

A Multi-Agent System for Semantic Mapping of Relational Data to Knowledge Graphs

Milena Trajanoska, Riste Stojanov, and Dimitar Trajanov

University Ss. Cyril and Methodius, Skopje, N.Macedonia
`{milena.trajanoska,riste.stojanov,dimitar.trajanov}@finki.ukim.mk`

Abstract. Enterprises often maintain multiple databases for storing critical business data in siloed systems, resulting in inefficiencies and challenges with data interoperability. A key to overcoming these challenges lies in integrating disparate data sources, enabling businesses to unlock the full potential of their data. Our work presents a novel approach for integrating multiple databases using knowledge graphs, focusing on the application of large language models as semantic agents for mapping and connecting structured data across systems by leveraging existing vocabularies. The proposed methodology introduces a semantic layer above tables in relational databases, utilizing a system comprising multiple LLM agents that map tables and columns to Schema.org terms. Our approach achieves a mapping accuracy of over 90% in multiple domains.

Keywords: Knowledge Graphs · Semantic AI Agents · Data Integration.

1 Introduction

In today’s data-driven enterprises, information is often distributed across multiple siloed databases, limiting the ability to fully leverage its value. Achieving seamless integration between heterogeneous data sources remains a major challenge, particularly due to differences in schema design, semantics, and data representation. Knowledge Graphs (KGs) [5] have emerged as a powerful paradigm for bridging these gaps by creating a unified, semantic view over disparate datasets.

Traditional data integration efforts, however, often require extensive manual work to align schemas and define mappings. Recent advances in Large Language Models (LLMs) [1] provide new opportunities to automate complex reasoning tasks, including semantic mapping and knowledge graph construction. LLMs have shown impressive performance in tasks that involve understanding relationships between concepts, making them ideal candidates for assisting in structured data integration.

In this paper, we present a novel semantic multi-agent system that leverages LLM agents to automate the integration of relational databases into a unified knowledge graph. Our approach introduces a semantic layer above relational data by mapping tables and columns to concepts from the Schema.org vocabulary [2].

Each agent in the system has a dedicated role: mapping relational structures to semantic concepts, establishing relationships between entities, and validating the generated knowledge graph.

We evaluate our system using the Spider dataset [14], a complex benchmark featuring databases from diverse domains. By mapping real-world relational schemas to Schema.org concepts, we demonstrate that our multi-agent system effectively automates the semantic integration process, achieving high mapping accuracy across several application domains. Our results highlight the potential of LLM-driven semantic agents to significantly reduce the manual burden of data integration and to build more interoperable enterprise knowledge bases.

2 Literature review

Recent research has explored various approaches to knowledge graph (KG) and ontology construction, and data integration, particularly in enterprise and industry contexts.

The Virtual Knowledge Graph (VKG) paradigm [12] offers an alternative to traditional data integration approaches by providing flexible, virtual graphs that embed domain knowledge over existing data sources. The approach provides tools for Ontology-based data access, which have significant use-cases in multiple industry applications.

In the context of Industry 4.0 (I4.0) where different systems exist, the demand for the creation of an integrated view increases. One such initiative is the Bosch Industry 4.0 Knowledge Graph (BI40KG) [4] which was developed with the goal of integrating data from different sources, improving interoperability and traceability across manufacturing systems.

Many efforts have been made for developing enterprise knowledge graphs (EKGs), in line with this, an Enterprise Knowledge Graphs Framework [3] was created, which aims to bridge the gap between the increasing need for EKGs and the lack of formal methods for realising them.

The use of Large Language Models (LLMs) in structured information extraction has gained significant traction recently. Findings suggesting that LLMs excel at reasoning tasks, and are being applied to information extraction tasks more often [13].

Our previous work in the field compares general-purpose LLMs like ChatGPT with specialized models such as REBEL for joint entity and relation extraction [11]. Using sustainability-related texts, pipelines for automatic Knowledge Graph construction were built. Results show that LLMs can improve accuracy in extracting structured knowledge, and foundational models also show promise for automated ontology generation.

3 Methodology

This section describes the methodology for creating the semantic multi-agent system as well as the data source used in evaluating the approach. The section is divided into the following subsections: Data description - stating the dataset used for evaluation and the semantic vocabularies, Graph-vector store construction - detailing the process of constructing a hybrid vector store containing subgraphs in a vector representation, and Multi-agent system architecture - detailing the entire flow of the mapping system along with the AI agents' roles.

3.1 Data description

The Yale Spider database[14] is a benchmark dataset used in semantic parsing and text-to-SQL tasks, featuring real-world relational data from various domains. The Spider challenge aims to create natural language interfaces for databases spanning multiple domains. It includes 10,181 questions and 5,693 distinct complex SQL queries across 200 databases with multiple tables, representing 138 domains.

