Towards Semantic Interpretation of Structured Data Sources in Privacy-Preserving Environments

Christina Karalka*, Georgios Meditskos and Nick Bassiliades

School of Informatics, Aristotle University of Thessaloniki, 54124, Greece

Abstract

As the use of sensitive data becomes increasingly prevalent, it is essential to ensure that privacy preserving technologies are effectively utilized to protect such data. Relational databases are commonly used for data storage, but they may not provide sufficient insights for identifying privacy vulnerabilities. Moreover, due to the various legal and technical terms, as well as the various actors involved, it is difficult to decide the privacy-preserving technology and the type of the configuration needed for a specific dataset. This short paper presents work in progress towards adding a semantic layer on top of structured data sources for efficient and intelligent use of data in privacy-preserving scenarios. More specifically, we present key research directions for the development of SemCrypt, a novel framework for schema-enrichment through semantic annotations and mappings to Knowledge Bases and domain ontologies so as to: a) interlink and contextually enrich schemata and data in an interoperable manner; b) use the underlying semantics to assist data owners in assessing privacy preserving technologies depending on the sensitivity of data in different use cases, such as in health, finance and cyber threat intelligence.

Keywords

knowledge graphs, ontologies, data sources, semantic interpretation, privacy preservation

1. Introduction

As data generation grows exponentially, it has become imperative to process it in a privacy-preserving manner, especially when dealing with sensitive personal information such as medical records or financial data. Despite relational databases (RDBs) being a crucial component of such information systems, they may not provide sufficient context for determining the appropriate privacy-preserving strategy, given the different actors involved and the lack of legal and technical terms. Therefore, more sophisticated data models are needed to ensure that such technologies are applied effectively.

Semantic lifting refers to the transformation of data into highly flexible and semantically rich representations, namely ontologies and knowledge graphs. It enables the extraction of the implicit meaning and relations between entities, which is critical for understanding data sensitivity and the risks of sharing it, but would otherwise remain concealed in traditional databases. Furthermore, a shared understanding of data can be established by reusing existing ontologies that define a common vocabulary of domain-specific concepts.

Fourth International Workshop On Knowledge Graph Construction

*Corresponding author.

kchristi@csd.auth.gr (C. Karalka); gmeditsk@csd.auth.gr (G. Meditskos); nbassili@csd.auth.gr (N. Bassiliades)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

Despite these advantages, mapping a RDB to a knowledge graph is not a straightforward task due to the inherit structural differences. Additionally, the automated identification of common concepts is hindered by the anticipated knowledge gap, the use of vague schema annotations and the great number of highly correlated attributes numerical attributes compared to categorical ones. Furthermore, access to data might be limited during the development process due to privacy concerns.

The aim of this short paper is to present work in progress towards addressing these challenges and assisting data owners in assessing the level of privacy required for their data. Specifically, we propose to semantically enrich the schema of a given relational database by mapping it to knowledge graphs and domain ontologies, while also developing a declarative framework to better understand the hidden relationships between entities and the sensitivity of individual attributes. It is expected that the utilization of semantic web technologies in this context will facilitate interoperability for privacy-preserving processing of similar data types across organizations, while it will also enhance the explainability of generated recommendations. SemCrypt is part of the ENCRYPT platform[1], introducing a layer of semantic abstraction to the underlying data facilitating intelligent management of information and decision support.

The rest of the paper is organised as follows: Section 2 presents related work on the domain of data annotation. Section 3 gives an overview of the framework and presents our motivation. In Section 4 we describe the basic concepts of the proposed framework, while in Section 5 we conclude our work, presenting next steps.

2. Related work

Annotating tabular data with semantic metadata from pre-existing knowledge graphs (KGs) and ontologies has gained prominence in research, with the SemTab challenge being a notable example. Specifically, table cells are identified as KG instances (CEA), columns as classes (CTA), and column pair relations as properties (CPA). Solutions generally follow a standard pipeline that includes data preprocessing, candidate generation and disambiguation [1]. These methods mostly rely on heuristics [2, 3] and their performance is linked with the compatibility of the input data with the KG [4]. In addition, learning-based approaches render more resilient to noise. Deep learning-based systems employ pretrained language models such as Word2Vec [5] and BERT [6, 7]. As word embeddings also capture semantic intricacies, word similarity is better reflected.

