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

Nan Liu[1,*], Mohamed-Anis Koubaa[1], and Wolfgang Suess[1]

[1]Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany
[*]Corresponding author: nan.liu@kit.edu.

Introduction

Energy systems are becoming increasingly complex as they interconnect and integrate multiple energy sources (such as energy storage systems, smart grids, etc.). The operation and optimization of such systems not only rely on a large amount of structured data, but also continuously generate extensive data resources. How to effectively share and reuse these data is important for open science research and data-driven energy management. However, there are still many challenges in real-world applications. For example, different energy systems and research projects often use their own terminology and nomenclature, due to the lack of unified semantic standards, the same concept will have different field names in different datasets, which thereby limits the reusability of the data for semantic search, automated reasoning, and cross-system integration. To achieve efficient reuse of energy data, semantic annotation is considered a core technology following the FAIR principles [1]. After the semantic annotation, the data has a clear semantic definition and can be released to an open data platform or knowledge graphs. However, manual semantic annotation is extremely difficult and error-prone. Without assistive tools, researchers need to perform term-by-term terminology queries and comparisons, and manually map them to the target ontology. Different people may have different understandings of the terminology, which may have inconsistent results. This semantic inconsistency will directly affect the subsequent data integration, information retrieval, and knowledge reasoning tasks.

With the extensive use of Natural Language Processing (NLP), especially pre-trained language models, more and more language understanding and semantic reasoning tasks can be automated by NLP models [2–6]. Semantic annotation maps the natural language descriptions to formalized domain ontologies and can also be improved with the help of NLP. The main contribution of this paper is the development of an ontology term recommendation system to assist or partially replace the manual semantic annotation process. The system automatically extracts descriptive information from energy-related tabular data sets and then uses a large language model to automatically enrich the data descriptions. By calculating the semantic similarity, the system will generate the top-K most relevant recommendation results. Through the proposed system, users can quickly select appropriate ontology terms for their data, thereby reducing annotation barriers and improving the accuracy and consistency of semantic annotation.

Methods

Based on the FAIR principles, we outline a data publication workflow, as shown in Figure 1. First, Energy Data Orchestrator (EDO) collects and organizes information from the Research Data Management Organizer (RDMO)[1] and other external resources. EDO compiles this information into structured metadata and provides a user interface where data providers can edit (meta)data descriptions. To ensure that data are FAIRly published, we design an ontology term recommendation system that achieves semi-automatic annotation, as shown in figure 2. After the EDO has collected the tabular data and its metadata, the system first performs entity-level extraction. However, the row metadata alone may lack semantic diversity, especially when column names are

