Towards the Definition of a Unified Metamodeling Semantics for Knowledge Graphs

Roberto Maria Delfino*

Sapienza University of Rome - Department of Computer, Control and Management Engineering

Abstract

The increasing importance of Knowledge Graphs (KGs) in various Artificial Intelligence (AI) contexts, e.g., Data Preparation and Neuro-Symbolic AI, highlights the need for a more robust and unified semantic foundation for KGs. Although various approaches exist for creating and managing KGs, each presents limitations. This work addresses these challenges by proposing a unifying semantic framework for knowledge graphs based on metamodeling. Our research aims to define a general semantic framework grounded in formal logic, capable of assigning a well-defined meaning to diverse types of KGs and providing a solid theoretical basis for querying them. Overall, this work aims to provide a contribution to enhancing data quality and interpretability in AI systems through a formally grounded semantic characterization of knowledge graphs.

Keywords

knowledge graphs, metamodeling semantics

1. Introduction

In recent years, the interest in Knowledge Graphs (KG) arose in both research and industry due the role that they can play in different contexts of Artificial Intelligence. Among all of such contexts, two particularly prominent ones where KGs are particularly relevant are represented by Data Preparation [1] and Neuro-Symbolic AI [2, 3].

Data Preparation is the process of collecting, cleaning, transforming, and organizing raw data in order to make it ready to be used by Machine Learning (ML) systems or to be analyzed. It is well-known that Data Preparation represents a highly costly phase, both in terms of required amount of time to be addressed, and of effort due to the complexity of challenges to be overcame, for all those AI processes which require high-quality data in order to provide better performances. Finding solutions to the general problem of Data Preparation is a challenging task, in that, while several different approaches have been proposed, their effectiveness can vary. It is a general opinion that standardization obtained through uniform data representations, vocabularies, and terminology in the data modeling phase can dramatically reduce the costs of Data Preparation, especially those derived from the need of repetitive actions which typically characterize it. In this context, KGs represent a useful tool, in that they can help in both data gathering and in integrating heterogeneous data sources by means of a proper semantic characterization, while also providing reasoning capabilities to be adopted to obtain useful insights.

Neurosymbolic AI has recently emerged as a field of research whose fundamental importance is motivated by the awareness that current AI systems, which are exclusively or predominantly based on statistical machine learning techniques, are characterized by a number of limitations. The extent of these limitations could be significantly reduced through the use of deductive semantic technologies inherent to symbolic AI, e.g., those based on formal languages and logic. One example of such limitations is the potential threat represented by biases originating from input data or within the internal mechanics of the black-box systems which may affect them, making them not completely trustable and reliable [4]. Such biases highlight the importance of providing tools that are capable of properly explaining and interpreting the results that those systems produce, in order to be able to identify potential issues and

SEBD 2025: 33rd Symposium On Advanced Database Systems, June 16–19, 2025, Ischia, IT

*Corresponding author.

delfino@diag.uniroma1.it (R. M. Delfino)

0000-0002-5492-5290 (R. M. Delfino)

make the systems themselves more trustable and reliable (this is also justified by the so-called "right to explanation" regulated by the EU's GDPR [5]). Notable examples of the use of knowledge graphs in symbolic and subsymbolic integration contexts are RAG LLMs [6, 7], where knowledge graphs, can be used as external knowledge bases to improve the accuracy, consistency, and interoperability of the model's responses. In particular, KGs represent an ideal tool, since, when provided with a correct semantics, they allow to perform both inductive and deductive reasoning tasks, thus representing a bridge between symbolic and sub-symbolic approaches, whose integration is a prerequisite for neuro-symbolic systems.

Several frameworks exist which are utilized to create and manage KGs, such as plain graph structures [8], graph databases equipped with some query language [9], RDFS graphs [10], or graphs as DL-based knowledge bases [11], but each of these approaches is characterized by forms of limitations that should be addressed to fully exploit the modeling and reasoning capabilities of KGs, in order to make them really effective in providing a tool to pursue high-quality data and explainability in AI systems.

