MHKG: AN ONTOLOGY-GUIDED, EVIDENCE-LINKED KNOWLEDGE GRAPH FOR METAL EXPOSURE AND HEALTH

Anonymous authors
Paper under double-blind review

ABSTRACT

Preventing metal exposure and related diseases is an important public health issue. However, related research involves various types of information and the relevant evidence is scattered across a large number of documents. The same concepts often appear in abbreviated forms, aliases, or different granular terms, which limits cross-study retrieval and evidence integration. We bulid a Metal-Health Knowledge Graph (MHKG) dataset, which organizes the literature extraction results into a queryable and auditable structured resource. MHKG links exposure factors with biological responses, health outcomes, and intervention measures, and the relationships retain links to the original articles and supporting paragraphs. We construct the dataset using a unified schema to organize heterogeneous entities and relationships. Meanwhile, cross-document entity standardization and evidence verification make the graph more coherent and the content more accessible for auditing. MHKG provides a structured and auditable resource for studying occupational metal exposure, supporting evidence-oriented retrieval, knowledge exploration, as well as future knowledge-enhanced applications.

1 INTRODUCTION

Occupational exposure to metals (Chaudhary et al., 2026; Khoshakhlagh et al., 2024) has long been a significant research topic in the fields of occupational health and environmental medicine. Over the past few decades, a great deal of research has been conducted on various aspects of this exposure spectrum. However, the relevant knowledge is typically scattered across different research topics and various scientific literature, lacking a unified structured organization, which limits the ability of researchers to retrieve coherent evidence chains (Donzelli et al., 2024; Grondin et al., 2016). Moreover, the fragmentation of knowledge restricts its practicality in advanced applications such as literature retrieval, cross-domain correlation analysis, and intelligent knowledge-driven decision support systems.

Despite the existence of numerous biomedical databases and knowledge graphs (Mattingly et al., 2012; Davis et al., 2023), exposure-related knowledge remains distributed across heterogeneous resources. Many established biomedical knowledge bases were originally organized around molecular relationships among chemicals, genes, phenotypes, and diseases, whereas detailed exposure context has required dedicated extensions and exposure-oriented ontologies. This limitation is particularly relevant to occupational health, where information about occupational settings, exposed populations, and exposure routes is crucial for linking environmental stressors with their health outcomes (Grondin et al., 2018). Consequently, general biomedical resources do not always provide the sufficient background information needed for occupational health research regarding specific exposure scenarios. In recent years, large language models (LLMs) (Vaswani et al., 2017) have provided a new technical approach for structured knowledge extraction (van Cauter & Yakovets, 2024) from scientific literature. However, ensuring the quality of the extracted content with unconstrained generative extraction is difficult (Huang et al., 2025). Furthermore, outputs from disparate sources require further processing before they can be transformed into stable, integratable resources for knowledge graph population (Zhang & Soh, 2024; Nicholson & Greene, 2020). Therefore, the process of constructing a final knowledge graph from original scientific literature involves not only

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

knowledge extraction, but also a unified semantic schema and a quality control mechanism running through the stages of data generation, review, and integration.

To overcome these limitations, we construct an ontology-constrained knowledge graph of occupational metal exposure and its associated health effects. First, we design a domain ontology (Mattingly et al., 2012; Pandithawatta et al., 2023) that focuses on the key elements of this research. This ontology covers key concepts such as industry scenarios, occupational populations, metals, exposure pathways, pathogenic mechanisms, biomarkers, diseases, clinical diagnostic methods, and preventive measures. It also describes the semantic relationships among these concepts. Next, we use relevant scientific literature as the main data source, processing original PDF documents via text parsing, structured segmentation, and content cleaning. Within the constraints of the established ontology, which is developed through expert discussion and review, we utilize LLMs to identify entities and their relationships from the documents (van Cauter & Yakovets, 2024; Oarga et al., 2026). Through this approach, we transform the knowledge originally scattered in natural language texts into structured candidate knowledge with unified entity types and semantic relationships (Olasunkanmi et al., 2026; Lan et al., 2026).

Our key contributions can be summarized as follows:

- We construct a unified framework for representing domain knowledge in the field of occupational metal exposure and health research, providing a structured data foundation for related studies.
- We establish a process for constructing knowledge graphs driven by scientific literature, integrating PDF document parsing, structured knowledge extraction, data auditing, entity normalization, entity disambiguation, entity fusion, and graph construction into a unified data production chain.
- We implement quality control of the knowledge generation process through a combination of pattern constraints, automated data auditing, and manual review, ensuring that the final resources can serve as the data basis for subsequent occupational health knowledge retrieval and discovery, and intelligent applications for knowledge enhancement.

2 Related Work

Biomedical and Occupational Health Knowledge Resources. In recent years, the development of biomedical knowledge graphs has gradually expanded from modeling single-type associations to integrating multi-source data. A systematic review of existing medical knowledge graph systems (Cui et al., 2025) indicates that the current resources have covered electronic health records, biomedical databases, scientific literature, as well as various clinical and biological knowledge, and have been further utilized for knowledge discovery, prediction, and decision support. Meanwhile, PubMed Knowledge Graph 2.0 (Xu et al., 2024a) connects over 36 million papers with scientific resources such as patents and clinical trials, demonstrating the value of the knowledge graph in organizing cross-source biomedical resources. Furthermore, eKG (Consoli et al., 2025) is constructed by leveraging multiple large language models to extract epidemiological information from the World Health Organization's disease outbreak reports.

Knowledge from Scientific Literature. Scientific literature is an important data source for domain knowledge graphs (DKGs). In recent years, several studies have begun to explore the use of LLMs to construct DKGs from unstructured biomedical texts. Existing research (Arsenyan et al., 2024) has systematically evaluated the ability of various LLMs to identify information such as diseases, treatments, and clinical manifestations from electronic medical record texts, and has further utilized the extracted results to construct biomedical knowledge graphs. There are also related studies (Xu et al., 2024b) that have combined the reviews of engineering and medical experts in the field of heart failure, and have constructed a disease KG. BioKGrapher (Schäfer et al., 2024) further explored the automatic construction of a biomedical knowledge graph from PubMed literature and conducted downstream application evaluation of the automatically obtained knowledge. Although LLMs significantly reduce the reliance of domain knowledge extraction on large-scale manually annotated data, the reliability of its generated results remains a key issue. GraphJudge (Huang et al., 2025) points out that noise in real documents, deviations in understanding of professional domain knowledge, and model hallucinations all affect the quality of the final knowledge graph.

