

Constructing a Knowledge Graph of Historical Mining Data

Basel Shbita*, Namrata Sharma, Binh Vu, Fandel Lin and Craig A. Knoblock

Information Sciences Institute, University of Southern California, Marina del Rey, California, USA

Abstract

The interpretation and analysis of historical mining data pose significant challenges due to its heterogeneous and scattered nature across various archival sources. This data is crucial for identifying new sources of critical minerals, understanding past resource utilization, and aiding in future project development, necessitating spatial, temporal, and semantic integration for comprehensive domain insights. In this paper, we detail our methodology for constructing, modeling, and semantically enriching a knowledge graph (KG) centered on historical mining data. Leveraging a custom ontology and semantic web technologies, we transform digitized archival records into a temporally and spatially aware, semantically rich KG. The resulting KG facilitates advanced temporal and spatial analyses through SPARQL queries, and enhances semantic richness by linking to additional data on the web. We demonstrate the application of our KG in the nuanced analysis of historical mining data and the generation of grade and tonnage models for two critical minerals: nickel and zinc. Our evaluation highlights the KG's effectiveness in spatial and temporal interpretation of mining data, underscores the strengths of our entity linking method with an open knowledge base, and details the performance analysis of query execution. We also make the resulting KG available as open linked data.

Keywords

knowledge graphs, geospatial linked data, data semantics, mineral mining data, semantic web

1. Introduction

Understanding historical mining data is a pursuit of geoscience research and a necessity for informed decision-making in resource management and environmental conservation. The ability to accurately analyze and interpret this data helps identify new sources of critical minerals and illuminates past resource utilization activities. With the increasing demand for mineral resources [1, 2], there is a growing need to draw upon historical mining data to make informed decisions about current and future mining projects.

Historical mining data is often heterogeneous in nature, existing in varied forms and scattered across numerous archival sources. In many cases, Subject Matter Experts (SMEs) and organizations such as the United States Geological Survey (USGS) are pivotal in organizing these data. They bring indispensable knowledge to critical tasks, such as mineral assessments [3]. However, historical mining data is scattered across multiple sources - ranging from quantitative

6th International Workshop on Geospatial Linked Data, co-located with the 21st Extended Semantic Web Conference (ESWC 2024), May 26–30, 2024, Hersonissos, Greece

*Corresponding author.

✉ shbita@isi.edu (B. Shbita); nsharma4@isi.edu (N. Sharma); binhvu@isi.edu (B. Vu); fandelli@isi.edu (F. Lin); knoblock@isi.edu (C. A. Knoblock)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

ore details in mine reports to spatial layers in existing databases - lacks structured organization, and suffers from issues such as quality, accuracy, and completeness [4, 5].

Knowledge Graphs (KG) are a popular way to represent information in a way that can be easily interpreted by both humans and machines. Emphasizing the value of transforming historical records into structured, queryable formats, KGs offer an effective solution for such transformation, combining expressivity, interoperability, and standardization in the semantic web stack, thus providing a strong foundation for querying and analysis.

The evolution and significance of the integration of geospatial data on the web and its extension to linked data has been extensively discussed by Janowicz et al. [6]. Recent technological advances [7] have greatly facilitated the integration of geospatial data from historical archives and maps, transforming these diverse datasets into structured, coherent knowledge bases [8, 9, 10, 11, 12]. A similar transformation is useful in addressing the challenges of mining data analysis, as it enables the consolidation of disparate data sources into a single, accessible, and simple linked data representation - a KG that can be materialized into triples, i.e., Resource Description Framework (RDF) data.

To address this task, this paper presents a methodology for the construction, modeling, and augmentation of a KG dedicated to historical mining data. Our approach involves creating a KG from diverse data sources, formulating a custom domain ontology (semantic model) to represent the data, and utilizing open knowledge from the web to enrich and contextualize data about mineral commodities. Moreover, by integrating additional attributes from geospatial databases, such as MRDS [13], we support spatial queries and visualizations pertaining to specific geo-locations, allowing the development of dedicated downstream applications.

The resulting KG unlocks significant value in generating grade and tonnage plots for mineral resources and commodities. These plots are pivotal, illustrating the relationship between the grade (mineral content) of a deposit and the available tonnage (quantity of ore), which are crucial for assessing the economic viability of mining projects and for accurate reporting of mineral resources and reserves. Figure 1 showcases a grade-tonnage model derived from our KG, highlighting the technology's ability to support sophisticated analyses and facilitate the extraction of domain insights with ease.