This paper presents an overview on our research on the formalization of a unifying semantic framework for knowledge graphs, starting from the state of the art with a mention on the motivations (Section 2) sustaining our research. In Section 3 we describe the general approach we adopted to reach the goal. In Section 4, we present some preliminary results. Finally, in Section 5 we present some research paths we intend to follow as a prosecution of our research.

2. Related works

Knowledge graphs have been extensively studied in different AI research areas. As of now, several different frameworks to create and manage knowledge graphs coexist. In some cases KGs are used as plain graph databases mostly utilized for inductive inference tasks based on the topology of the graph itself. An example of such tasks is represented by KG embeddings for link prediction [12, 13]. In this approach a KG is seen as a database structured as a directed graph, possibly equipped with some additional syntactic features, such as labels and properties associated to nodes or edges (property graphs [14]). Domain elements are represented by nodes, while edges represent the relationships existing between them. In some cases, these KGs come with specialized query languages, e.g., Cypher used for the main property graphs [9]. One important limitation of this kind of KGs comes from the fact that they are not equipped with a rich semantics: in most cases they are just plain graph databases providing extensional knowledge exclusively, while not specifying any kind of intensional knowledge describing the relations between classes of objects or relationships. This severely limits the capabilities to model sophisticated domain features, to check for consistency or to apply deductive reasoning on the graph. The corresponding query languages are typically based on the notion of pattern-matching, i.e., a query over the KG takes into account a single logical model, corresponding to the graph itself, and it looks for bindings to be used for matching the patterns specified in the query. This can be a serious limitation in terms of semantic characterization for query answering.

An alternative formalism is represented by the RDFS framework [15]. An graph of this kind is represented as a set of triples of the form $\langle \mathtt{s}\ \ \mathtt{p}\ \ \mathtt{o} \rangle$ where p represents a relationship (called *predicate*) existing between a *subject* s and an *object* o. Every triple can be seen as a subgraph, where s and o represent nodes and p represents a directed edge from s to o. RDFS introduces a vocabulary of special symbols characterized by a specific semantics, which can be used to represent intensional knowledge in terms of classes, predicates, ISA relationships between pairs of classes or predicates, class and property membership. Despite this, the standard semantics of RDFS falls short in adequately capturing some semantic aspects (see [16]), and, moreover, it lacks forms of incomplete information in domain modeling related to some symbols of the vocabulary (e.g., `rdf:Statement`).

Another approach is represented by KGs as DL knowledge bases. A DL knowledge base consists in a set of axioms describing both intensional (TBox) and extensional (ABox) knowledge, typically written in a first-order logic fragment. DL knowledge bases have been extensively studied and they are typically provided with a semantics based on formal logic, which would overcome some of the problems of both

the previous approaches (e.g., KGs as property graphs and RDFS KGs). Nevertheless, real systems based on the classical DL approach, do not offer metamodeling capabilities, i.e., semantic aspects involving domain elements which can simultaneously play the role of class and property, or of class and individual. Such systems often adopt the so-called *Direct Semantics* making use of *punning*, according to which two occurrences of the same syntactic domain element in different positions (e.g, class position and individual position) are treated as if they refer to different elements. It has been observed that this does not constitute a proper form of metamodeling semantics, which has been proposed for OWL 2 QL ontologies in [17]. As already discussed in [18], we argue that a unifying semantic characterization of KGs is required in order to overcome the limits posed by preexisting approaches, while also studying possible extensions (e.g., adding forms of negations in both KGs and query language, or by means of epistemic queries).

