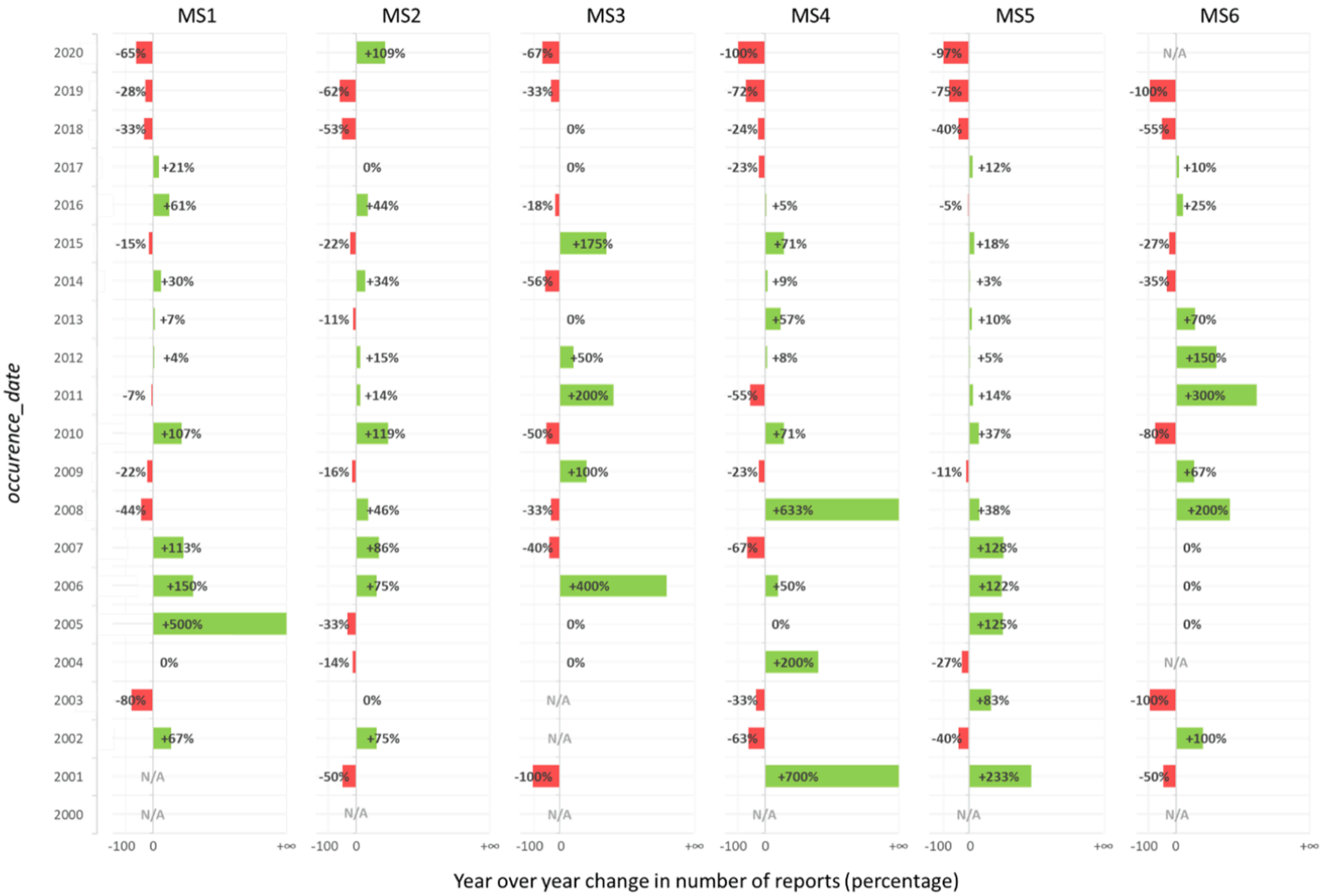

Fig. 12. Temporal analysis by *occurence_date*. Year over year percentage change of the number of reports by macro-sectors.

analysis, Supervision, Resources, Software. **Patrizia Agnello:** Data curation, Formal analysis, Validation, Writing – review & editing. **Giulio Di Gravio:** Writing – review & editing, Validation, Project administration. **Riccardo Patriarca:** Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing, Validation, Formal analysis, Methodology, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgements

The authors thank Roberto Chiapella, Alessandro Delfino, Laura Patronetti, Monica Giannetti, and Paolo Bragatto for giving their precious support in achieving the results presented in this paper. This work is part of the XXXVII Cycle Italian PhD project "Cognitive and Physical Modelling of Complex Industrial Processes for Vulnerability Analysis and Resilient Design" (PhD student: Francesco Simone) funded by INAIL in collaboration with the Department of Mechanical and Aerospace Engineering at "La Sapienza" University of Rome.

Appendix A

(See Tables A1-A2).

Table A1
Year over year percentage change in average completeness per macro-sector (by *collection_date*).

	MS1	MS2	MS3	MS4	MS5	MS6
2016 – N/A	N/A	N/A	N/A	N/A	N/A	N/A
2017 – 2016	−0.06%	−6.17%	+29.50%	+3.00%	+0.19%	N/A
2018 – 2017	−14.32%	+2.22%	−19.84%	−7.80%	+5.50%	−11.45%
2019 – 2018	+24.56%	−2.14%	+27.70%	−2.16%	+3.34%	+14.52%
2020 – 2019	−30.39%	−0.13%	−34.87%	N/A	+14.47%	N/A
Average	−5.05%	−1.55%	+0.62%	−2.32%	+5.87%	+1.53%
Variance	0.04	0.00	0.08	0.00	0.00	0.02

Table A2
Year over year percentage change in number of reports per macro-sector (by *collection_date*).

	MS1	MS2	MS3	MS4	MS5	MS6
2016 – N/A	N/A	N/A	N/A	N/A	N/A	N/A
2017 – 2016	+137.78%	+18.47%	+400.00%	+41.03%	−2.98%	N/A
2018 – 2017	+103.74%	−52.15%	−91.67%	+154.55%	40.44%	−11.76%
2019 – 2018	−67.89%	−14.61%	+80.00%	–32.86%	−69.74%	0.00%
2020 – 2019	−38.57%	−60.53%	−11.11%	−100.00%	−90.36%	−100.00%
Average	+33.76%	−27.20%	+94.31%	+15.68%	−30.66%	−37.25%
Variance	0.78	0.10	3.48	0.89	0.27	0.20

Appendix B

(See Tables B1-B2).

Table B1
Year over year percentage change in number of reports per macro-sector (by *occurence_date*).

	MS1	MS2	MS3	MS4	MS5	MS6
2000 – N/A	N/A	N/A	N/A	N/A	N/A	N/A
2001 – 2000	N/A	+20.92%	N/A	−4.84%	−8.60%	+55.04%
2002 – 2001	−1.56%	−2.36%	N/A	−15.68%	−8.89%	−29.00%
2003 – 2002	−13.10%	+13.51%	N/A	−17.39%	+25.34%	N/A
2004 – 2003	−57.53%	−9.04%	−12.79%	+25.15%	−20.77%	N/A
2005 – 2004	+116.13%	−1.80%	+22.67%	+14.02%	+6.18%	+16.46%
2006 – 2005	−10.55%	+8.81%	−38.91%	+0.67%	−6.02%	−14.13%
2007 – 2006	+5.18%	−19.89%	−35.35%	+9.31%	+17.29%	0.00%
2008 – 2007	+7.36%	−1.67%	+89.91%	−16.08%	−8.19%	+3.38%
2009 – 2008	+18.11%	+16.03%	−1.81%	−5.46%	+3.94%	−14.04%
2010 – 2009	−13.24%	+3.70%	−4.80%	+4.66%	−2.23%	+9.69%
2011 – 2010	+4.82%	−3.46%	−18.60%	−7.88%	−3.68%	−8.44%
2012 – 2011	−3.89%	−2.57%	+40.95%	+2.55%	+4.08%	+15.60%
2013 – 2012	−0.52%	+1.42%	+10.06%	+7.34%	−4.33%	–22.30%
2014 – 2013	+7.97%	+1.08%	−6.38%	+4.81%	−0.08%	+25.96%
2015 – 2014	+3.99%	−4.43%	−37.29%	−6.46%	+0.64%	+10.16%
2016 – 2015	−15.51%	−6.66%	+20.36%	+6.26%	+3.28%	−26.66%
2017 – 2016	+19.79%	+4.40%	+14.86%	−13.10%	−0.98%	+11.63%
2018 – 2017	−9.60%	+7.86%	−31.43%	−3.93%	+9.31%	+5.65%
2019 – 2018	+11.86%	−6.29%	−14.71%	−9.03%	−10.81%	N/A
2020 – 2019	−34.51%	−2.39%	+69.40%	N/A	+20.86%	N/A
Average	+1.85%	+0.86%	+3.89%	−1.32%	+0.82%	+2.44%
Variance	0.10	0.01	0.12	0.01	0.01	0.04