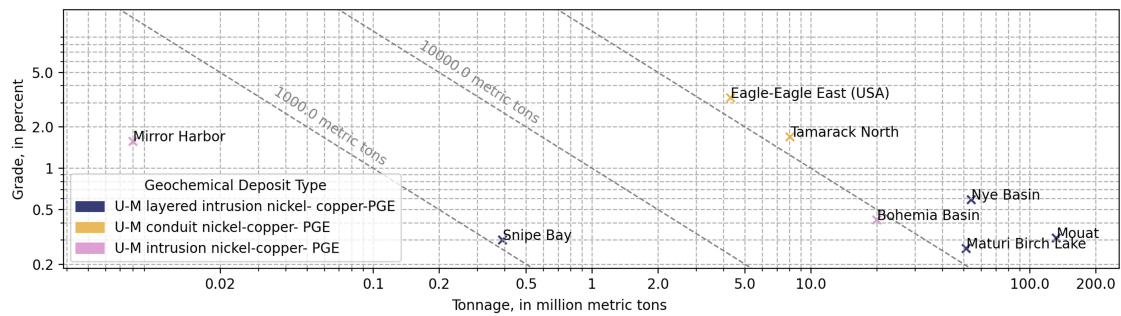


Figure 1: Grade-tonnage model of nickel mineral deposits built from a KG query (SPARQL) response, categorized by their Critical Minerals Mapping Initiative (CMMI) deposit classification. Specific sites are marked to illustrate the variability in grade and tonnage among these deposit types.

Through a well-crafted SPARQL [14] query, grade and tonnage data for various minerals can be efficiently retrieved from the KG. SPARQL is a powerful RDF query language and protocol designed for querying and manipulating RDF data (triples) in the KG, enabling users to precisely extract and analyze information based on specific criteria and the most current data. Furthermore, the KG's flexible structure allows users to tailor queries to specific minerals, geographic areas, or time frames, showcasing its adaptability and comprehensive support for the economic assessments of mining projects.

We evaluate the application of our KG through a detailed analysis of historical mining sites, focusing on two critical minerals: zinc and nickel. Our evaluation demonstrates the KG's ability to generate grade and tonnage plots using SPARQL queries to visualize this data. It highlights the flexibility and robustness of our system in handling complex queries across multiple dimensions. For example, by aggregating mineral sites by their past resource classifications, geochemical deposit classification, or by restricting it to specific geographic regions. Additionally, we conduct a rigorous evaluation of our entity linking method, demonstrating its effectiveness in accurately matching commodities to their corresponding entities in an external knowledge base.

We list the contributions of this paper as follows:

1. We present a pipeline for the integration of extracted quantitative, spatial, and semantic information from historical mining data archives, resulting in an integrated KG.
2. We introduce a method to identify and retrieve instances of a given type from a publicly available KG, specifically entity matching commodities with open linked data.
3. We assess the applicability of the resulting KG by designing queries to automatically generate grade-tonnage models for two critical minerals: zinc and nickel.
4. We make the resulting KG publicly available in the form of linked data (queryable RDF via a SPARQL endpoint).¹

The rest of the paper is organized as follows: Section 2 outlines our comprehensive methodology for constructing, linking and augmenting the KG, detailing our entity linking process and the integration of diverse data sources to populate the KG. Section 3 presents our evaluation framework, employing real-world datasets focused on two critical minerals. Section 4 describes related work. Finally, Section 5 concludes the paper by summarizing our findings and outlining future work.

2. Constructing the Knowledge Graph

This section describes the methodological framework for constructing the KG for historical mining data. Our methodology is characterized by the formulation of a unique semantic model (Section 2.1), linking with external linked data on the web (Section 2.2), and finally the materialization of the data into a KG (Section 2.3). Each component is meticulously planned to ensure the integration of semantic relationships, spatial data, and temporal dimensions, thereby facilitating a robust analysis of historical and contemporary mining data.

¹<https://minmod.isi.edu/sparql>

2.1. Defining the Semantic Model

Central to the semantic enrichment and structural integrity of our KG is developing a custom ontology and semantic model tailored to the unique characteristics and relationships inherent to historical mining data. This model is designed to capture the domain-specific attributes and complex relationships essential for accurately representing mining information about mineral commodities, such as mineral sites and inventories.

Utilizing RDF as a foundational framework, our approach leverages its structural flexibility and suitability for representing diverse metadata forms, including using existing standards that adhere to universally accepted conventions. This is particularly beneficial for supporting spatial queries, a capability enhanced by integrating the OGC GeoSPARQL standard [15]. The use of this standard enriches our model by providing a vocabulary for representing geospatial data on the web, facilitating qualitative spatial reasoning and quantitative spatial computations.

As illustrated in Figure 2, the semantic model delineates the primary entities and their relationships. The model identifies `:MineralSite` entities, characterized by spatial information (`:LocationInfo`) encoded using GeoSPARQL namespace and Well-Known Text (WKT) notation for describing geometries. Mineral deposits, representing natural occurrences of minerals, are linked to mineral sites. Each site may contain multiple `:MineralInventory` items, representing the quantity (`:Ore`) and grade (`:Grade`) of commodities (`:Commodity`) present. The model allows for aggregating these commodities by their `:ResourceCategory` (enumeration from a predefined list, e.g., indicated, measured, inferred), and the association of each site with a specific `:DepositType`, that is also selected from a predefined list, according to the CMMI standards [16].