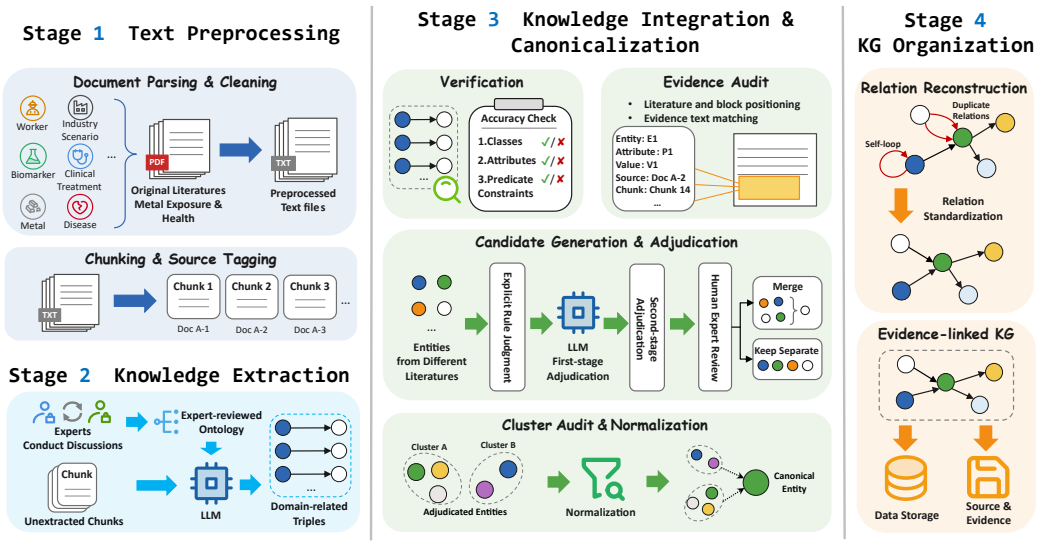

Figure 1: Overview of our work. Literature collected from PubMed is parsed and segmented into text fragments, which are then associated with their sources, from which ontology-constrained entities, relations, and supporting evidence are extracted. The extracted records then undergo validation, normalization, and disambiguation before being integrated into canonicalized entities and relations, while the source records are kept separately.

Entity Normalization and Cross-Document Resolution. Biomedical literature refers to the same concept through abbreviations, synonyms, and orthographic variants, making entity normalization essential to the construction of coherent knowledge resources (French & McInnes, 2023). Methods such as BioSyn and SapBERT learn representations that align diverse mentions with biomedical concepts (Sung et al., 2020; Liu et al., 2021). In a literature-derived knowledge graph, normalization also determines whether relations extracted from different documents converge on the same entity while preserving distinctions among chemically or clinically different concepts.

Ontology-Guided KG Construction. Predefined ontology is not only used to describe the semantic structure of the final knowledge graph, but also can serve as an explicit constraint in the knowledge extraction process. SPIRES (Caufield et al., 2024) recursively calls LLMs based on the user-defined knowledge patterns to conduct structured information extraction, and uses the existing ontologies and controlled vocabularies to perform semantic alignment on the extracted results. OSKGC (Wang & Iwaihara, 2025) clearly defined ontology schema-based knowledge graph construction as a task of constructing a knowledge graph from text, requiring the extracted results to simultaneously satisfy the accuracy of text facts and the consistency of the predefined ontology structure. OntoGRC (Zhang et al., 2026) further embeds the concepts, relationships, and domain–range constraints in the ontology into the multi-stage LLM prompts, improving the semantic consistency of the extracted results through the generation, reflection, and correction processes. Based on this idea, we use the ontology that has been reviewed by domain experts and remains stable as the unified semantic framework for the entire resource construction process. In the knowledge extraction stage, we constrain the entity categories and semantic relationships, and further conduct data auditing, entity normalization, and disambiguation after extraction to ensure that cross-document knowledge can enter the final knowledge graph in a unified pattern.

3 KG Construction

In this study, we construct a literature-based knowledge graph to systematically organize entities and relations relevant to heavy metal exposure and health effects, transforming scattered information from heterogeneous sources into structured knowledge with unified identifiers and clear provenance. Our workflow consists of four main stages: text processing of the literature, ontology-guided knowledge extraction, cross-document knowledge fusion, and graph construction with provenance-

preserving storage, as shown in the overview of our work (Figure 1). Throughout the extraction and validation process, we apply domain ontologies that had been audited by experts to ensure semantic consistency. To support complete traceability, we save the original literature and extraction logs separately, enabling reviewers and future users to trace the original evidence for generating the knowledge (Sikos & Philp, 2020).

3.1 TEXT PREPROCESSING

We use Unstructured to parse the full text of the raw PDFs of the target literature collected from PubMed, which we then organize into coherent chunks based on the documented heading hierarchy. To balance input length constraints against the preservation of local semantic context, we set a maximum chunk size of 1,800 characters. New chunks preferentially initiate once the running length exceeds 1,400 characters. Chunks shorter than 500 characters are merged according to our chunking rules, while overly long segments are split with a 150-character overlap to maintain continuity. Our text cleaning procedure involves removing isolated page numbers, recurring patterns indicative of headers and footers, redundant whitespace, and any identifiable encoding artifacts. To minimize interference from non-essential content, we truncate the text at the first occurrence of terminal section headings such as References and Acknowledgments. For each retained text chunk, we record three pieces of metadata: the literature identifier, the cleaned chunk index, and the corresponding original chunk index. This design enables all extracted results to be traced back to the exact source text without relying on entity name-based reverse retrieval.

3.2 KNOWLEDGE EXTRACTION

We develop the MHKG ontology as the framework for extracting and integrating knowledge on metal exposure and health-related topics from various literature. First, we obtain the relevant ontology for this topic through discussions and audits by several domain experts, which is used to organize knowledge about metal exposure, health impacts, prevention and control measures. The ontology covers key concepts such as industrial scenarios, occupational exposure populations, exposure pathways, heavy metals, pathogenic mechanisms, biomarkers, diseases, diagnosis, treatment, and preventive interventions. Figure 2 shows the part of the ontology the we define. These categories are interrelated through predefined relationships, while class attributes are used to distinguish chemical forms, sample types, detection items, and intervention plans. To extract the defined patterns, it also serves as an authoritative source for subsequent data validation, ensuring that the knowledge generation is semantically consistent with the preset norms. We integrate the ontology definition, extraction rules, output structure, and a set of concise examples into a unified prompt template, and then use this template to instruct the LLM to extract entities, relationships, and supporting evidence from each text. Some prompt fragments are shown in Appendix D. The LLM is instructed to generate records based solely on the content of the current fragment, without introducing external knowledge or supplementing unsupported facts. For each extracted entity and relationship, we require the model to retain the original text evidence and its corresponding fragment source. The more detailed extraction process can be found in Appendix C.