In this study, the database serves as the foundation for evaluating our multi-agent approach, allowing us to demonstrate the effectiveness of mapping relational data to a knowledge graph.

For creating the mappings, we utilize existing semantic vocabularies, which are defined and maintained by established organizations over a long time period. The reason for this is to avoid hallucinating terms that do not have definitions in any well-known ontologies.

With the goal of including as many general concepts as possible, we use the Schema.org vocabulary [2], which represents a collection of schemas (types and properties) that can be used to markup content on the Web, enabling search engines to better understand the context of Web pages. It describes entities, their relationships, and actions, and can be readily expanded using a clearly defined extension framework. As of 2024, more than 45 million websites have used Schema.org to annotate their pages, resulting in over 450 billion structured data objects [2].

3.2 Graph-vector store construction

To enable our use case for generating correct semantic mappings, we have created a custom graph-vector index for retrieval augmented generation [7] using the term definitions in Schema.org.

The index creation process is depicted in Figure 1. For each term, we extract its URI, type, human-readable comment, and label, as well as the domain and range for property terms. Based on this information, we construct a one-hop subgraph for each term. This enriched subgraph offers a more detailed semantic context. Finally, each subgraph is transformed into a vector representation using a configurable text embedding transformer model [8].

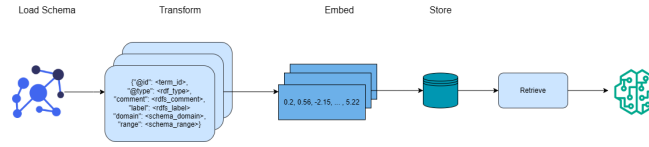


Fig. 1. Graph - Vector store construction with ontology sub-graphs

The generated vectors are stored in a vector store. The subgraphs are retrieved from the vector store during the table mapping process, depending on their semantic similarity to the columns and values in the table being mapped.

3.3 Multi-agent system architecture

The system consists of multiple GPT-4o-mini [6] agents with different responsibilities:

1. **Mapping agent:** Maps columns to corresponding Schema.org terms.
2. **Relation agent:** Determines primary and foreign keys in tables, creating links between entities in the knowledge graph.
3. **Validator agent:** Validates and corrects the mapping and relations produced by the two previous agents.

The mapping process is fully represented in Figure 2. The process operates in a table-wise fashion. For each table, the Mapping agent receives the table name, all its column names, summary statistics for each column's values, a sample of the first K rows, and any available textual descriptions for the table or its columns.

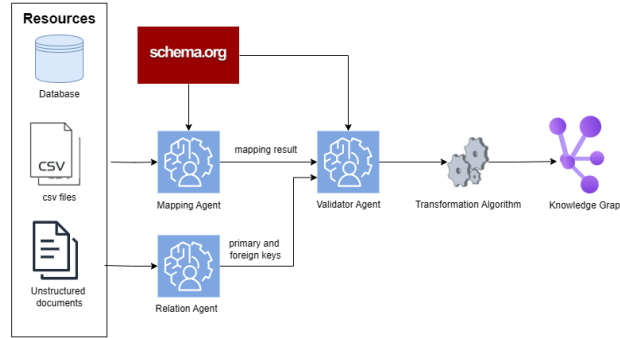


Fig. 2. Data flow of the semantic multi-agent system

In addition, the mapping LLM is provided with the most similar terms from Schema.org to the table and column in order to map them accordingly. The most

similar terms are retrieved from the custom graph-vector store described in the previous subsection.

In the subsequent step, all mapped tables are integrated into a unified knowledge base by accurately identifying primary and foreign key relationships through the Relation agent. This structured integration enables seamless interaction with the knowledge base by both internal and external systems.

In the final step, the Validator agent inspects and refines each table mapping and the final relation mapping to ultimately produce a more accurate result. The Validator agent may remove or re-map any relations it deems necessary to improve the accuracy.

All the prompts for the different types of agents are available in our GitHub repository [10] under the prompts folder.

All of the agents in the system output a confidence variable which can be one of the three following categories: HIGH, MEDIUM or LOW, to describe their confidence in the completion of their dedicated task. The categories are averaged from all steps to produce the final confidence class.

4 Results and discussion

The mapping system was evaluated on multiple databases from different domains present in the Spider data source. The accuracy of the generated mappings was assessed through manual validation, evaluating whether the Schema.org terms chosen by the system correctly reflected the semantics of the corresponding database elements. The evaluation dataset can be found in our GitHub repository [9] in the eval folder.

This section summarizes the findings for the entire execution of the system, including the time to generate the mapping result, as well as the accuracy of the mapping process.