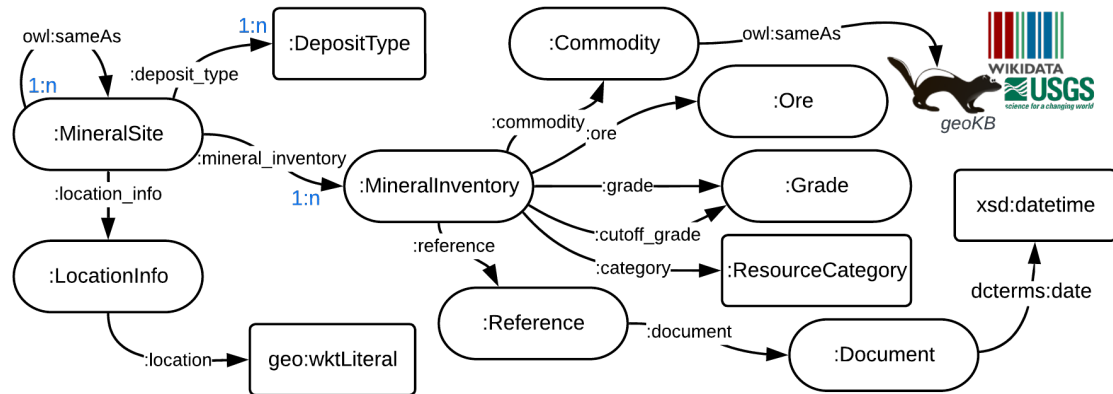


Figure 2: Semantic model of the mining data structure. Cardinalities (in blue) show one-to-many node relationships. Circular nodes represent instances, rectangular nodes represent literals (or enumerations). The “:” denotes our namespace.

Including missing information or expert inputs on deposit types underscores the KG’s ability to facilitate data classification and aggregation by spatial or attribute data. Each inventory item is associated with a `:Reference` to `:Document` provenance, ensuring data veracity and enabling effective interaction with the KG. `owl:sameAs` enhances the integration with external sources (e.g., geoKB) and within our KG (`:MineralSite` instances). Furthermore, inventory items are

tagged with `dc:terms:date` properties to indicate temporal aspects, adhering to the Dublin Core Metadata Initiative and W3C recommendations, enhancing the model’s comprehensiveness.

2.2. Entity Linking with geoKB

Our approach to enhancing the KG with links going to additional sources involves a simple entity linking process with geoKB², the Geoscience Knowledge Base developed by the USGS. This process is pivotal in enriching our KG with validated and scientifically relevant data, leveraging the extensive earth systems science portfolio within geoKB. Figure 3 demonstrates how the nickel mineral is depicted in geoKB, highlighting the depth of linked data that enriches our KG with details on mineral species and their historical classifications.

GEOKB Nickel (Q162561)

mineral species in the Iron Group sourced from Mindat and the Geoscience Ontology

Niccolum | nikle | Nickel | نيكيل | Nikelo | ニッケル | Nikal | Nichel | Nikiel | Nikl | IMA1966-039 | 니켈 | Niķelis | Nikkeli | نيكيل | Niqel | Nikelj | निकेल | ნიკელი | Никел | Nikèl | Никѣль | Nickel | Iztāctepoztlī | نيكيل | نيكيل | Niqel | Никл | نيكيل | Nickyl | نيكيل | Никель | Nikel | Nikil | Nikelis | Νικέλιο | Néckel | Никель | Kopukōreko | Niken | Nichele | 自然镍 | Nikkel

Statements

- subclass of: 01.AA.05 - Copper-cupalite family (2 references), mineral material (1 reference)
- has chemical element: nickel (1 reference)
- member of: Iron Group (1 reference)
- same as: <http://www.mindat.org/min-2895.html> (0 references), <https://w3id.org/gso/mineral/nickel> (0 references)

Figure 3: Illustration of the nickel mineral species in geoKB, showcasing the depth of information our entity linking method accesses, enriching our KG with metadata from external sources.

At the heart of our methodology is a constrained search within geoKB, using SPARQL queries designed to find commodities classified under certain instance types. This approach narrows the scope of potential matches to enhance their relevance. Listing 1 presents a query aimed at fetching candidate mineral commodity instances of nickel from geoKB, marking the first step in our entity linking process. The query filters for entities that are instances of (P1) mineral commodities (Q406) as seen in line 3 (where `gkbt` and `gkbi` represent namespaces for predicates and instances within geoKB, respectively). The `FILTER` clause (line 4) conducts a case-insensitive search to align the entity labels with the commodity string in question, exemplifying our strategy for extracting semantically related instances.