Incorporating contextual information also leads to more accurate semantic annotation of table elements. To leverage intra-column relations in CEA, [8] applied a TF-IDF scoring function based on characteristics of column cell candidates. [5] employed convolution networks on the pre-trained word embeddings of cell values in CTA. To also capture inter-column context, [6] applied BERT on a multi-column serialized form of the input table. However, single-table mapping approaches are not directly applicable to RDBs, as they consist of multiple tables of different types with complex interrelationships.

Training data scarcity is a common issue in real-world applications. Therefore, instead of leveraging an annotated dataset, [7] fine-tuned BERT for identifying equivalent classes using

[1]https://encrypt-project.eu/

sets of synonym and non-synonym pairs, generated based on the given ontologies and same-domain auxiliary ontologies. However, the utilization of contextual information is limited in [7]. Meanwhile, [9] trained a GNN model on the underlying ontology graph to enhance pretrained word embeddings with structural information.

3. Key Concepts and Motivation

Currently, there is no universally accepted method for securely exchanging business data that upholds the principles of traceability, data owner privacy, and data sovereignty. Business information is often communicated using various standards and formats, and is frequently isolated in organizational silos, which makes it difficult to extract reusable knowledge and promotes local collaboration only. ENCRYPT is a Horizon Europe project that aims to provide a scalable, practical, adaptable privacy-preserving framework facilitating the GDPR-compliant processing of such data stored in federated cross-border data spaces. To this end, it develops an intelligent and user-centric platform for the confidential processing of privacy-sensitive data via configurable, optimizable, and verifiable privacy preserving techniques, in three domains: a) Health (management of medical records in an oncology department); b) Cyber theat intelligence (server and database logs); c) Fintech (dept collection services).

SemCrypt aims to enrich ENCRYPT with a semantic annotation layer. From one hand, SemCrypt acts as the semantic middleware, capturing, interlinking and semantically enriching schemata and data. On the other hand, it develops a declarative framework on top of the enriched schemata to identify situations, i.e. entities and relations of interest, so as to achieve privacy awareness and foster personalised suggestions on privacy preserving technologies depending on the sensitivity of data. All in all, SemCrypt aims to address the following challenges:

Knowledge gap - The concepts defined in distinct data models might overlap but a complete alignment cannot be expected [10]. This is particularly true for domain-specific databases where the entities may not have already been modeled in existing ontologies. Therefore, it might not be feasible to obtain accurate and complete semantic annotation of the input sources when the process solely relies on the existence and compatibility with domain ontologies. Leveraging encyclopedic KGs, such as DBpedia, may alleviate this issue, however the vast scope of such sources might increase ambiguity and complexity of the mapping process [4].

Data privacy concerns - While the schema of a database might be available during development, its contents may not be accessible due to confidentiality or privacy concerns. Therefore, instead of also utilizing the values of attributes to infer a richer class taxonomy when generating an ontology from the given database, the mapping must be performed solely based on its schema [11]. This can limit the disambiguation ability of the framework and can lead to incomplete or incorrect mappings. The use of synthetic data is a possible step towards resolving this issue. However, the generation of datasets that reflect the characteristics of real data is a time-consuming process.

Annotation of ambiguous data sources - The use of generic and non-descriptive naming conventions for schema elements can complicate the mapping process. While there are some generally accepted naming conventions, no standardized set have been established and different organizations can follow their own preferred conventions. However, when the matching of an

ontology term and a database field related to the ambiguous element has been established, it may be possible to identify the proper annotation among the entities and properties related to this ontology term. Additionally, the presence of highly correlated numerical fields is a common occurrence in RDBs. Defining mappings that capture the nuances of fields with closely related semantics is essential for disambiguation.

Domain independence - While the proposed mapping framework is initially being developed with specific use cases in mind, it has the potential to be extended to other domains where privacy-preserving computations are needed. Achieving domain independence requires a universal representation layer that can facilitate data exchange and support decision-making processes across various sectors. The framework's interoperability can be enhanced by utilizing ontology embedding methods [12] to detect similar entity roles and patterns among different domain ontologies. Based on these observations, a layer of abstract concepts can be extracted.