3. Contribution and Approaches

We propose a semantic characterization of knowledge graphs based on meta-modeling. Our main goal is to define a general semantic framework for KGs based on formal logic, allowing to assign a well-founded meaning to any kind of KGs, and to provide the necessary theoretical foundations to the main reasoning problems, starting from the one of answering queries over KGs in the proposed semantic framework. The necessity of a higher-order modeling for KGs is motivated by the inherent characteristics of the data that is typically represented through KGs and that may require a higher than normal flexibility in terms of data modeling capabilities. In particular, a semantics based on meta-modeling allows to precisely capture the meaning of metaclasses and metaproperties, where a metaclass is a class having classes among its instances, and a metaproperty can represent relations holding between classes or properties instead of just individuals. In other words, a semantics based on metamodeling allows for elements to simultaneously play different roles, e.g., possibly acting at the same time as a class and an individual, or as a property and a class. As an example, one might want to represent a domain about Universities with classes *Associate Professor*, *Full Professor*, whose instances are associate professors and full professors, respectively, *Student*, i.e., the class of students, and *FacultyRole*, whose instances are specific faculty roles as *Associate Professor* and *Full Professor*. In this case *FacultyRoles* is a metaclass, in that its instances are classes themselves. At the same time, *Associate Professor* and *Full Professor* simultaneously play the role of classes and the role of individuals.

Some existing frameworks, like RDF(S) already support metamodeling capabilities, in that they explicitly allow elements to play different roles, e.g., that of a class and of a property, but their semantics is not necessarily based on formal logic, and when it comes to properly reason over KGs, answering queries might yield to unsound or incomplete results (as we show through an example in the next Section).

In our work, we define a metamodeling semantics for KGs based on the notion of interpretation, starting from a specific type of KGs, which we call RKG, and which is characterized by a vocabulary of pre-defined symbols with an assigned semantics which captures the basic relationship types, like ISA relationships between classes and between properties, typing constraints, and membership assertions. We also address the main reasoning tasks, starting from that of query entailment, for which we provided an algorithm and a complexity characterization of the problem.

4. Preliminary Results

Our work on the formalization of a framework for KGs based on metamodeling began from the definition of a restricted class of KGs, called RKGs, and characterized by a vocabulary of pre-defined symbols, namely `Resource`, `Class`, `Property`, `type`, `domain`, `range`, `subClassOf`, and `subPropertyOf`, which is a subset of the RDFS vocabulary which we argue represents a basic yet effective tool for KGs domain modeling. We then proceeded with a metamodeling formalization of RKGs based on the notion of interpretation. In particular, an interpretation $\mathcal{I}$ for an RKG G is a pair $\langle W, \cdot^{\mathcal{I}} \rangle$, where W is a

structure consisting of the tuple $\langle \Delta^{\mathcal{I}}, \cdot^{C}, \cdot^{P} \rangle$, with $\Delta^{\mathcal{I}}$ being the interpretation domain, and $\cdot^{C}$ and $\cdot^{P}$ two partial functions defining the possible extensions of domain elements as classes or properties, respectively.

Absence of a Universal Model

A deeper study of the logical properties of the defined framework led to a first interesting result: RKGs do not admit a universal model in the general case. This is a rather surprising result, since lightweight languages typically admit a canonical model which can be used for reasoning tasks, in that it can be view as a representative of all the models of the knowledge base. Such result has an interesting implication, which is illustrated by the following example.
Consider the RKG $G = \{\langle a\ R\ b\rangle, \langle b\ R\ a\rangle, \langle t\ \texttt{type}\ b\rangle, \langle a\ \texttt{type Class}\rangle\}$ and the Boolean query Q

```
ASK WHERE { x R y. z type y. x subClassOf b }
```