Following retrieval, we apply the Jaccard [17] similarity measure for set-based comparison between the commodity strings and geoKB entity labels. This process, based on the intersection over the union of the derived sets, helps determine the top instance among the candidates. The selected instance is then leveraged to enrich our KG, infusing it with additional semantics

²<https://geokb.wikibase.cloud/>

```
1 SELECT ?entity ?entityLabel WHERE {  
2   ?entity rdfs:label ?entityLabel.  
3   ?entity gkbt:P1 gkbi:Q406. # instance of mineral commodity  
4   FILTER(CONTAINS(LCASE(?entityLabel), "nickel")) }
```

Listing 1: A SPARQL query example targeting the nickel mineral commodity in geoKB serving as a foundational step for entity linking.

and metadata from sources like Wikidata [18], the Geoscience Ontology [19], and geoKB itself. Consequently, our KG is augmented with comprehensive information on mineral species and historical mineral classifications, among other data, thereby enhancing its semantic richness and inter-connectedness.

2.3. Transforming the Data into Triples

Our methodology for populating the KG leverages data sourced from semi-structured (tabular format) and structured (JSON files) data related to mining activities and mineral commodities. The initial step in the KG construction involves meticulous data cleaning and normalization, including entity deduplication and URI (Uniform Resource Identifier) mapping to ensure each entity, such as a mineral site, inventory, or document, is uniquely identifiable and accessible on the web.

URIs are generated using a hash function (e.g., MD5) to create a de-referenceable URI from the unique combination of the defining attributes of an instances, including feature type, location data, and temporal information. For instance, a `:MineralInventory` instance's URI is constructed using a concatenation of its commodity URI, category, the referring mineral site URI, and any associated document URIs. A similar approach is taken for `:MineralSite` and `:Document` URIs, using their respective source identifiers and bibliographic data.

In our KG, entities such as commodities and deposit types are linked to external knowledge bases or predefined lists, like the CMMI, while other entities are represented as blank nodes within our namespace. The transformation of data into RDF triples is facilitated by automated tools like D-REPR [20] and SAND [21], ensuring seamless integration and update of information.

Geospatial data are materialized using dedicated namespaces, notably GeoSPARQL, which augments the KG's capacity for spatial analysis. Additionally, the ontology supports OWL [22] for representing complex inter-entity relationships and attributes.

A validation layer, crucial for maintaining data integrity and consistency, is implemented through an automated system using SHACL [23]. This validation ensures that our KG not only accurately represents the data but also adheres to the predefined semantic model (Section 2.1), enabling reliable and sophisticated queries.

3. Evaluation and Discussion

Our evaluation framework includes qualitative and quantitative analyses over a KG built from a dataset covering two mineral commodities, focusing on the KG's adherence to the semantic model and completeness (Section 3.1), entity linking with geoKB (Section 3.2), and its utility

and performance in advanced data analysis (Section 3.3), particularly in generating grade and tonnage models.

The extensive dataset we use covers over 50 NI 43-101 technical reports (International Strategic Mineral Inventory reports) on nickel (2001 to 2021) and zinc (2002 to 2019), supplemented by spatial data from the MRDS (Mineral Resources Data System) and USMIN (US Mineral Deposit Database) databases. Sources such as Mudd and Jowitt’s compiled work on zinc from 2017 [24], and on nickel from 2022 [25], provide additional data and extensive coverage on various commodities as well, offering a rich blend of geospatial, geological, and economic data from various global locations.

3.1. Evaluation on the Semantic Model

Our evaluation confirms the KG’s adherence to the semantic model, reflecting accurate domain representation in compliance with RDF standards for enhanced query performance and data interoperability. Our resulting KG characteristics are described in Table 1.

The resulting KG hosts a significant number of instances and blank nodes, suggesting a rich network of connected data. Diving deeper, the KG encapsulates vital entities including `:MineralSite`, `:MineralInventory`, `:Commodity`, `:DepositType`, and `:Reference`. The entities form a network that reflects the complex interactions and interplay between various facets of mining data. For instance, `:MineralSite` entities are geospatially positioned through `:LocationInfo` relationships, while each site can encompass multiple `:MineralInventory` records, detailing the reserves in terms of quantity and grade for each `:Commodity`. Moreover, the KG adheres to CMMI standards for deposit classification and manages data provenance through `:Reference` links to detailed source citations, as expected.

Table 1

Historical mining data knowledge graph characteristics.

Characteristic	Count
Total Triples	2,397,708
Distinct Classes	16
Instances (Non-literals)	226,267
Geospatial Instances	2,884
Blank Nodes	1,518,981

The granular representation of the data in the resulting graph, with 1,112 zinc and 1,132 nickel reserve and resource measurements, alongside 3,809 zinc and 2,021 nickel mineral site instances, demonstrates the KG’s quantitative depth and utility. This structured model captures key attributes, ensuring interoperability and alignment with semantic data representation best practices. The approach we present is complete and follows linked data principles by:

- Generating URIs as names for things, without modifying previously published identifiers
- Maintaining existing relations (predicates) between instances (“backward compatibility”)
- Generating machine-readable structured data
- Using standard namespaces and semantics (e.g., OWL, Dublin Core, GeoSPARQL)
- Linking to additional resources on the web (e.g., geoKB)

3.2. Evaluation on Entity Linking

To evaluate the effectiveness of our proposed entity-linking method, we conducted an evaluation using a dataset of 135 extracted commodity labels mapped by human experts to geoKB. This dataset serves as a benchmark for determining the success of our entity linking method.