We define ten top-level classes as follows. (1) *Industry Scenario* refers to a human industrial or occupational setting in which metal exposure may occur, encompassing specific workplaces and production processes; (2) *Occupational Population* represents groups of workers engaged in particular occupations or job activities who may be exposed to metals in the workplace; (3) *Exposure Route* denotes the pathway through which a metal enters or comes into contact with the human body, including inhalation, ingestion, or dermal contact; (4) *Metal* refers to either a single metal element or a mixture of multiple metals that may pose occupational health risks; (5) *Pathogenic Mechanism* represents the biological or pathophysiological process through which metal exposure causes adverse health effects; (6) *Biomarker* refers to a measurable biological indicator associated with metal exposure, toxic effects, or individual susceptibility; (7) *Prevention Measure* represents an intervention adopted to reduce metal exposure or prevent and control related diseases; (8) *Disease* denotes a pathological condition or adverse health outcome associated with occupational metal exposure, characterized by the affected organ or system and its corresponding clinical manifestations; (9) *Clinical Treatment* refers to a therapeutic intervention or treatment protocol used to alleviate, control, or manage diseases caused or aggravated by metal exposure; and (10) *Clinical Diagnostic*

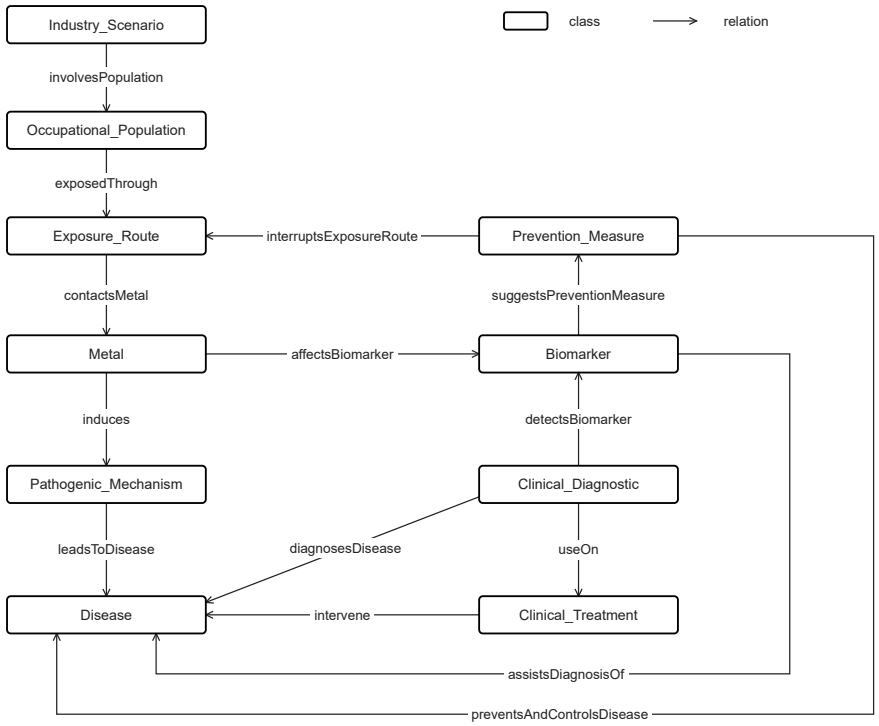

Figure 2: A part of the MHKG ontology.

represents the methods and procedures used to identify metal exposure, detect relevant biomarkers, and diagnose associated diseases.

We also define sixteen relations, which are further categorized into four types as follows.

Occupational Exposure Relation. This relation type models the industrial and occupational contexts in which heavy-metal exposure occurs. We define five occupational exposure relations. Specifically, `hasWorkplaceAndProcess` represents the relation between an industry scenario and its associated workplace or production process. `involvesPopulation` describes the occupational population involved in a particular industry scenario. `hasJobType` links an occupational population to its specific job type or work responsibility. `exposedThrough` represents the exposure pathway through which an occupational population comes into contact with a metal, such as inhalation, ingestion, or dermal contact. `contactsMetal` links an exposure route to the corresponding metal involved in the exposure process.

Pathogenic and Biomarker Relation. This relation type models the biological effects of metal exposure and the association between metals, pathogenic mechanisms, diseases, and biomarkers. We define four relations in this category. Specifically, `induces` represents the relation in which a metal induces a pathogenic mechanism. `leadsToDisease` describes the relation in which a pathogenic mechanism contributes to or leads to a disease. `affectsBiomarker` links a metal to a biomarker whose level or expression is affected by metal exposure. `assistsDiagnosisOf` represents the relation in which a biomarker supports the diagnosis or evaluation of a disease.

Prevention and Control Relation. This relation type describes measures used to reduce metal exposure and prevent or control exposure-related diseases. We define three prevention-related relations. Specifically, `suggestsPreventionMeasure` represents the relation in which a biomarker indicates or suggests that a specific prevention measure should be adopted.

`interruptsExposureRoute` links a prevention measure to the exposure route that it interrupts or reduces. `preventsAndControlsDisease` describes the relation in which a prevention measure is used to prevent or control a disease associated with metal exposure.

Clinical Diagnosis and Treatment Relation. This relation type covers the clinical procedures used to detect exposure-related biomarkers, diagnose diseases, and guide therapeutic interventions. We define four relations in this category. Specifically, `detectsBiomarker` represents the relation in which a clinical diagnostic method or examination detects a biomarker. `diagnosesDisease` links a clinical diagnostic method to the disease that it helps identify. `useOn` describes the relation in which a clinical diagnostic method is used to select, guide, or evaluate a clinical treatment. intervene represents the relation in which a clinical treatment acts on or manages a disease. The complete class hierarchy and the object- and data-property specifications of our ontology are provided in Appendices A and B, respectively. The complete class hierarchy and the object- and data-property specifications of our ontology are provided in Appendices A and B, respectively.

At this stage, we aggregate extracted records without asserting entity equivalence solely from identical names across chunks. Cross-document synonymy and coreference resolution were deferred until the subsequent fusion process.