We are interested in the problem of query entailment, i.e., the problem of establishing whether the query Q is true in every model of G, denoted $G \models Q$. According to the semantics of SPARQL, answering the query requires to find a binding making the query true in every model of G. Intuitively, there exist two types of such bindings, namely those where $x \leftarrow a$ and $y \leftarrow b$, and those where $x \leftarrow b$ and $y \leftarrow a$. The bindings of the first type do not make the query true in every model, since there could be models where a is not a subclass of b, and thus the third atom of the query would not be true in these models. Similarly, the bindings of the second type can not make the query true in the models where “a" is an empty class; in these models the second atom of the query would be false. The conclusion would be that the query Q is not entailed by G. We argue that such a way of reasoning is not coherent with proper formal logic, in that in order for Q to be entailed by G, it is necessary that, for every model of G there exists a binding making the query true, which is different from the requirement that there exists a binding making the query true in every model. In light of this observation, one could conclude that Q is indeed entailed by G, since models can be divided into classes, for each of which there exists an assignment making Q true. For all the models where a is an empty class, Q is satisfied by the first type of binding, while for those models where a is not empty, the second type of bindings satisfies Q. We found that the absence of universal models is caused by specific forms of uncertainty that may appear in RKGs (as the one on the element “a" of the previous example, which can be either empty or non-empty in every model). We provided a classification of RKGs allowing to clearly distinguish those RKGs admitting a universal model from those that do not admit one.

Query Entailment over RKGs

The absence of a universal model working as a representative of all the models of G to be queried, required to redefine the task of query entailment through a procedure based on reasoning “by cases”. We studied the problem of query entailment in the new framework according to the metamodeling semantics we defined, and we provided an algorithm for properly solving the problem of query entailment for those RKGs requiring reasoning by cases, thus establishing an upperbound for the problem, which is Π_2^p in combined complexity and in AC^0 in data complexity. We also illustrated a polynomial time reduction from an instance of the NP-compelte problem of 3SAT to the complement of the query entailment problem, thus establishing a lowerbound of coNP in combined complexity and AC^0 in data complexity for the problem of query entailment over RKGs.

Negations in queries

Parallel to previous results, our research has also addressed the problem of answering conjunctive queries (CQs) containing specific forms of negation (such as *safe negation* and *inequality atoms*) over DL-Lite$^{\neg}_{RDFS}$, i.e., the DL counterpart of RDFS knowledge bases with the addition of disjunction. Previous results had shown that answering CQs with inequality atoms over DL-Lite$^{\neg}_{RDFS}$ knowledge bases is Π_2^p-complete in the general case, while NP-complete in combined complexity and and P-complete in data complexity, when CQs contain at most one inequality atom [19, 20]. We extended

the study and found that the presence of two inequality atoms or two safe negations (a safe negation is a negated atom whose terms also appear in at least one positive atom of the query) is enough to make the problem Π_2^p-complete in combined complexity and coNP-complete in data complexity. Related results also had an impact on the problem of query containment, in that we proved that the problem of query containment is Π_2^p-complete when the contained query contains at most one inequality atom, and the containing query contains at most two inequality atoms.

5. Future Work and Research Directions

The research described so far offers several possible goals to be pursued. In particular we aim at:

- proceeding in the direction of a general parametric formalization of a framework for knowledge graphs with a semantics based on formal logic, and including meta-modeling capabilities. Said general framework should be able to capture suitable constraints in order to provide proper formalizations for KGs of different types (e.g., plain graph databases or KGs based on DLs).
- capturing forms of incomplete information arising in the different variants of KGs, e.g., RDFS knowledge graphs, where some constructs (e.g., rdf:Statement) yield to forms of incompleteness when interpreted according to formal logic.
- extending both the modeling constructs of KGs and the query languages with additional and more powerful features, e.g., negation, inequality, integrity constraints and epistemic queries. Epistemic queries would allow to capture weak forms of reasoning and integrate them with more powerful forms of reasoning.

Acknowledgments

This work has been supported by MUR under the PNRR project FAIR (PE0000013).

This work has been carried out while the author was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome.