To provide a comprehensive evaluation, we contrasted our approach against three distinct baseline methods within geoKB. The first two baselines involve a generalized, string search strategy with SPARQL, then label comparison – one utilizing the Jaro [26] string similarity measure and the other employing the Jaccard measure – based purely on textual relevance. The third baseline adopts a constrained search strategy (instance based) combined with the Jaro similarity measure, similar in part to our proposed method, refining the search scope yet still leveraging Jaro’s well-regarded efficiency in measuring string similarities for short texts, such as names.

This multifaceted comparison shows substantial gains in matching accuracy with our proposed method – instance-based constrained search followed by Jaccard similarity measure for set-based comparisons – over the baselines. Specifically, our method has demonstrated significant performance enhancements, showcasing the advantage of a constrained search strategy combined with the precision of Jaccard similarity. This approach not only refines the selection of potential matches but also ensures a high degree of textual similarity between the commodity strings and the linked geoKB entities. The results are summarized in Table 2.

Table 2

Evaluation results for the entity linking experiments with geoKB.

Method	MRR	Hits@1	Hits@3	Hits@5
String search, then Jaro	0.557	0.459	0.659	0.659
String search, then Jaccard	0.648	0.637	0.659	0.659
Instance search, then Jaro	0.801	0.689	0.926	0.956
Instance search, then Jaccard (proposed)	0.940	0.904	0.978	0.978

The results underscore the superior performance of our proposed entity linking approach, with a Mean Reciprocal Rank (MRR) of 0.940 and impressive Hits@1, Hits@3, and Hits@5 rates of 0.904, 0.978, and 0.978 again, respectively. These metrics notably surpass those of the baseline methods, thereby affirming the utility of combining constrained search with the Jaccard similarity measure. The MRR value, being substantially higher than that of the baselines, indicates that our method consistently identifies the most relevant geoKB entity at the top rank. The high Hits@1 value signifies that the correct entity is identified as the top match in a significant majority of cases, a critical metric for applications relying on precision. Similarly, the near-perfect Hits@3 and Hits@5 scores suggest that if the top match isn’t the exact entity, it is highly likely to be within the top 3 or 5 candidates, offering a valuable safety net for ensuring data quality in the KG. These results collectively justify our method’s design, which meticulously tailors the search and comparison phases to optimize for both accuracy and relevance in entity linking.

3.3. Evaluation on Querying the KG

To assess the query performance, utility, and effectiveness of our KG in extracting relevant information for creating grade-tonnage models, we executed a series of SPARQL queries. These queries, aimed at testing the KG's ability to retrieve grade and tonnage data under diverse constraints and scenarios, were conducted using RDF triples hosted on Apache Jena³, a lightweight and programmable environment with geospatial query support. The baseline query, shown in Listing 2, is crucial for fetching grade and tonnage data along with site and inventory identifiers, and serves as the foundation for developing more complex queries. By building on this foundational query, we introduced three distinct constraint types in our subsequent queries – textual, temporal, and spatial – thereby not only testing the KG's flexibility in meeting varied query requirements but also showcasing its capability to provide precise and contextually relevant data across different analytical dimensions.

```
1 SELECT ?ms ?mi ?ms_name ?mi_cat ?ore ?grade
2 WHERE {
3   ?ms :mineral_inventory ?mi .
4   OPTIONAL { ?ms rdfs:label|:name ?ms_name . }
5   ?mi :category ?mi_cat .
6   ?mi :ore [ :ore_value ?ore;
7             :ore_unit ?ore_unit] .
8   ?mi :grade [ :grade_value ?grade;
9               :grade_unit ?grade_unit] . }
```

Listing 2: Baseline SPARQL query for grade and tonnage data.

In the first type of query we retrieve ore and tonnage data for a specific commodity. This query aims to retrieve ore grade and tonnage data for all inventories associated with a specified commodity. It demonstrates the KG's capability to filter data based on commodity type, which is essential for users interested in specific mineral insights. Listing 3 shows the added clause needed to retrieve entries for a given mineral commodity name, nickel in the example.

```
1 ?mi :commodity/:name "nickel"@en .
```

Listing 3: SPARQL clause for filtering by commodity type: this clause filters inventory items to retrieve data specific to the nickel commodity, demonstrating how to tailor queries for particular mineral.