Table 1 shows that the execution time generally increases with the number of tables and columns being mapped. This is likely due to the complexity of larger tables and databases in the relationships they model.

Table 1. System execution time based on the number of tables and columns being mapped.

Number of tables	Number of columns	Execution time (sec)
3	10	122
6	25	179
6	31	191
11	42	342
12	47	295

Concerning the validity of the generated mappings, we have conducted an evaluation for five different fields, namely: retail, movies, automotive, apartments, and delivery. We gathered all the datasets from these domains present

in the Spider database and manually evaluated the accuracy of the mappings in each domain.

The accuracy of the multi-agent system across the five domains is reported in Table 2. Apart from the mapping result, the multi-agent system returns the confidence of the mapping process, which can be one of the following categories: HIGH, MEDIUM, and LOW. For all confidence levels, the results are displayed in the table.

Table 2. Accuracy of the multi-agent system on the five chosen domains presented by human evaluators

Sector	Overall (%)	HIGH (%)	MEDIUM (%)	LOW (%)
Retail	78.72	95.83	64.29	55.56
Movies	90	85.71	100.00	/
Appartments	93.54	100.00	88.89	50.00
Automotive	84	100.00	80.00	25.00
Delivery	80.95	96.30	66.67	0

The highest accuracy (93.54%) was observed in the 'Appartments' domain, while the lowest (78.72%) occurred in the 'Retail' domain. Even for retail, the score is still satisfactory. It is evident that the mappings with certainty equal to HIGH are, in general, the most correct ones.

5 Conclusion

This paper introduced a semantic multi-agent system for enterprise data integration using Knowledge Graphs and LLMs. By aligning relational schemas with Schema.org, our approach enhances interoperability and enables unified, semantically rich graphs across domains.

LLM agents were assigned specialized roles—mapping, extraction, and validation—resulting in a scalable and automated integration pipeline. Experiments on the Spider benchmark showed high mapping accuracy and cross-domain applicability.

The main contributions of this work are:

- The design and implementation of a **semantic multi-agent system** for automating relational-to-graph data integration.
- The innovative use of **LLM agents** to perform semantic mapping and reasoning over structured data using existing vocabularies.
- The empirical evaluation on a real-world benchmark, validating the system’s practicality and generalization capability.

Future work includes extending the system to support custom domain ontologies, enhancing the reasoning capabilities of the agents through fine-tuning, and investigating scaling strategies for integrating larger and more heterogeneous datasets.

6 Acknowledgements

This publication is based upon work from COST Action CA23147 GOBLIN - Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology, <https://www.cost.eu>).

References

1. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al.: A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* **15**(3), 1–45 (2024)
2. schema.org Community Group. Schema.org, W.: <https://schema.org/docs/about.html>
3. Galkin, M., Auer, S., Kim, H., Scerri, S.: Integration strategies for enterprise knowledge graphs. In: 2016 IEEE Tenth International Conference on Semantic Computing (ICSC). pp. 242–245. IEEE (2016)
4. Grangel-González, I., Lösch, F., ul Mehdi, A.: Knowledge graphs for efficient integration and access of manufacturing data. In: 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). vol. 1, pp. 93–100. IEEE (2020)
5. Hogan, A., Blomqvist, E., Cochez, M., d’Amato, C., Melo, G.D., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S., et al.: Knowledge graphs. *ACM Computing Surveys (Csur)* **54**(4), 1–37 (2021)
6. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024)
7. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* **33**, 9459–9474 (2020)
8. Patil, R., Boit, S., Gudivada, V., Nandigam, J.: A survey of text representation and embedding techniques in nlp. *IEEE Access* **11**, 36120–36146 (2023)
9. Trajanoska, M.: Knowledge graph agents - eval. <https://github.com/MilenaTrajanoska/knowledge-graph-agents/blob/main/examples/data/eval/Semantic%20Agents%20Evaluation.csv> (2025), accessed: 2025-06-08
10. Trajanoska, M.: Knowledge graph agents - prompts. <https://github.com/MilenaTrajanoska/knowledge-graph-agents/tree/main/src/graphindex/common/prompts> (2025), accessed: 2025-06-08
11. Trajanoska, M., Stojanov, R., Trajanov, D.: Enhancing knowledge graph construction using large language models. *arXiv preprint arXiv:2305.04676* (2023)
12. Xiao, G., Ding, L., Cogrel, B., Calvanese, D.: Virtual knowledge graphs: An overview of systems and use cases. *Data Intelligence* **1**(3), 201–223 (2019)
13. Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., Chen, E.: Large language models for generative information extraction: A survey. *Frontiers of Computer Science* **18**(6), 186357 (2024)
14. Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., et al.: Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *arXiv preprint arXiv:1809.08887* (2018)