Knowledge Graphs for Data-Driven Computational Materials Research

Abril Azócar Guzmán[1], Elaheh Akhoundi[1], Stefan Sandfeld[1]
[1]Institute for Advanced Simulations – Materials Data Science and Informatics (IAS-9), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

a.azocar.guzman@fz-juelich.de

The advancement of artificial intelligence in materials science is closely linked to the availability of structured, interoperable, and machine-readable data. In computational materials research, heterogeneous simulation environments, diverse file formats, and incomplete metadata descriptions hinder data reuse, cross-study aggregation, and automated analysis. To enable AI-driven materials discovery, simulation data must be represented in a semantically consistent manner. We present an ontology-driven framework for the formal representation of computational materials samples and atomistic simulation workflows as application-level knowledge graphs. The Computational Materials Sample Ontology (CMSO) [1] provides a structured description of material structures and crystallographic defects, while the Atomistic Simulation Methods Ontology (ASMO) [2] captures methodological and workflow metadata. Semantic annotation is embedded directly within simulation workflows using atomRDF [3], enabling automatic generation of knowledge graphs.

We demonstrate the scientific utility of this approach through two use cases based on atomistic simulation data. First, we perform the extraction, aggregation, and comparison of materials properties, including defect formation energies and thermodynamic quantities, from heterogeneous simulation datasets. Second, we extend the knowledge graph beyond structured workflows by integrating information extracted from scientific literature using large language models (LLMs). The resulting knowledge graph supports complex queries, cross-study aggregation, and identification of implicit relationships that are not readily accessible in conventional data storage formats. This ontology-aligned representation establishes a scalable foundation for FAIR [4] and AI-ready materials data infrastructures. By combining formal ontological modeling, workflow-level semantic annotation, and LLM-assisted extraction from unstructured sources, this work provides a robust infrastructure for data-driven and AI-supported materials research.

References

[1] A. Azócar Guzmán, S. Sandfeld, Zenodo (2026) DOI: 10.5281/zenodo.10805535
[2] A. Azócar Guzmán, S. Sandfeld, Zenodo (2026) DOI: 10.5281/zenodo.10805591
[3] S. Menon, A. Azócar Guzmán, atomRDF, GitHub repository, Version 0.11.0 (2026), https://github.com/pyscal/atomRDF
[4] M. D. Wilkinson et al., Sci. Data 3 (2016) 160018.

Figures

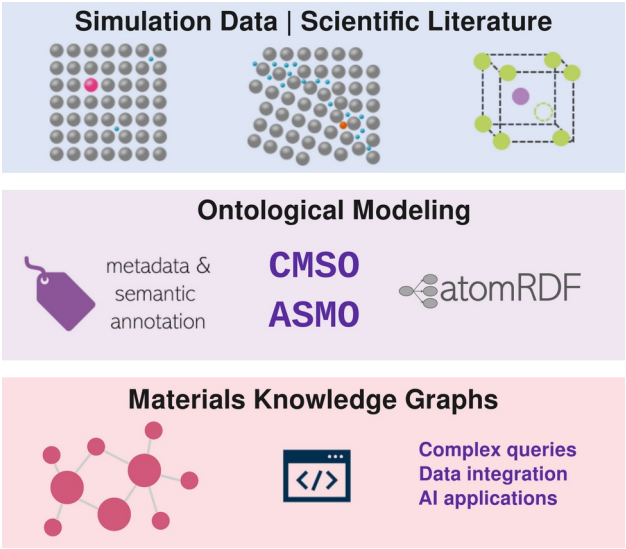

Figure 1. Ontology-aligned representation of simulation and literature data for scalable materials knowledge graphs and AI integration.