In the second type of query we retrieve ore and tonnage data with an emphasis on a temporal constraint on document provenance, from which the data originated. This query filters ore and tonnage data based on the publication date of the source documents. Such a query is useful for researchers interested in how grade and tonnage estimates could have changed over time or analyze specific data within a specific timeframe. Listing 4 shows the added clause needed to retrieve inventory items pertaining to specific time ranges. In this case we are fetching inventories from documents published between the year 2000 to 2010.

³<https://jena.apache.org/>

```

1 ?mi :reference/:document [ dcterms:date ?date ] .
2 FILTER(?date >= "2000"^^xsd:gYear && ?date <= "2010"^^xsd:gYear) .

```

Listing 4: SPARQL clause for temporal filtering: this clause applies a temporal filter to select inventory items based on their document’s publication year between 2000 and 2010, showcasing the KG’s ability to analyze historical data over a specific time range.

Utilizing Apache Jena’s support for GeoSPARQL, the third query retrieves tonnage data from inventories at mineral sites within a certain distance from a given point. It exemplifies the KG’s spatial querying capabilities, which are crucial for geographical analyses and decision-making. Listing 5 shows the added clause needed to retrieve mineral sites, with inventory items, that are within a specific distance from a given point data in WKT format. In this example we are searching for mines that are within 500 miles from given coordinates.

```

1 ?ms :location_info/:location ?loc_wkt .
2 FILTER(geof:distance(?loc_wkt, "POINT(-118.57 47.56)"^^geo:wktLiteral, unit:mile) < 500)

```

Listing 5: SPARQL clause for spatial proximity filtering: this clause leverages GeoSPARQL to find mineral sites within a 500-mile radius of a specified point, exemplifying spatial querying capabilities for geographical analysis. The geof and unit namespaces are standard namespaces utilized for specifying distance measurements and units, respectively.

Table 3 presents a summary of the query-time performance, including average, minimum, and maximum times, effectively showcasing the efficiency of our KG when operating under various query constraints. This efficiency is underscored by the execution of hundreds of similar queries across a diverse range of values for each constrained scenario, further demonstrating the robustness and adaptability of our system in handling retrieval tasks.

Table 3
Query time statistics (in milliseconds)

Query Constraint Type	Avg	Min	Max
Textual	450	369	649
Temporal/Numeric	438	388	607
Spatial	708	501	811

The results outlined in Table 3 showcase the KG’s performance across different query constraints, with query times measured in milliseconds. The average query time for textual constraints was notably efficient at 450 ms, reflecting the rapid response to straightforward textual searches. Temporal queries, with an average time of 438 ms, highlight the KG’s adept handling of quantitative and temporal data retrieval, facilitating temporal analysis. Spatial queries, while more computationally intensive due to the nature of geospatial data processing, still performed admirably with an average time of 708 ms. This demonstrates the system’s capacity to efficiently manage spatial reasoning tasks, a crucial aspect for mining data analysis where geographical context is vital.

These results are not only indicative of the KG’s robust performance but also validate our methodological choices and architecture. The swift response times, especially for spatial queries, are a testament to the efficiency of integrating GeoSPARQL and our custom semantic model, facilitating advanced spatial analyses. Furthermore, the accurate retrieval of information across all query types confirms the KG’s utility in supporting complex queries for critical tasks such as generating grade-tonnage models, as we demonstrate in Figure 1.

The application of SPARQL queries against our KG exemplifies the invaluable insights gained from the fusion of semantic web technology with spatial visualization techniques, enabling the straightforward interpretation of otherwise complex geographic data. For example, Figure 4 presents a detailed visualization of nickel mineral sites across the United States, categorized according to CMMI standards and overlaid on a topographic map, demonstrating the expansive coverage of our KG. By integrating this classification with other geospatial data, such as geological formations and stratigraphy information, we can significantly enhance the multi-dimensional analytical capabilities available to SMEs, allowing for predictive modeling of mineral potential and helping to identify unexplored areas with high resource prospects.

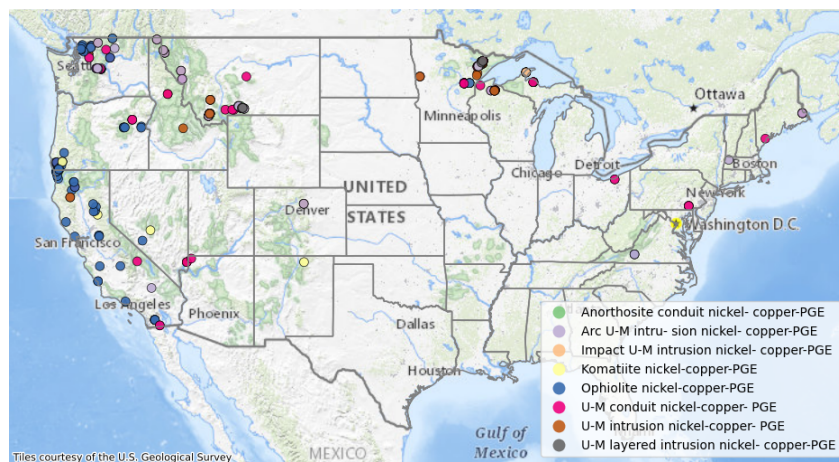


Figure 4: Nickel mineral sites in the US by CMMI classification on a topographic map background. Example of a spatial visualization based on data derived from the KG and retrieved via SPARQL, showcasing nickel mine distribution.