3.3 Knowledge Integration and Canonicalization

First, we validate the extracted results against schema constraints by excluding relationships that violate endpoint or type constraints without using the model to infer missing endpoints or rewrite unrecognized predicates. Attribute anomalies are logged and filtered during the final node generation. Evidence is verified by locating source chunks; unlocatable records are flagged but not automatically deleted, thereby separating provenance issues from knowledge errors. We then normalize entity names by unifying case, whitespace, and punctuation, and match them against a domain dictionary using labels and category-specific identity attributes (Sung et al., 2022; Liu et al., 2021). Unmatched entities receive the first non-empty identity attribute in priority order, or fall back to the label. We generate temporary identifiers from the canonical name and category, retaining only the first occurrence per document. Document-level alignment is only perform within the same category.

Based on these normalized and category-constrained representations, we subsequently identify potentially equivalent entities and subject them to a multi-stage resolution process. Candidate resolution follows a three-tiered pipeline comprising deterministic rules, LLM judgment (Ye & Mitchell, 2025), and human review. The rules accept punctuation and spacing, safe plural variants, and known spelling variants while rejecting candidates with substantive semantic differences, such as sample types, laterality, numerical subtypes, and metal speciation. Candidates that the rules could not determine are assessed by the model using labels, attributes, and up to two evidence excerpts per entity. Decisions classify as SAME CONCEPT, DIFFERENT CONCEPT, or INSUFFICIENT EVIDENCE. First-round outcomes are recorded by category, and unresolved candidates proceed to a second round.

In the second round, we add labels from already-accepted entity clusters and existing rejection relationships to the input of LLM, which let the LLM use prior alignment outcomes, not just individual names. Then LLM return a decision d, a confidence score q, and a conflict indicator F. The state update rules were defined as follows:

$$S = \begin{cases} \text{ACCEPTED}, & d = \text{ACCEPT},\ q \geq \tau_{\text{acc}},\ F = \varnothing, \\ \text{REJECTED}, & d = \text{REJECT},\ q \geq \tau_{\text{rej}}, \\ \text{REVIEW}, & \text{otherwise.} \end{cases}$$

We use asymmetric decision thresholds, with $\tau_{\text{acc}} > \tau_{\text{rej}}$, because an accepted identity link can propagate through the entity clusters. An incorrect acceptance may merge distinct concepts and consolidate their associated relations and provenance under a single canonical entity. Therefore, acceptance requires both higher model-reported confidence and the absence of a detected cluster conflict ($F = \varnothing$). A rejection leaves the entities separate without propagating an identity link. Cases that satisfy neither condition remain in REVIEW for manual adjudication.

After candidate-level adjudication, we group the accepted edges together with dictionary-equivalence edges to form connected components, or entity clusters. We then examine these clusters

for risks of transitivity-induced merging. Only clusters that are deemed safe, or those that had received manual approval, proceeded to canonical entity selection. For canonical name selection, we prioritize dictionary-preferred terms when available. Otherwise, candidates are ranked based on their frequency of occurrence, average similarity to other labels within the cluster, and textual quality. To mitigate identifier instability caused by label adjustments or repeated runs, we assign canonical identifiers through a persistent registry.

3.4 KG Organization

Based on the reviewed mappings, we convert the temporary entity identifiers into final canonical identifiers and update the corresponding relationship endpoints accordingly. Then we recheck the predicate and type constraints on the mapped relationships, discarding those with unmappable endpoints, those associated with excluded entities, and those that formed self-loops after fusion. We merge records that share the same canonical subject, predicate, and canonical object into a single triple. To prevent evidence loss during relation deduplication, we store the graph structure and provenance records separately. We assign a unique relation identifier to each canonical triple, linking it to all retained provenance records, including literature and chunk information, evidence text, relation attributes, and the original confidence scores. The final output consists of canonical nodes, deduplicated relations, and the corresponding relation-level provenance data.

4 MHKG Resource

Following ontology-guided extraction, cross-document entity resolution, and provenance-preserving graph organization, MHKG provides a canonical representation of knowledge about occupational metal exposure and its health effects while retaining explicit links from graph statements to their source evidence. In this section, we demonstrate the scale and semantic coverage of the resulting graph, the representation of supporting evidence, and the formats through which the resource can be queried and reused.

4.1 Resource Scale and Statistics

MHKG was constructed from full texts of 1,086 scientific publications, which were parsed into 34,410 text chunks. Ontology-guided extraction identified 165,791 entity mentions and 76,040 relation-source records. The details are as shown in Table 1. An entity mention is an expression as it appears in a source document, such as `Pb` and `lead`. A canonical entity is the unified graph node produced after normalization, disambiguation, and fusion, such as `Lead`. This distinction is essential because repeated mentions, abbreviations, spelling variants, and context-dependent expressions may refer to the same canonical entity.

4.2 Evidence and Provenance Representation

MHKG represents each deduplicated graph statement as a canonical relation with a stable relation identifier, a canonical subject identifier, a predicate, and a canonical object identifier. The graph table and the provenance table are stored separately and joined by the relation identifier which maintains the compactness of the graph without collapsing the evidence records that gave rise to the canonical triple (Kuhn et al., 2018). As a result, one graph edge can be inspected at the level of individual source documents and text chunks.

This separation serves three resource functions. First, relation deduplication does not discard repeated or independently reported evidence. Second, users can trace a graph statement to its source document and local text context without reconstructing provenance from entity labels. Third, the same representation supports evidence-oriented retrieval, manual auditing, and incremental updates, because an individual provenance record can be reviewed or replaced without altering the identity of an otherwise unchanged canonical relation.

The large number of entity mentions and relation-source records is a normal characteristic of evidence-linked resources. A canonical entity can integrate expressions from various mentions and publications, and the same holds for canonical relations, but normalization can eliminate redundant representations in the graph.

Table 1: Overall statistics of MHKG.

Resource Component	Count
Scientific publications	1,086
Text chunks	34,410
Entity mentions	165,801
Canonical entities	46,384
Canonical relations	60,663
Relation-source records	76,041

4.3 KNOWLEDGE COVERAGE

MHKG organizes occupational metal-exposure knowledge under ten top-level ontology classes. These classes represent a broader knowledge space than a collection of chemical-disease associations. The central exposure-to-outcome chain can be traversed as follows: `Industry Scenario` → `Occupational Population` → `Exposure Route` → `Metal` → `Pathogenic Mechanism` → `Disease`. Complementary branches connect metals to biomarkers, biomarkers to diseases, diagnostic procedures to diseases and so on. These paths support movement between occupational context, exposure, biological response, health outcome, and intervention within a common semantic structure.