Table B2
Year over year percentage change in number of reports per macro-sector (by *occurence_date*).

	MS1	MS2	MS3	MS4	MS5	MS6
2000 – N/A	N/A	N/A	N/A	N/A	N/A	N/A
2001 – 2000	N/A	−50.00%	−100.00%	+700.00%	+233.33%	−50.00%
2002 – 2001	+66.67%	+75.00%	N/A	−62.50%	−40.00%	+100.00%
2003 – 2002	−80.00%	0.00%	N/A	–33.33%	+83.33%	−100.00%
2004 – 2003	0.00%	−14.29%	0.00%	+200.00%	−27.27%	N/A
2005 – 2004	+500.00%	–33.33%	0.00%	0.00%	+125.00%	0.00%
2006 – 2005	+150.00%	+75.00%	+400.00%	+50.00%	+122.22%	0.00%
2007 – 2006	+113.33%	+85.71%	−40.00%	−66.67%	+127.50%	0.00%
2008 – 2007	−43.75%	+46.15%	–33.33%	+633.33%	+38.46%	+200.00%
2009 – 2008	–22.22%	−15.79%	+100.00%	–22.73%	−11.11%	+66.67%
2010 – 2009	+107.14%	+118.75%	−50.00%	+70.59%	+36.61%	−80.00%
2011 – 2010	−6.90%	+14.29%	+200.00%	−55.17%	+14.38%	+300.00%
2012 – 2011	+3.70%	+15.00%	+50.00%	+7.69%	+4.57%	+150.00%
2013 – 2012	+7.14%	−10.87%	0.00%	+57.14%	+10.38%	+70.00%
2014 – 2013	+30.00%	+34.15%	−55.56%	+9.09%	+3.47%	−35.29%
2015 – 2014	−15.38%	−21.82%	+175.00%	+70.83%	+18.18%	−27.27%
2016 – 2015	+60.61%	+44.19%	−18.18%	+4.88%	−5.26%	+25.00%
2017 – 2016	+20.75%	0.00%	0.00%	–23.26%	+11.97%	+10.00%
2018 – 2017	–32.81%	−53.23%	0.00%	−24.24%	−40.46%	−54.55%
2019 – 2018	−27.91%	−62.07%	–33.33%	−72.00%	−75.00%	−100.00%
2020 – 2019	−64.52%	+109.09%	−66.67%	−100.00%	−97.44%	N/A
Average	+40.31%	+17.80%	+29.33%	+67.18%	+26.64%	+26.36%
Variance	1.45	0.26	1.33	4.22	0.56	1.02

References

Abu-Salih, B., 2021. Domain-specific knowledge graphs: a survey. J. Netw. Comput. Appl. 185 https://doi.org/10.1016/j.jnca.2021.103076.

Ansaldi, S.M., Agnello, P., Pirone, A., Vallerotonda, M.R., 2021. Near miss archive: a challenge to share knowledge among inspectors and improve seveso inspections. Sustainability 13. https://doi.org/10.3390/su13158456.

Brunelli, M., 2015. Introduction to the Analytic Hierarchy Process, Springer Briefs in Operations Research. Springer Cham. 10.1007/978-3-319-12502-2.

Bugalia, N., Maemura, Y., Ozawa, K., 2021. A system dynamics model for near-miss reporting in complex systems. Saf. Sci. 142 https://doi.org/10.1016/j.ssci.2021.105368.