By structuring historical and current mining data within a KG, we enable powerful query capabilities through SPARQL, facilitating the retrieval and representation of complex data sets easily and quickly. The query results above establish high confidence in our model, showing that we can easily and effectively answer complex queries in a robust manner. Furthermore, the integration of our commodity data with geoKB enhances our KG’s utility by enabling federated SPARQL queries, which allow us to fetch additional data from external sources such as Wikidata. This capability significantly broadens the scope of our analysis, providing access to a wealth of information that complements our existing datasets. Overall, we demonstrated that our approach and the proposed pipeline can be effectively used to automatically construct effective and contextualized open KGs and linked data from historical and contemporary mining data, as well as support both temporal and spatial analysis.

4. Related Work

Recent advancements in geology and earth science data analysis have been significantly propelled by the application of machine learning techniques, which have enabled the enhancement of data mining and extraction for geology and mineral data [27]. These developments have shown considerable promise in various applications, ranging from prospectivity mapping to knowledge organization in the natural sciences [28, 29]. However, the full utilization of semantic and spatial relationships in historical mining data remain largely underexplored, indicating a gap in the current research landscape.

In the domain of geoscientific knowledge graphs, our work complements existing knowledge bases such as GeoKB and the Geoscience Ontology [19], by addressing the nuanced gaps in the semantic enrichment and spatial analysis of historical mining data, areas often overlooked in the broader context of such applications. This gap presents a unique opportunity to contribute to the field by leveraging semantic web technologies with spatial and temporal data analysis to enrich our understanding of historical mining activities and their implications for contemporary and future mining endeavors.

5. Conclusion and Future Work

In this paper, we introduce a comprehensive approach for constructing, modeling, and enriching a Knowledge Graph (KG) that captures the spatial and temporal dynamics, along with the complex semantic relationships, within historical mining data. Our approach enriches our understanding of historical mining operations and resource utilization and provides invaluable insights for academic research and practical applications in the mining industry.

Looking ahead, the continued development of our KG opens several promising avenues for further exploration. Plans are underway to integrate a broader range of data covering additional critical minerals, which promises to significantly enhance and expand the analytical capabilities of our KG alongside the inclusion of more diverse historical datasets. Furthermore, integrating advanced machine learning algorithms with our KG can reveal novel insights from historical data, revolutionizing the mining sector's strategic planning and operational efficiencies. Further enhancements could also explore the improvement of the semantic enrichment process to enhance the accuracy and relevance of the extracted and linked information, providing even more detailed insights into the complex history of mining data. Additionally, applying our methodologies to other historical contexts, such as assessing environmental impacts, represents a promising direction for extending the impact of our work beyond the mining domain.

Acknowledgments

This material is based upon works supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112390132 and Contract No. 140D0423C0093. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA); or its Contracting Agent, the U.S. Department of the Interior, Interior Business Center, Acquisition Services Directorate, Division V.

We thank Dr. Graham W. Lederer (United States Geological Survey) and Dr. Simon M. Jowitt (University of Nevada Reno, Nevada), who provided insights and expertise in the geology that greatly assisted the research.