5 CONCLUSION

In this paper, we propose MHKG, aiming to systematically extract and integrate scattered knowledge from large-scale domain literature, and to structurally model the domain entities and their complex semantic relationships. In order to guarantee the quality of the graph, we construct an automated process covering pattern-constrained knowledge extraction, structure and evidence auditing, entity fusion and normalization, as well as relationship reconstruction. This process enables a reliable conversion of knowledge from unstructured literature to a structured graph and traceability of the sources. By providing standardized, traceable, and reusable domain knowledge resources, this research not only provides a data foundation for related knowledge discovery and knowledge enhancement tasks, but also its construction method can be extended to other professional fields with large-scale unstructured literature, thereby promoting the systematic organization and utilization of domain knowledge.

DATA AVAILABILITY

MHKG dataset used in this study is available at https://github.com/ymirAkarin/MHKG. The repository contains entity, relationship, and relation-provenance files together with the RDF/XML ontology.

AI USE STATEMENT

In this work, we used generative AI tools for polishing language and optimization of the transitions between paragraphs in the paper. We have not used generative AI tools for the design of knowledge graph ontology. We applied generative AI tools for KG data extraction, auditing and standardization. Additionally, we used generative AI tools for assisting with writing and debugging of some of the code for data processing and knowledge graph construction. We have reviewed all AI-assisted work. All the code generated by AI assistance is reviewed, tested and verified by the author manually. It is also modified according to the actual data. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Vahan Arsenyan, Spartak Bughdaryan, Fadi Shaya, Kent Wilson Small, and Davit Shahnazaryan. Large language models for biomedical knowledge graph construction: Information extraction from EMR notes. In Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii (eds.), *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pp. 295–317, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.bionlp-1.23. URL https://aclanthology.org/2024.bionlp-1.23/.

J. Harry Caufield, Harshad Hegde, Vincent Emonet, Nomi L. Harris, Marcin P. Joachimiak, Nicolas Matentzoglu, HyeongSik Kim, Sierra Moxon, Justin T. Reese, Melissa A. Haendel, Peter N. Robinson, and Christopher J. Mungall. Structured prompt interrogation and recursive extraction of semantics (spires): a method for populating knowledge bases using zero-shot learning. *Bioinformatics (Oxford, England)*, 40(3):btae104, 2024. doi: 10.1093/bioinformatics/btae104. URL https://doi.org/10.1093/bioinformatics/btae104.

Meghna Dipakkumar Chaudhary, Rajvi D. Chaudhary, Kshitij Sale, Chaitanya Chaudhary, Pranav Manek, Nirav Parmar, Miral Mehta, and Dhaval Niranjan Mehta. Occupational exposure to heavy metals: A review of health outcomes and safety interventions. *Bioinformation*, 22(1):547–552, 2026. ISSN 0973-2063. doi: 10.6026/973206300220547. PMID: 41960540.

Sergio Consoli, Pietro Coletti, Peter V. Markov, Lia Orfei, Indaco Biazzo, Lea Schuh, Nicolas Stefanovitch, Lorenzo Bertolini, Mario Ceresa, and Nikolaos I. Stilianakis. An epidemiological knowledge graph extracted from the world health organization's disease outbreak news. *Scientific Data*, 12(1):970, 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-05276-2. URL https://doi.org/10.1038/s41597-025-05276-2.

Hejie Cui, Jiaying Lu, Ran Xu, Shiyu Wang, Wenjing Ma, Yue Yu, Shaojun Yu, Xuan Kan, Chen Ling, Liang Zhao, Zhaohui S. Qin, Joyce C. Ho, Tianfan Fu, Jing Ma, Mengdi Huai, Fei Wang, and Carl Yang. A review on knowledge graphs for healthcare: Resources, applications, and promises. *Journal of Biomedical Informatics*, 169:104861, 2025. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2025.104861. URL https://www.sciencedirect.com/science/article/pii/S1532046425000905.

Allan Peter Davis, Thomas C Wiegers, Robin J Johnson, Daniela Sciaky, Jolene Wiegers, and Carolyn J Mattingly. Comparative toxicogenomics database (ctd): update 2023. *Nucleic Acids Research*, 51(D1):D1257–D1262, 01 2023. ISSN 0305-1048. doi: 10.1093/nar/gkac833. URL https://doi.org/10.1093/nar/gkac833.

Gabriele Donzelli, Ronette Gehring, Sivakumar Murugadoss, Tom Roos, Alexandra Schaffert, and Nunzia Linzalone. A critical review on the toxicological and epidemiological evidence integration for assessing human health risks to environmental chemical exposures. *Reviews on Environmental Health*, 40:427–436, 12 2024. doi: 10.1515/reveh-2024-0072.

Evan French and Bridget T. McInnes. An overview of biomedical entity linking throughout the years. *Journal of Biomedical Informatics*, 137:104252, 2023. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2022.104252. URL https://www.sciencedirect.com/science/article/pii/S153204642200257X.

Cynthia J. Grondin, Allan Peter Davis, Thomas C. Wiegers, Benjamin L. King, Jolene A. Wiegers, David M. Reif, Jane A. Hoppin, and Carolyn J. Mattingly. Advancing exposure science through chemical data curation and integration in the comparative toxicogenomics database. *Environmental Health Perspectives*, 124(10):1592–1599, Oct 2016. doi: 10.1289/EHP174.

Cynthia J. Grondin, Allan Peter Davis, Thomas C. Wiegers, Jolene A. Wiegers, and Carolyn J. Mattingly. Accessing an expanded exposure science module at the comparative toxicogenomics database. *Environmental Health Perspectives*, 126(1):014501, 2018. ISSN 0091-6765. doi: 10.1289/EHP2873. PMID: 29351546; PMCID: PMC6014688.

Haoyu Huang, Chong Chen, Zeang Sheng, Yang Li, and Wentao Zhang. Can LLMs be good graph judge for knowledge graph construction? In Christos Christodoulopoulos, Tanmoy Chakraborty,

Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10929–10948, Suzhou, China, nov 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.554. URL https://aclanthology.org/2025.emnlp-main.554/.

Amir Hossein Khoshakhlagh, Safiye Ghobakhloo, and Agnieszka Gruszecka-Kosowska. Inhalational exposure to heavy metals: Carcinogenic and non-carcinogenic risk assessment. *Journal of Hazardous Materials Advances*, 16:100485, 2024. ISSN 2772-4166. doi: https://doi.org/10.1016/j.hazadv.2024.100485. URL https://www.sciencedirect.com/science/article/pii/S277241662400086X.

