Accelerating Advanced Aluminium Alloy Design via Ontology-Driven Semantic Modelling and Large Language Models

Anton Bakuteev[1], Jaijith Sreekantan[2], Karim Houni[3] and Khuram Pervez[4]
1. Principal Data Scientist
2. Senior Principal Data Scientist
3. Head of Digital Innovation,
4. Director Data Science and AI
Emirates Global Aluminium (EGA) - Industry 4.0, United Arab Emirates
Corresponding authors: abakuteev@ega.ae, jsreekantan@ega.ae
https://doi.org/10.71659/icsoba2025-ch009

Abstract

One of the major challenges in advancing discovery of new aluminium alloys is the heterogeneous nature of the process, experimental and simulation datasets. This data is often fragmented with inconsistent terminologies necessitating an integrated semantic modelling framework for robust data harmonisation to derive meaningful insights. In our work, we develop a multi-layer semantic integration architecture that employs standardised ontologies implemented through frameworks such as Resource Description Framework (RDF) and Web Ontology Language (OWL) to systematically encode critical variables including alloy composition, processing parameters, and performance metrics into interoperable semantic entities. This structured approach facilitates precise data aggregation, automated inference, and advanced query capabilities.

To ensure standardization and compatibility with broader material science data, we align with key standards relevant to materials modelling in general and the aluminium industry in specific, such as relevant ISO standards (ISO 3522 and ISO 7722) for aluminium and aluminium alloys casting, European Materials Modelling Ontology (EMMO) for materials modelling, the Materials Design Ontology (MDO) focusing on material structures and compositions, Materials Mechanics Ontology (MMO) capturing mechanical properties, and ChEBI (Chemical Entities of Biological Interest) providing standardised terminology for chemical elements and compounds. This structured approach facilitates precise data aggregation, automated inference, and advanced query capabilities.

Integral to our methodology is the incorporation of a domain-specific large language model (LLM) that operates within these rigorously defined ontologies. The integrated semantic layer enables language models to better interpret complex experimental protocols mitigating hallucinations and improving reliability and enables users to query multiple data sources with natural language. In this work, we present a comprehensive semantic modelling framework that combines standardised ontologies with LLM-driven natural language querying to accelerate and enhance the design of novel aluminium alloys.

Keywords: Large Language Model (LLM), Semantic modelling, Ontology, Aluminium alloys, Web Ontology Language (OWL).

1. Introduction

The development of novel aluminium alloys is a time-consuming, resource-intensive process, traditionally relying on empirical approaches and expert knowledge. The complexity of alloy design stems from the high-dimensional compositional space, intricate processing-structure-property relationships, and the heterogeneity of available data. Accelerating this process requires

advanced computational techniques that can integrate disparate data sources while maintaining scientific rigor [1].

In the aluminium industry, especially for cast alloys, data fragmentation and inconsistent terminologies present significant challenges. Information about composition, microstructure, mechanical and thermal properties, processing parameters, and application-specific performance exists in various formats across research papers, industrial reports, internal databases, and simulation outputs. This fragmentation makes it difficult to establish comprehensive relationships between composition, processing, structure, and properties [2].

Traditional data integration approaches often struggle with the semantic complexity of materials science concepts. For example, a property like "strength" can refer to yield, tensile, compressive, or fatigue strength, each measured under specific conditions. Without proper semantic context, data integration leads to erroneous comparisons and conclusions that hinder alloy development [3].

Recent advances in semantic modelling and large language models (LLMs) offer promising solutions. Ontologies provide formal, explicit specifications of shared conceptualisations, enabling precise definition of entities, properties, and relationships in a domain. LLMs excel at natural language understanding and generation, enabling more intuitive interactions with complex knowledge bases. However, their application in scientific domains is often hindered by hallucination, limited domain knowledge, and a lack of specific reasoning capabilities [4].

We present an integrated approach that combines ontology-driven semantic modelling with domain specific LLMs to accelerate aluminium alloy design. Our approach focuses on aluminium cast alloys, critical for aerospace, automotive, and general engineering sectors [5]. We use the Web Ontology Language (OWL) for formalisation and integrate standards such as the Resource Description Framework (RDF), European Materials Modelling Ontology (EMMO) [6], Materials Design Ontology (MDO) [7], Chemical Entities of Biological Interest (ChEBI) [8], and Quantities, Units, Dimensions, and Types (QUDT) [9]. We also adhere to industry standards for aluminium alloys, including ISO 3522 for chemical composition and mechanical properties [10] and ISO 7722 for the global designation system of castings [11]. For all subsequent mentions, we use only the abbreviations.

1.1 Ontologies and LLMs

Ontologies provide a formal, explicit way to define shared conceptualisations, enabling precise definition of entities, properties, and relationships in a domain. LLMs excel at natural language understanding and generation, enabling more intuitive interactions with complex knowledge bases.