References

- [1] K. J. Schulz, Critical mineral resources of the United States: economic and environmental geology and prospects for future supply, Geological Survey, 2017.
- [2] S. M. Fortier, N. T. Nassar, G. W. Lederer, J. Brainard, J. Gambogi, E. A. McCullough, Draft critical mineral list—Summary of methodology and background information—US Geological Survey technical input document in response to Secretarial Order No. 3359, Technical Report, US Geological Survey, 2018.
- [3] C. J. Green, G. W. Lederer, H. L. Parks, M. L. Zientek, Grade and tonnage model for tungsten skarn deposits—2020 update, Technical Report, US Geological Survey, 2020.
- [4] W. C. Day, The Earth Mapping Resources Initiative (Earth MRI): Mapping the Nation’s critical mineral resources, Technical Report, US Geological Survey, 2019.
- [5] A. H. Hofstra, V. Lisitsin, L. Corriveau, S. Paradis, J. Peter, K. Lauzière, C. Lawley, M. Gadd, J.-L. Pilote, I. Honsberger, et al., Deposit classification scheme for the Critical Minerals Mapping Initiative Global Geochemical Database, Technical Report, US Geological Survey, 2021.
- [6] K. Janowicz, S. Scheider, T. Pehle, G. Hart, Geospatial semantics and linked spatiotemporal data—past, present, and future, *Semantic Web 3 (2012)* 321–332.
- [7] Y.-Y. Chiang, S. Leyk, C. A. Knoblock, A survey of digital map processing techniques, *ACM Computing Surveys (CSUR)* 47 (2014) 1–44. doi:10.1145/2557423.
- [8] M. Alirezaie, M. Långkvist, M. Sioutis, A. Loutfi, Semantic referee: a neural-symbolic framework for enhancing geospatial semantic segmentation, *Semantic Web 10 (2019)* 863–880. doi:10.3233/SW-190362.
- [9] Z. Li, Y.-Y. Chiang, S. Tavakkol, B. Shbita, J. H. Uhl, S. Leyk, C. A. Knoblock, An automatic approach for generating rich, linked geo-metadata from historical map images, Association for Computing Machinery, New York, NY, USA, 2020, pp. 3290–3298. doi:10.1145/3394486.3403381.
- [10] J. H. Uhl, S. Leyk, Z. Li, W. Duan, B. Shbita, Y.-Y. Chiang, C. A. Knoblock, Combining remote-sensing-derived data and historical maps for long-term back-casting of urban extents, *Remote Sensing* 13 (2021) 3672. doi:10.3390/rs13183672.
- [11] B. Shbita, C. A. Knoblock, W. Duan, Y.-Y. Chiang, J. H. Uhl, S. Leyk, Building spatio-temporal knowledge graphs from vectorized topographic historical maps, *Semantic Web 14 (2023)* 527–549. doi:10.3233/SW-222918.
- [12] Y.-Y. Chiang, M. Chen, W. Duan, J. Kim, C. A. Knoblock, S. Leyk, Z. Li, Y. Lin, M. Namgung, B. Shbita, et al., GeoAI for the digitization of historical maps, in: *Handbook of Geospatial Artificial Intelligence*, CRC Press, 2023, pp. 217–247.
- [13] E. McFaul, G. Mason, W. Ferguson, B. Lipin, US Geological Survey mineral databases; MRDS and MAS/MILS, Technical Report, US Geological Survey, 2000.
- [14] W. W. W. Consortium, et al., SPARQL 1.1 overview, Technical Report, World Wide Web Consortium, 2013.
- [15] N. J. Car, T. Homburg, GeoSPARQL 1.1: Motivations, details and applications of the decadal update to the most important geospatial lod standard, *ISPRS International Journal of Geo-Information* 11 (2022) 117.
- [16] K. D. Kelley, D. L. Huston, J. M. Peter, Toward an effective global green economy: The

- critical minerals mapping initiative (cmmi), SGA News 8 (2021) 1–5.
- [17] P. Jaccard, The distribution of the flora in the alpine zone. 1, *New phytologist* 11 (1912) 37–50.
 - [18] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, *Communications of the ACM* 57 (2014) 78–85.
 - [19] B. Brodaric, S. M. Richard, The geoscience ontology, in: *AGU Fall Meeting Abstracts*, volume 2020, 2020, pp. IN030–07.
 - [20] B. Vu, J. Pujara, C. A. Knoblock, D-REPR: a language for describing and mapping diversely-structured data sources to rdf, in: *Proceedings of the 10th International Conference on Knowledge Capture*, 2019, pp. 189–196.
 - [21] B. Vu, C. A. Knoblock, SAND: A tool for creating semantic descriptions of tabular sources, in: *European Semantic Web Conference*, Springer, 2022, pp. 63–67.
 - [22] D. L. McGuinness, F. Van Harmelen, et al., OWL web ontology language overview, *W3C recommendation* 10 (2004) 2004.
 - [23] W. W. W. Consortium, et al., Shapes constraint language (SHACL), Technical Report, World Wide Web Consortium, 2017.
 - [24] G. M. Mudd, S. M. Jowitt, T. T. Werner, The world’s lead-zinc mineral resources: scarcity, data, issues and opportunities, *Ore Geology Reviews* 80 (2017) 1160–1190.
 - [25] G. M. Mudd, S. M. Jowitt, The new century for nickel resources, reserves, and mining: Reassessing the sustainability of the devil’s metal, *Economic Geology* 117 (2022) 1961–1983.
 - [26] M. A. Jaro, Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida, *Journal of the American Statistical Association* 84 (1989) 414–420.
 - [27] Y. Qun, X. Linfu, L. Yongsheng, W. Rui, W. Bo, D. Ke, W. Jianbang, Mineral prospectivity mapping integrated with geological map knowledge graph and geochemical data: A case study of gold deposits at raofeng area, shaanxi province, *Ore Geology Reviews* (2023) 105651.
 - [28] Y. Zhu, W. Zhou, Y. Xu, J. Liu, Y. Tan, et al., Intelligent learning for knowledge graph towards geological data, *Scientific Programming* 2017 (2017).
 - [29] C. Wang, X. Ma, J. Chen, J. Chen, Information extraction and knowledge graph construction from geoscience literature, *Computers & geosciences* 112 (2018) 112–120.