Tobias Kuhn, Albert Meroño-Peñuela, Alexander Malic, Jorrit H. Poelen, Allen H. Hurlbert, Emilio Centeno Ortiz, Laura I. Furlong, Núria Queralt-Rosinach, Christine Chichester, Juan M. Banda, Egon Willighagen, Friederike Ehrhart, Chris Evelo, Tareq B. Malas, and Michel Dumontier. Nanopublications: A growing resource of provenance-centric scientific linked data. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, pp. 83–92, 2018. doi: 10.1109/eScience.2018.00024.

Wuyang Lan, Siqi Zhang, Wenzheng Wang, Ke Hu, Tianrun Gao, Lei Shi, Zongbo Han, Yanjun Chen, Hao Zhang, Song Wu, Xiaohong Liu, and Guangyu Wang. A unified framework and benchmark for generalizable biomedical knowledge extraction and applications with large language models. *Cell Reports Medicine*, 7(8):102975, 2026. ISSN 2666-3791. doi: https://doi.org/10.1016/j.xcrm.2026.102975. URL https://www.sciencedirect.com/science/article/pii/S2666379126003927.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4228–4238, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.334. URL https://aclanthology.org/2021.naacl-main.334/.

Carolyn J. Mattingly, Thomas E. McKone, Michael A. Callahan, Judith A. Blake, and Elaine A. Cohen Hubal. Providing the missing link: The exposure science ontology exo. *Environmental Science & Technology*, 46(6):3046–3053, Mar 2012. doi: 10.1021/es2033857.

David N. Nicholson and Casey S. Greene. Constructing knowledge graphs and their biomedical applications. *Computational and Structural Biotechnology Journal*, 18:1414–1428, 2020. doi: 10.1016/j.csbj.2020.05.017. URL https://spj.science.org/doi/abs/10.1016/j.csbj.2020.05.017.

Alexandru Oarga, Matthew Hart, Andres M. Bran, Magdalena Lederbauer, and Philippe Schwaller. Scientific knowledge graph and ontology generation using open large language models. *Digital Discovery*, 5(3):1269–1279, 03 2026. ISSN 2635-098X. doi: 10.1039/d5dd00275c. URL https://doi.org/10.1039/d5dd00275c.

Olawumi Olasunkanmi, Matthew Satusky, Hong Yi, Chris Bizon, Harlin Lee, and Stanley Ahalt. RELATE: Relation extraction in biomedical abstracts with LLMs and ontology constraints. In Peniel Argaw, Haoran Zhang, Sarah Jabbour, Payal Chandak, Jerry Ji, Sumit Mukherjee, Olawale Salaudeen, Trenton Chang, Elizabeth Healey, Fabian Gröger, Amin Adibi, Stefan Hegselmann, Benjamin Wild, and Ayush Noori (eds.), *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297 of *Proceedings of Machine Learning Research*, pp. 1178–1193. PMLR, 13–14 Dec 2026. URL https://proceedings.mlr.press/v297/olasunkanmi26a.html.

Sonali Pandithawatta, Seungjun Ahn, Raufdeen Rameezdeen, Christopher W. K. Chow, Nima Gorjian, and Tae Wan Kim. Development of a knowledge graph for automatic job hazard analysis: The schema. *Sensors*, 23(8), 2023. ISSN 1424-8220. URL https://www.mdpi.com/1424-8220/23/8/3893.

Henning Schäfer, Ahmad Idrissi-Yaghir, Kamyar Arzideh, Hendrik Damm, Tabea M.G. Pakull, Cynthia S. Schmidt, Mikel Bahn, Georg Lodde, Elisabeth Livingstone, Dirk Schadendorf, Felix Nensa, Peter A. Horn, and Christoph M. Friedrich. Biokgrapher: Initial evaluation of automated knowledge graph construction from biomedical literature. *Computational and Structural Biotechnology Journal*, 24:639–660, 2024. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2024.10.017. URL https://www.sciencedirect.com/science/article/pii/S2001037024003386.

Leslie F. Sikos and Dean Philp. Provenance-aware knowledge representation: A survey of data models and contextualized knowledge graphs. *Data Science and Engineering*, 5(3):293–316, 2020. ISSN 2364-1541. doi: 10.1007/s41019-020-00118-0. URL https://doi.org/10.1007/s41019-020-00118-0.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. Biomedical entity representations with synonym marginalization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3641–3650, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.335. URL https://aclanthology.org/2020.acl-main.335/.

Mujeen Sung, Minbyul Jeong, Yonghwa Choi, Donghyeon Kim, Jinhyuk Lee, and Jaewoo Kang. Bern2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics*, 38(20):4837–4839, 10 2022. ISSN 1367-4811. doi: 10.1093/bioinformatics/btac598. URL https://doi.org/10.1093/bioinformatics/btac598.

Zeno van Cauter and Nikolay Yakovets. Ontology-guided knowledge graph construction from maintenance short texts. In Russa Biswas, Lucie-Aimée Kaffee, Oshin Agarwal, Pasquale Minervini, Sameer Singh, and Gerard de Melo (eds.), *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pp. 75–84, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.kallm-1.8. URL https://aclanthology.org/2024.kallm-1.8/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Dali Wang and Mizuho Iwaihara. Oskgc: A benchmark for ontology schema-based knowledge graph construction from text. *CEUR Workshop Proceedings*, 4041, 2025. ISSN 1613-0073. Publisher Copyright: © 2025 Copyright for this paper by its authors.; Joint of the 3rd Workshop on Knowledge Base Construction from Pre-Trained Language Models and the 4th Challenge on Language Models for Knowledge Base Construction, KBC-LM+LM-KBC 2025 ; Conference date: 02-11-2025 Through 02-11-2025.

Jian Xu, Chao Yu, Jiawei Xu, Ying Ding, Vetle I. Torvik, Jaewoo Kang, Mujeen Sung, and Minju Song. Pubmed knowledge graph 2.0: Connecting papers, patents, and clinical trials in biomedical science. *ArXiv*, abs/2410.07969, 2024a. URL https://api.semanticscholar.org/CorpusID:279448162.

Tianhan Xu, Yixun Gu, Mantian Xue, Renjie Gu, Bin Li, and Xiang Gu. Knowledge graph construction for heart failure using large language models with prompt engineering. *Frontiers in Computational Neuroscience*, Volume 18 - 2024, 2024b. ISSN 1662-5188. doi: 10.3389/fncom.2024.1389475. URL https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2024.1389475.