1.2 Integrated Approach

The integrated approach combines ontology-driven semantic modelling with domain-specific LLMs to accelerate aluminium alloy design. Our approach focuses specifically on aluminium cast alloys, which are critical for applications in aerospace, automotive, and general engineering sectors.

2. Methodology

Our approach to accelerating aluminium alloy design is built on a multi-layer semantic integration architecture that combines standardised ontologies, graph-based knowledge representation, and LLM-powered natural language interfaces.

6. Limitations and Challenges

In this section, we outline the main limitations and challenges identified in our approach:

- Ontology construction and automation: Our semi-automated workflow accelerates ontology development by combining LLM-driven parsing with a Python ontology library. However, manual verification and domain expert oversight remain essential to ensure semantic correctness and prevent errors. The Python-centric outputs can also limit adaptability to other frameworks.

- Reasoning and explainability: While the system excel at information retrieval and basic comparative analyses, complex causal reasoning about alloy behaviour demands multi-step inference. Providing transparent explanations and provenance for these reasoning chains remains an open challenge.

- Validation and trust: Ensuring the reliability of system recommendations, especially for high-stakes applications, requires robust validation workflows, transparent auditing of inference steps, and clear provenance tracking.

- Scalability and context limitations: As the ontology schema grows, embedding the entire model in an LLM context window is impractical. Our agentic graph-query approach addresses this but introduces additional overhead in tool calls and multi-hop reasoning. Improving performance, reducing latency, and enhancing tool-use proficiency for large-scale ontology traversal remain critical challenges.

7. Conclusions

We presented an integrated approach to accelerating aluminium alloy design through ontology-driven semantic modelling and LLMs. Our main contributions:

- Development of a comprehensive ontology for aluminium cast alloys using a semi-automated, LLM-assisted workflow.
- Integration of the ontology with a property graph database for advanced semantic querying.
- Enabling natural language querying of the knowledge graph using LLMs.
- Demonstration of the effectiveness with real-world alloy data and use cases.

This integration of semantic technologies with LLMs addresses data integration and knowledge accessibility challenges in materials science, making alloy knowledge more accessible and actionable, and accelerating materials innovation for diverse applications.

8. References

1. Ankit Agrawal and Alok Choudhary, Perspective: materials informatics and big data: realisation of the fourth paradigm of science in materials science, *APL Materials*, 4 (2016) 053208, http://doi.org/10.1063/1.4946894.
2. Toshihiro Ashino, Materials ontology: an infrastructure for exchanging materials information and knowledge, *Data Science Journal*, 9 (2010), 54–61, http://doi.org/10.2481/dsj.008-041.
3. Janna Hastings et al., Chebi in 2016: improved services and an expanding collection of metabolites, *Nucleic Acids Research*, 44 (2016) D1214–D1219, https://www.ebi.ac.uk/chebi/init.do.
4. Emanuele Ghedini et al., EMMO, the European materials and modelling ontology, https://github.com/emmo-repo/EMMO.

5. ISO 3522:2007, Aluminium and aluminium alloys – castings – chemical composition and mechanical properties, *International Organization for Standardization*, 2007.
6. ISO 7722:2000, Aluminium and aluminium alloy castings – global designation system for castings, *International Organization for Standardization*, 2000.
7. J. Gilbert Kaufman, Introduction to aluminium alloys and tempers, *ASM International*, (2000).
8. Youssra El-Korchi et al., The materials design ontology, https://github.com/w3id.org/mdo.
9. Model Context Protocol (MCP), official documentation and introduction, https://modelcontextprotocol.io/introduction.
10. Luca Montanelli et al., High-throughput extraction of phase–property relationships from literature using natural language processing and large language models, *Integrating Materials and Manufacturing Innovation*, 13 (2024), 396–405, http://doi.org/10.1007/s40192-024-00344-8.
11. Ralph Hodgson et al., QUDT – quantities, units, dimensions and types, https://qudt.org.
12. Lukas Twist et al., LLMs love Python: a study of LLMs' bias for programming languages and libraries, arXiv preprint arXiv:2503.17181, 2025, https://arxiv.org/abs/2503.17181.
13. Yanpeng Ye et al., Construction and application of materials knowledge graph in multidisciplinary materials science via large language model, *NeurIPS* 2024, http://doi.org/10.48550/arXiv.2404.03080.
14. Wayne Xin Zhao et al., A survey of large language models, *arXiv* preprint arXiv:2303.18223, (2023), http://doi.org/10.48550/arXiv.2303.18223.
15. Shunyu Yao et al., ReAct: synergizing reasoning and acting in language models, Advances in Neural Information Processing Systems (NeurIPS), 2022, https://arxiv.org/abs/2210.03629.