KONDA: An LLM-based Tool for Semantic Annotation and Knowledge Graph Creation Using Ontologies for Research Data

Achieving semantic interoperability of research data is a critical enabler for cross-domain data integration, reuse, and knowledge discovery [1]. Although it is widely recognized that there is a need to align heterogeneous datasets using shared vocabularies and ontologies, doing so remains a considerable challenge for many researchers in practice [2], [3].

Barriers include the following challenges:

- (C1) Lack of expertise in ontologies: Many researchers are unfamiliar with ontology engineering and the conceptual foundations of semantic annotation.
- (C2) Absence of established domain ontologies: While in some domains, such as medicine, controlled vocabularies are well established, there exist other domains, such as production engineering, where suitable ontologies do not exist or are not widely adopted. This can make it difficult to identify reusable ontologies.
- (C3) Technical barriers: The knowledge required to work with ontology-related technologies, such as RDF or mapping tools, often presents an entry barrier for researchers.
- (C4) Tool heterogeneity: The need to operate and integrate multiple disconnected tools adds cognitive and technical overhead.
- (C5) Limited resources: Researchers typically face time constraints, making it difficult to invest in learning complex tools or developing custom pipelines for research data management.
- (C6) Proprietary solutions: Many existing tools for semantic mapping (e.g., Talend [4]) are proprietary and may therefore not be appropriate to use for scientific work.

To address these challenges, we present KONDA, an LLM-based tool designed to support researchers in enriching their datasets with metadata from ontologies and in constructing explorable knowledge graphs from the enriched datasets, all within an accessible, user-centered workflow.

The researcher interacts with the tool according to the following structure:

1. An interface prompts the user to upload their research dataset, including optional supplementary documents (e.g., protocols, DMPs, README files) to provide the tool with a broader context.
2. The user is supported in the selection of suitable ontologies via a direct integration with the TIB Terminology Service [5], with the option to add custom ontologies.
3. The tool performs an automated LLM-based semantic annotation of the dataset based on the previously uploaded context material and selected ontologies. A feedback screen enables the user to correct faulty annotations.
4. An output is generated in RDF format with an immediate visualization of the annotated data in the form of a knowledge graph.

KONDA's architecture comprises a user interface guiding the interaction with the tool, a server backend managing user sessions and data processing, and an API layer that connects the tool to an LLM, where the process of semantic enrichment is conducted with techniques such as named entity recognition, relation extraction, and ontology-based annotation.

Through KONDA, a guided, interactive tool is provided in which users receive LLM-assisted suggestions and the opportunity to intuitively explore their enriched data directly through automated knowledge graph creation, thus reducing the need for formal training in semantic technologies (C1). The discovery of reusable ontologies is enabled through the integration of terminology services (C2). Users of the tool interact over a clickable interface, thereby reducing technical requirements to a minimum (C3). KONDA unifies all steps of the enrichment pipeline within a single, cohesive environment (C4). The tool's semi-automated workflow provides fast and visually supported results with minimal manual effort (C5) while retaining opportunities for human feedback to ensure output quality. Finally, KONDA's modular backend supports the deployment of both proprietary and open LLMs to power the semantic annotation functionality (C6).

KONDA empowers researchers to semantically enrich their datasets with minimal technical effort, offering an integrated, transparent, and adaptable solution. Future research directions of the tool include persistent graph storage, automated ontology recommendations, and in-depth evaluation in real-world settings. By building on the strengths of LLMs and prioritizing usability, KONDA provides a robust foundation for advancing data interoperability across disciplines.

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2023 Internet of Production – 390621612.

References

[1] H. P. Sustkova, K. M. Hettne, P. Wittenburg, et al., "Fair convergence matrix: Optimizing the reuse of existing fair-related resources," Data Intelligence, vol. 2, no. 1-2, pp. 158–170, Jan. 2020. DOI: 10.1162/dint_a_00038.

[2] H. Liyanage, P. Krause, and S. de Lusignan, "Using ontologies to improve semantic interoperability in health data," Journal of innovation in health informatics, vol. 22, no. 2, pp. 309–315, 2015. DOI: 10.14236/jhi.v22i2.159.

[3] I. M. Putrama and P. Martinek, "Heterogeneous data integration: Challenges and opportunities," Data in Brief, vol. 56, p. 110 853, Oct. 2024. DOI: 10.1016/j.dib.2024.110853.

[4] Q. I. AB. "Talend." (first published 2006), [Online]. Available: https://www.talend.com (visited on 04/16/2025).

[5] L. I. C. for Science and T. U. Library. "Tib terminology service - ontologies." (first published 2023), [Online]. Available: https://terminology.tib.eu/ts/ontologies (visited on 04/15/2025).