Declaration on Generative AI

(based on the activity taxonomy in ceur-ws.org/genai-tax.html)
During the preparation of this work, the author used Chat-GPT-4 and Google Gemini in order to: Grammar and spelling check, Paraphrase and reword, Formatting assistance. After using these tool(s)/service(s), the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] S. Tiwari, F. N. Al-Aswadi, D. Gaurav, Recent trends in knowledge graphs: theory and practice, Soft Computing 25 (2021) 8337–8355.
[2] P. Hitzler, A. Eberhart, M. Ebrahimi, M. K. Sarker, L. Zhou, Neuro-symbolic approaches in artificial intelligence, National Science Review 9 (2022).
[3] A. Oltramari, J. Francis, C. Henson, K. Ma, R. Wickramarachchi, Neuro-symbolic architectures for context understanding, in: Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges, IOS Press, 2020, pp. 143–160.
[4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 1–42.

[5] European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), Technical Report, European Union, 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj, accessed: 2025-03-09.
[6] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 2 (2023).
[7] S. Wang, W. Fan, Y. Feng, X. Ma, S. Wang, D. Yin, Knowledge graph retrieval-augmented generation for llm-based recommendation, arXiv preprint arXiv:2501.02226 (2025).
[8] F. N. Stokman, P. H. de Vries, Structuring knowledge in a graph, in: Human-Computer Interaction: Psychonomic Aspects, Springer, 1988, pp. 186–206.
[9] N. Francis, A. Green, P. Guagliardo, L. Libkin, T. Lindaaker, V. Marsault, S. Plantikow, M. Rydberg, P. Selmer, A. Taylor, Cypher: An evolving query language for property graphs, in: Proceedings of the 2018 international conference on management of data, 2018, pp. 1433–1445.
[10] W. Ali, M. Saleem, B. Yao, A. Hogan, A.-C. N. Ngomo, A survey of rdf stores & sparql engines for querying knowledge graphs, The VLDB Journal (2022) 1–26.
[11] F. Baader, I. Horrocks, C. Lutz, U. Sattler, An Introduction to Description Logic, Cambridge University Press, 2017.
[12] M. Wang, L. Qiu, X. Wang, A survey on knowledge graph embeddings for link prediction, Symmetry 13 (2021) 485. URL: https://doi.org/10.3390/sym13030485. doi:10.3390/SYM13030485.
[13] A. Rossi, D. Barbosa, D. Firmani, A. Matinata, P. Merialdo, Knowledge graph embedding for link prediction: A comparative analysis, ACM Trans. Knowl. Discov. Data 15 (2021). URL: https://doi.org/10.1145/3424672. doi:10.1145/3424672.
[14] R. Angles, The property graph database model., in: AMW, 2018.
[15] W3C, RDF schema 1.1, https://www.w3.org/TR/rdf-schema/, 2014.
[16] H. J. ter Horst, Completeness, decidability and complexity of entailment for rdf schema and a semantic extension involving the owl vocabulary, Web Semant. 3 (2005) 79–115. URL: https://doi.org/10.1016/j.websem.2005.06.001. doi:10.1016/j.websem.2005.06.001.
[17] M. Lenzerini, L. Lepore, A. Poggi, Metamodeling and metaquerying in owl 2 ql, Artificial Intelligence 292 (2021) 103432.
[18] R. M. Delfino, M. Lenzerini, A. Poggi, On the need of a formal meta-modeling semantics for knowledge graphs, in: Proceedings of the 23rd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024), CEUR-WS.org, Bolzano, Italy, 2024. URL: https://ceur-ws.org/Vol-3915/Paper-10.pdf.
[19] V. Gutiérrez-Basulto, Y. Ibáñez-García, R. Kontchakov, E. V. Kostylev, Queries with negation and inequalities over lightweight ontologies, Journal of Web Semantics 35 (2015) 184–202.
[20] G. Cima, M. Lenzerini, A. Poggi, Answering conjunctive queries with inequalities in dl-lite$_r$, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 2782–2789.