Christophe Ye and Cassie S. Mitchell. LLM as entity disambiguator for biomedical entity-linking. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 301–312, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.25. URL https://aclanthology.org/2025.acl-short.25/.

Bowen Zhang and Harold Soh. Extract, define, canonicalize: An LLM-based framework for knowledge graph construction. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9820–9836, Miami, Florida, USA, nov 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.548. URL https://aclanthology.org/2024.emnlp-main.548/.

Chunju Zhang, Chenxi Li, Shu Wang, Zhao Zhang, Yunqiang Zhu, RuiXue Xiong, Gang Wang, Maohong Chen, RenYi Chen, and Yankai Yu. Ontogrc: ontology-guided mineral exploration knowledge extraction from geological texts using generate–reflect–correct prompting. *Ore Geology Reviews*, 195:107411, 2026. ISSN 0169-1368. doi: https://doi.org/10.1016/j.oregeorev.2026.107411. URL https://www.sciencedirect.com/science/article/pii/S0169136826003100.

A Ontology Class Hierarchy

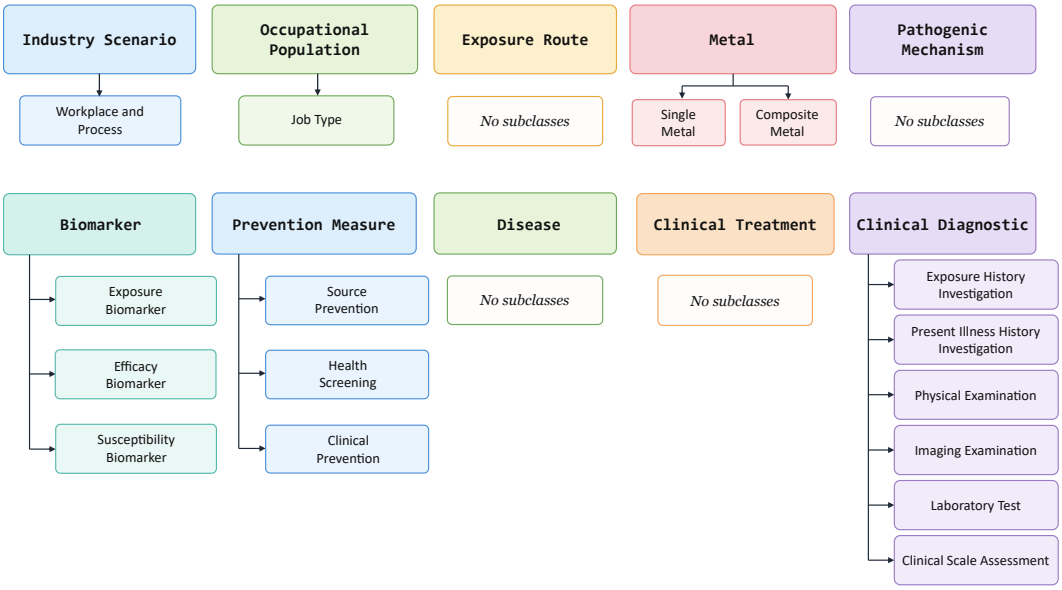

Figure 3: Class hierarchy of the MHKG ontology.

The ontology used in MHKG comprises 26 classes, including 10 top-level classes and 16 subclasses. Figure 3 shows the hierarchical organization of these classes. Several top-level classes are further specialized into domain-specific subclasses. For instance, `Metal` is divided into `Single_Metal` and `Composite_Metal`; `Biomarker` includes 3 types of subclasses; and `Clinical_Diagnostic` contains six subclasses corresponding to different diagnostic activities. Other top-level classes without further specialization remain as independent classes in this ontology.

B Data Properties

In addition to the class hierarchy described in Appendix A, the MHKG ontology specifies the relations permitted between entity classes and the data properties associated with individual entity types. This section summarizes these schema-level specifications. Object properties define directed semantic relations between classes and are constrained by predefined domains and ranges, whereas data properties describe the attributes that may be attached to entities of a given class.

Table 2: Object properties and domain–range constraints in the MHKG ontology.

ID	Subject	Predicate	Object
P1	Industry_Scenario	hasWorkplaceAndProcess	Workplace_and_Process
P2	Industry_Scenario	involvesPopulation	Occupational_Population
P3	Occupational_Population	hasJobType	Job_Type
P4	Occupational_Population	exposedThrough	Exposure_Route
P5	Exposure_Route	contactsMetal	Metal
P6	Metal	induces	Pathogenic_Mechanism
P7	Pathogenic_Mechanism	leadsToDisease	Disease
P8	Metal	affectsBiomarker	Biomarker
P9	Biomarker	assistsDiagnosisOf	Disease
P10	Biomarker	suggestsPreventionMeasure	Prevention_Measure
P11	Clinical_Diagnostic	detectsBiomarker	Biomarker
P12	Clinical_Diagnostic	diagnosesDisease	Disease
P13	Prevention_Measure	interruptsExposureRoute	Exposure_Route
P14	Prevention_Measure	preventsAndControlsDisease	Disease
P15	Clinical_Diagnostic	useOn	Clinical_Treatment
P16	Clinical_Treatment	intervene	Disease

B.1 Object Properties and Domain–Range Constraints

Our ontology defines 16 directed object properties. Each object property is associated with a predefined domain and range, which jointly constrain the admissible subject and object types of a relation. Relations that do not conform to the predefined predicate vocabulary or domain–range constraints are not included in the materialized graph. The complete set of object properties is shown in Table 2.

B.2 Entity Data Properties

Besides inter-entity relations, our ontology defines class-specific data properties for representing descriptive information associated with individual entities. Data properties are populated only when the corresponding information is explicitly supported by the source text. Table 3 lists the data properties permitted for each entity class.

C Knowledge Extraction Guidelines

C.1 Extraction Input

Knowledge extraction is performed at the document chunk level. Each input instance contains the title of the source paper, an internal chunk identifier, the identifier of the corresponding source chunk, and the textual passage used for extraction. The textual content of the chunk is treated as the source of biomedical knowledge. The paper title and chunk identifiers are retained exclusively for provenance tracking.

Table 3: Entity classes and associated data properties in the MHKG ontology.