Caspi, H., Perlman, Y., Westreich, S., 2023. Managing near-miss reporting in hospitals: the dynamics between staff members' willingness to report and management's handling of near-miss events. Saf. Sci. 164, 106147 https://doi.org/10.1016/j.ssci.2023.106147.

Casson Moreno, V., Marroni, G., Landucci, G., 2022. Probabilistic assessment aimed at the evaluation of escalating scenarios in process facilities combining safety and security barriers. Reliab. Eng. Syst. Saf. 228, 108762 https://doi.org/10.1016/j.ress.2022.108762.

Dekker, S., 2019. Foundations of safety science: a century of understanding accidents and disasters. Routledge. https://doi.org/10.4324/9781351059794.

EU Council, 2012. DIRECTIVE 2012/18/EU On the control of major accident hazards involving dangerous substances. Off. J. Eur. Union L197, 1–37.

Hughes, P., Robinson, R., Figueres-Esteban, M., van Gulijk, C., 2019. Extracting safety information from multi-lingual accident reports using an ontology-based approach. Saf. Sci. 118, 288–297.

Khan, N., Ma, Z., Ullah, A., Polat, K., 2022. Categorization of knowledge graph based recommendation methods and benchmark datasets from the perspectives of application scenarios: a comprehensive survey. Expert Syst. Appl. 206 https://doi.org/10.1016/j.eswa.2022.117737.

Li, X., Lyu, M., Wang, Z., Chen, C.-H., Zheng, P., 2021. Exploiting knowledge graphs in industrial products and services: a survey of key aspects, challenges, and future perspectives. Comput. Ind. 129 https://doi.org/10.1016/j.compind.2021.103449.

Misuri, A., Landucci, G., Cozzani, V., 2021. Assessment of risk modification due to safety barrier performance degradation in Natech events. Reliab. Eng. Syst. Saf. 212, 107634.

Newman, M., 2010. Networks: An Introduction, Networks: An Introduction. Oxford University Press. 10.1093/acprof:oso/9780199206650.001.0001.

Pedrosa, M.H., Guedes, J.C., Dias, I., Salazar, A., 2022. New approaches of near-miss management in industry: a systematic review. Stud. Syst. Decis. Control 406, 109–120. https://doi.org/10.1007/978-3-030-89617-1_10.

Peng, F.-L., Qiao, Y.-K., Yang, C., 2023. Building a knowledge graph for operational hazard management of utility tunnels. Expert Syst. Appl. 223, 119901 https://doi.org/10.1016/j.eswa.2023.119901.

Phimister, J.R., Oktem, U., Kleindorfer, P.R., Kunreuther, H., 2003. Near-miss incident management in the chemical process industry. Risk Anal. 23, 445–459. https://doi.org/10.1111/1539-6924.00326.

Saaty, T.L., 1990. How to make a decision: the analytic hierarchy process. Eur. J. Oper. Res. 48, 9–26. https://doi.org/10.1016/0377-2217(90)90057-I.

Simone, F., Ansaldi, S.M., Agnello, P., Patriarca, R., 2023. Industrial safety management in the digital era: constructing a knowledge graph from near misses. Comput. Ind. 146, 103849 https://doi.org/10.1016/j.compind.2022.103849.

Studer, R., Benjamins, V.R., Fensel, D., 1998. Knowledge engineering: principles and methods. Data Knowl. Eng. 25, 161–197. https://doi.org/10.1016/S0169-023X(97)00056-6.

Wang, Q., Mao, Z., Wang, B., Guo, L., 2017. Knowledge graph embedding: a survey of approaches and applications. IEEE Trans. Knowl. Data Eng. 29, 2724–2743. https://doi.org/10.1109/TKDE.2017.2754499.

Zhang, P., Wang, T., Yan, J., 2022. PageRank centrality and algorithms for weighted, directed networks. Phys. A Stat. Mech. its Appl. 586 https://doi.org/10.1016/j.physa.2021.126438.

Zhu, R., Hu, X., Bai, Y., Li, X., 2021. Risk analysis of terrorist attacks on LNG storage tanks at ports. Saf. Sci. 137 https://doi.org/10.1016/j.ssci.2021.105192.