Karlsruhe Institute of Technology

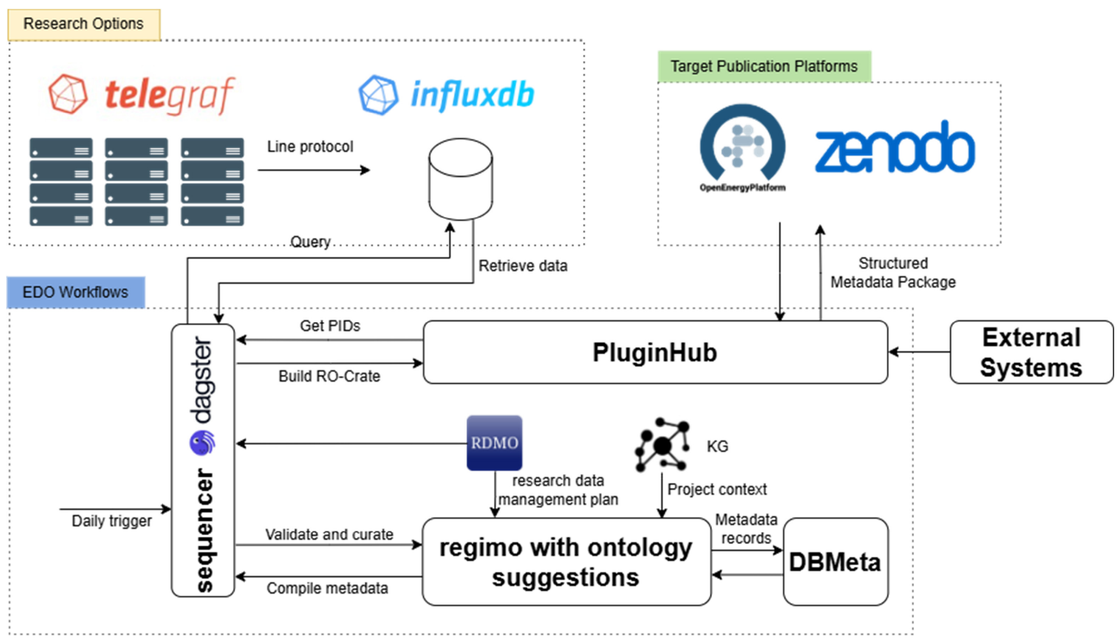
Institute for Automation and Applied Informatics (IAI)

From Strings to Semantics: Semi-Automatic Ontology Suggestions for FAIR Energy Data Publication Workflows

Nan Liu, Mohamed-Anis Koubaa, Wolfgang Suess

Motivation

- **FAIR (Findable, Accessible, Interoperable, and Reusable)** data publications are important for enabling open energy research across interdisciplinary domains.
- The realization of the FAIR principle for data still faces many challenges, such as the diversity of data formats, semantic heterogeneity, lack of formalized ontologies, and error-prone manual annotation of data. These challenges impede the effective sharing and integration of energy data.
- To address these issues, we propose an automated ontology term recommendation system based on ontology embedding and semantic similarity, aiming to facilitate FAIR data publication in energy research.

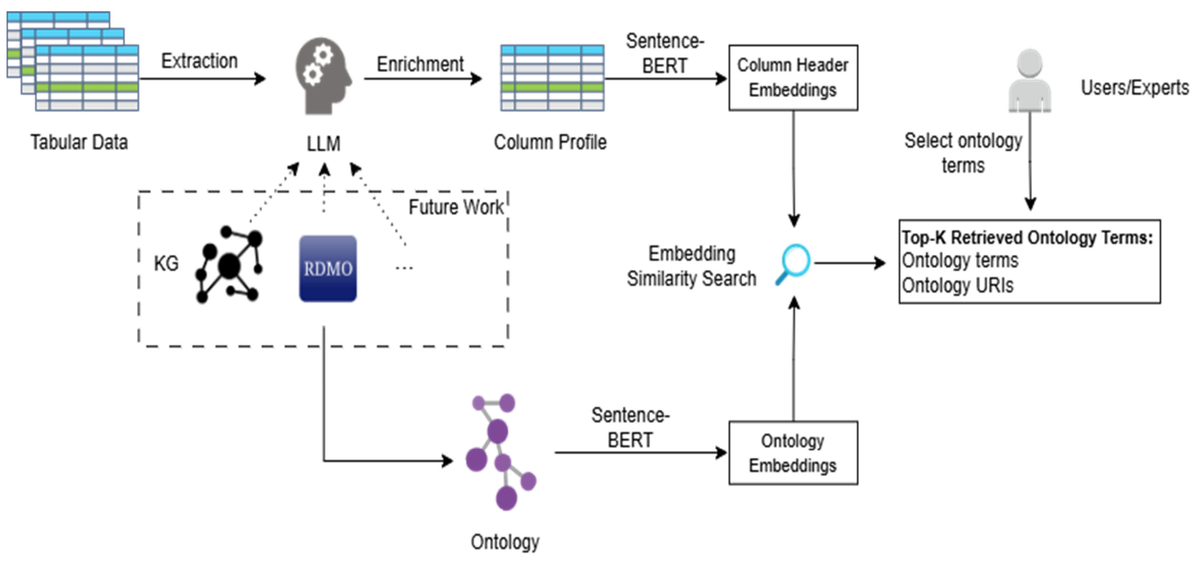


Energy Data Orchestrator (EDO)

- EDO is a machine-actionable data management plan, which ensures energy-related data and metadata documentation, sharing, and reuse
- Integrates data, metadata, standards, and ontologies
- Toward semi-automatic metadata preparation and semantic annotation
- Provides a central hub for energy data management
- End-to-end support for FAIR data publication

Benefits

- **Efficiency:** accelerate FAIR data publishing
- **Consistency:** standardize annotations across projects
- **Interoperability:** unify heterogeneous energy datasets
- **Scalability:** support large and evolving datasets
- **Foundation** for knowledge graph compilation and construction



Ontology Annotation Support

- Inputs: data table column headers and project context from external knowledge base (RDMO, KG, …)
- Outputs: Top-K terms with URI
- LLM helps to make columns machine-understandable
  - Normalize Units
  - Abbreviation expansion
  - Use project context to disambiguate
  - Build a compact column profile for matching
- Embedding-based semantic search in the target ontology

Conclusion

In this work, we present an embedding-based automatic semantic annotation system that aims to help researchers reduce the semantic annotation barriers that they face during the FAIR data publication process. Semantic similarity is used to match the tabular metadata field name with respective ontology terms and recommend the top-K most relevant ontology terms to users. The system reduces manual effort, improves annotation consistency, and accelerates the data publication workflow.

Nan Liu – nan.liu@kit.edu
https://www.iai.kit.edu/2154_4782.php
Karlsruhe, Germany

Acknowledgments

This work was also supported by the Helmholtz Metadata Collaboration (HMC, https://www.helmholtz-metadata.de/), an incubator platform of the Helmholtz Association within the framework of the Information and Data Science strategic initiative
The authors would like to thank the German Federal Government, the German State Governments, and the Joint Science Conference (GWK) for their funding and support as part of the NFDI4Energy consortium. The work was funded by the German Research Foundation (DFG) – 501865131 within the German National Research Data Infrastructure (NFDI, www.nfdi.de)