Entity Class	Data Properties
Industry Scenario	
Industry_Scenario	Industry_Name, Industry_Category, Industry_Description
Workplace_and_Process	Workplace_Type, Process_Name
Occupational Population	
Occupational_Population	Population_Description, Age_Range
Job_Type	Job_Title, Job_Responsibilities
Exposure Route	
Exposure_Route	Route_Category
Metal	
Single_Metal	Metal_Name, Element_Symbol
Composite_Metal	Metal_Mixture
Pathogenic Mechanism	
Pathogenic_Mechanism	Mechanism_Pathway, Pathophysiology_Description
Biomarker	
Exposure_Biomarker	Exposure_Biomarker_Name, Exposure_Biomarker_Sample_Type, Reference_Range
Efficacy_Biomarker	Biomarker_Name, Relevant_Organ, Reference_Range
Susceptibility_Biomarker	Biomarker_Name, Relevant_Organ, Reference_Range
Prevention Measure	
Prevention_Measure	Prevention_Measure_Name
Source_Prevention	Prevention_Measure_Name, Measure_Type, Target_Population
Health_Screening	Prevention_Measure_Name, Screened_Population, Screening_Tests
Clinical_Prevention	Prevention_Measure_Name, Prevention_Protocol, Outcome_Metrics
Disease	
Disease	Disease_Name, Target_Organ/System, Clinical_Manifestations
Clinical Treatment	
Clinical_Treatment	Treatment_Protocol, Treatment_Metrics
Clinical Diagnostic	
Exposure_History_Investigation	Questionnaire_Items
Present_Illness_History_Investigation	Symptom_Description, Onset_Time, Severity_Level
Physical_Examination	Physical_Examination_Items
Imaging_Examination	Imaging_Examination_Type
Laboratory_Test	Sample_Type, Test_Items
Clinical_Scale_Assessment	Scale_Name, Measured_Domain, Score_Range

C.2 Ontology-Constrained Extraction

Entity and relation extraction are constrained by the predefined MHKG ontology. Extracted mentions are assigned to the most specific ontology class supported by the source text. Overly generic labels are avoided when more specific expressions are available. Synonymous and coreferential

mentions occurring within the same text chunk are merged when they refer to the same concept, and standard identifiers are preserved when they are explicitly reported in the source text.

Relation extraction is restricted to the predefined relation vocabulary (P1-P16). Relations or intermediate concepts are not introduced when they are not explicitly supported by the source text. For instance, if a passage directly associates a metal with a disease but does not describe a pathogenic mechanism, no additional `Pathogenic_Mechanism` entity is inferred. Each extracted entity and relation is associated with concise textual evidence from the source passage, and the evidence supporting a relation must be sufficient to identify both its subject and object.

To preserve uncertainty in the extraction output, relation-level confidence is represented by three categories: high, medium, and low. These categories correspond respectively to stated evidence, indirect or moderately supported evidence, and tentative or weakly supported evidence. Entity labels and attribute values retain the language of the source text, whereas relation predicates follow the fixed English ontology vocabulary.

C.3 Output Representation

The extraction results are stored in a structured JSON format that contains separate entity and relation records. Each entity record includes an entity identifier, display label, ontology class, class-specific data properties, supporting evidence, and provenance information. Relation records contain the subject identifier, predicate, object identifier, optional relation-level properties, supporting evidence, provenance metadata, and a confidence label. The provenance fields preserve the source paper and chunk identifiers, enabling every extracted item to be traced back to the text passage from which it was generated.

D Prompt for MHKG Construction

We compile the prompt used for extracting knowledge from the literature during the MHKG construction process. Some of the details are as follows:

Prompt

You are a biomedical knowledge graph construction expert specializing in metal exposure and health effects. Extract entities and relations from user-provided academic text, including abstracts, full texts, or passages. Strictly follow the ontology schema below and output only valid JSON. Do not use external knowledge or invent unsupported information. The user input is a

JSON object representing one text chunk from a JSONL file.
Input fields:
- paper_title: title of the source paper, derived from the PDF filename.
- chunk_id: sequential chunk id in the generated JSONL file.
- source_chunk_id: original chunk id from the PDF partitioning process.
- text: the actual source text to extract from.

Only extract biomedical entities and relations from the ‘text‘ field.
Do not treat ‘paper_title‘, ‘chunk_id‘, or ‘source_chunk_id‘ as biomedical content.
Copy ‘paper_title‘, ‘chunk_id‘, and ‘source_chunk_id‘ exactly into the output provenance fields.

Extraction Rules

1. Extract all entities matching the ontology. Assign each entity to the most specific possible class.
2. Do not extract overly generic terms such as `workers`, `disease`, `toxicity`, `screening`, or `treatment` if a more specific expression is available.
3. `Industry_Scenario`, `Workplace_and_Process`, `Occupational_Population`, and `Job_Type` are human-only occupational or industrial classes. Do not use them for animal

experiments, animal models, cell lines, tissue experiments, laboratory protocols, or non-human subjects such as mice, rats, zebrafish, HepG2 cells, or animal exposure groups.
4. Merge synonymous or coreferential mentions, such as `lead`, `Pb`, and `plumbum`.
5. Preserve disease names from the source text. If ICD-10, MeSH, or other standard terms are explicitly given, store them in properties.
6. Extract data properties only when stated in the text. Do not infer or supplement missing information.
7. Do not create a `Pathogenic_Mechanism` entity merely to connect a metal directly associated with a disease.
8. Record concise but sufficient *evidence_text* for every entity and relation. For relations, evidence must mention or clearly support both subject and object.
9. Entity labels and property values should use the same language as the source text. Predicates must remain in English.
10. For long texts, process paragraph by paragraph, merge duplicates, and extract all relevant metals, diseases, populations, routes, biomarkers, diagnostics, treatments, and prevention measures.
11. Preserve chunk provenance for every output. Each entity and relation must include a *source* object copied exactly from the input:
- source.paper_title = input.paper_title
- source.chunk_id = input.chunk_id
- source.source_chunk_id = input.source_chunk_id

Output Format

Output only JSON. Do not output explanations, Markdown, or text outside the JSON. Omit fields with no corresponding information.

```
{
  "entities": [
    {
      "id": "unique_entity_identifier",
      "label": "entity_display_name",
      "class": "class_name_or_most_specific_subclass",
      "properties": {
        "property_name": "property_value_string_or_string_array"
      },
      "evidence_text": "source text span",
      "source": {
        "paper_title": "input paper_title",
        "chunk_id": number,
        "source_chunk_id": number
      }
    }
  ],
  "relations": [
    {
      "subject": "subject_entity_id",
      "predicate": "relation_name_corresponding_to_P1-P16",
      "object": "object_entity_id",
      "evidence_text": "source text span",
      "source": {
        "paper_title": "input paper_title",
        "chunk_id": number,
        "source_chunk_id": number
      },
      "confidence": "high/medium/low"
    }
  ]
}
```