[1]`https://rdmorganiser.github.io/`

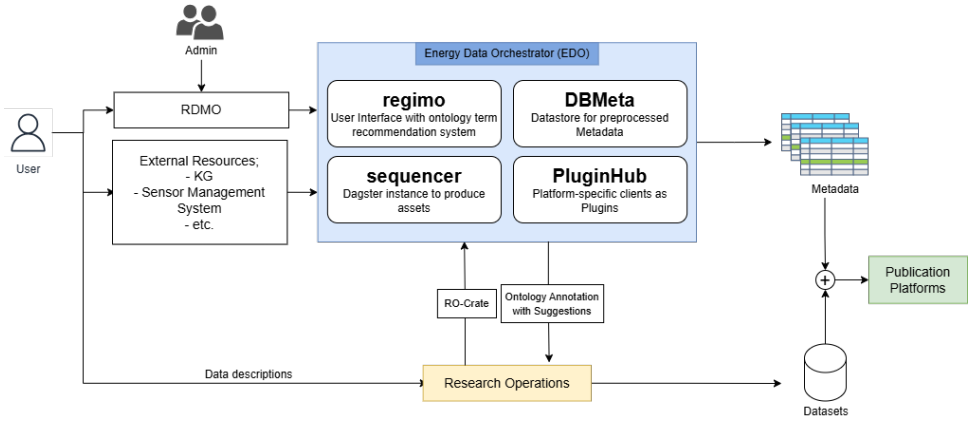

Figure 1: FAIR Energy Data Publication Pipeline with EDO

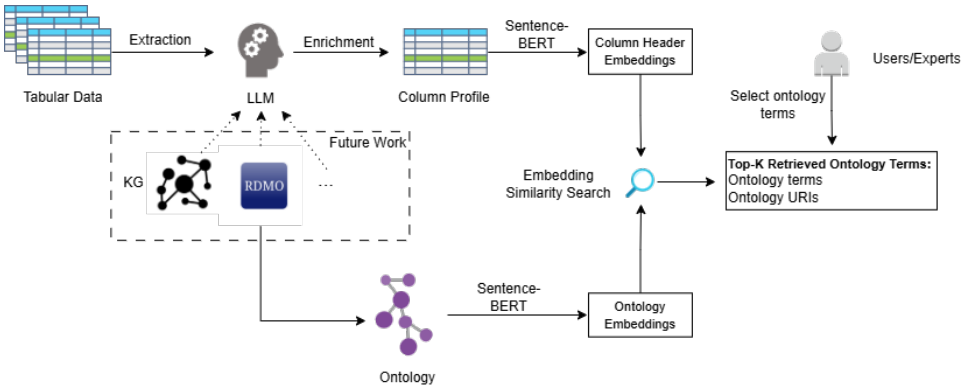

Figure 2: Ontology Recommendation Workflow

abbreviated, ambiguous, or domain-specific. Since most embedding models are not pre-trained on such domain-specific corpora, the resulting embeddings may fail to capture useful semantic information from them. To address this problem, we introduce large language models (LLMs) to enhance the expression of metadata descriptions. A pre-trained Sentence-Bidirectional Encoder Representations from Transformer (SBERT) [7] has been used to encode both LLM-enriched data descriptions and ontology into the embeddings in the same embedding space. Then we use a recommendation strategy based on semantic similarity retrieval. We calculate the similarity score between the embeddings of metadata descriptions and the pre-calculated ontology embedding. We renk all similarity scores and return the first top-K terms with the highest scores as the recommendation results.

Discussion

In this paper, we propose an ontology term recommendation system that aims to help researchers reduce the semantic annotation barriers that they face during the FAIR data publication process. The goal of the proposed system is to minimize the misannotation by users due to a lack of ontology knowledge, thereby enhancing the interoperability and reusability of the energy research data, which supports open science research. 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.

However, the system still has some limitations. For example, it may still have some errors when dealing with semantic ambiguity. In future work, we will introduce a user feedback mechanism to make the recommendation process more interactive and accurate. Besides, we will consider adding an ontology matching module to the system to support cross-ontology term recommendation and annotation, which can enhance the data interoperability and better meet the goal of FAIR data management.

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten,

L. B. da Silva Santos, P. E. Bourne *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016.

[2] S. Selva Birunda and R. Kanniga Devi, "A Review on Word Embedding Techniques for Text Classification," in *Innovative Data Communication Technologies and Application*, J. S. Raj, A. M. Iliyasu, R. Bestak, and Z. A. Baig, Eds. Springer, pp. 267–281. doi: 10.1007/978-981-15-9651-$3_2$3 .

[3] D. Ganguly, D. Roy, M. Mitra, and G. J. Jones, "Word embedding based generalized language model for information retrieval," in *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 2015, pp. 795–798.

[4] K. A. Hambarde and H. Proenca, "Information retrieval: recent advances and beyond," *IEEE Access*, vol. 11, pp. 76 581–76 604, 2023.

[5] M. Esposito, E. Damiano, A. Minutolo, G. De Pietro, and H. Fujita, "Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering," *Information Sciences*, vol. 514, pp. 88–105, 2020.

[6] K. Nassiri and M. Akhloufi, "Transformer models used for text-based question answering systems," *Applied Intelligence*, vol. 53, no. 9, pp. 10 602–10 635, 2023.

[7] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. doi: 10.18653/v1/D19-1410 . [Online]. Available: https://aclanthology.org/D19-1410/