4. Methodology

SemCrypt involves the generation of the ENCRYPT Knowledge Graphs (EKG) by mapping, integrating, and correlating information and data at various levels of granularity. The EKG will encapsulate hierarchies of concepts, reusing existing domain ontologies whenever possible, such as the Financial Industry Business Ontology (FIBO) in the fintech domain, to facilitate the semantic lifting of the input data and extract semantic dependencies among resources. This will allow for the exchange of standardized cybersecurity intelligence, knowledge, data types, and models across different organizations. In addition, tools will be developed to enhance the EKG by deriving additional entities and relationships and semantically enriching the EKG to enable advanced decision-making, recommendations, and justifications.

The conceptual architecture of the proposed approach is illustrated in Figure 1. More specifically:

- The input database and the target ontology are transformed into graphs and random walks are generated for each ontology entity.
- The matching score of an ontology/database term pair in the context of the RWs is assessed by a BERT-based model and candidates are selected through inexact graph matching.
- A SPARQL/SHACL rule set will be employed to identify privacy vulnerabilities and patterns and recommend solutions.

In the following we present details for each component of the architecture.

4.1. Relational database and ontology graphs

Our aim is to establish semantic correspondences between the concepts of the given RDB and the target domain ontology by associating the entities, attributes, and relationships described in the schema with the classes and properties of the ontology. However, the complexity of RDBs lies in the interdependencies between tables and the distribution of entity characteristics across multiple tables. To mitigate this issue, we propose to transform the RDB schema into a graph structure, or putative ontology G_{RDB} following the direct mapping guidelines established by

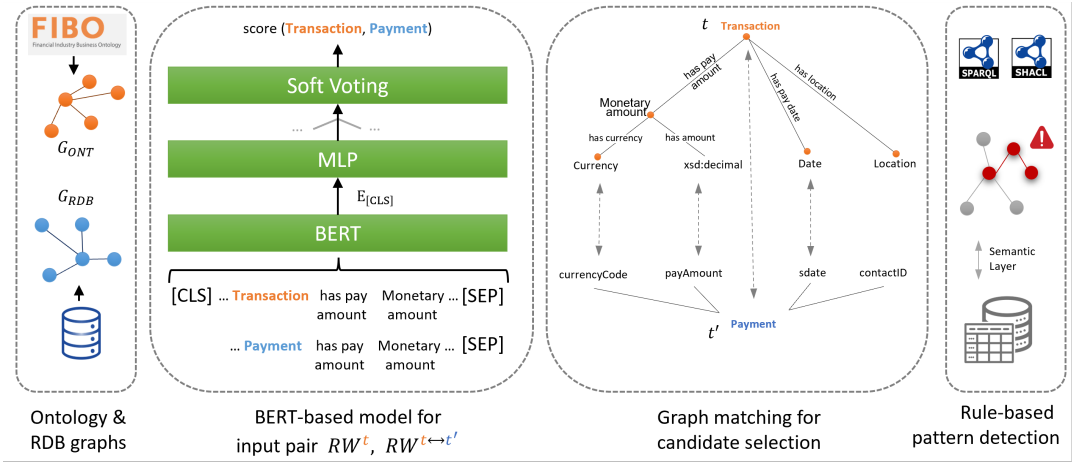

Figure 1: An overview of the main components of our proposed architecture.

W3C [13]. This conversion will provide a more concise and unified view of the RDB's structure, allowing for easier identification of entities and their relationships. Moreover, the flexibility of this structure permits the inclusion of taxonomical relations derived from the database's contents, if available [11].

Similarly, a graph can be derived from the ontology, based on associations between entity pairs established by object properties, hierarchies, restrictions or axioms, and reasoning. Despite the intuitive correlation between SQL constrains and property restrictions, such as *NOT NULL* being equivalent to *owl:minCardinality 0*, considering constraints or qualifiers during the mapping process might lead to confusion due to design discrepancies. Therefore, an ontology graph G_{ONT} is simply defined as a heterogeneous graph of entity nodes connected by ontology properties.

4.2. Capturing semantic context with random walks

The task of identifying equivalent concepts is formulated as the distinction between synonym and non-synonym terms or phrases. The underlying assumption is that a pair of synonyms can be used interchangeably in a sentence without significantly altering its meaning. In the context of ontologies, this implies that the label of a term can be replaced by the label of an equivalent concept in an RDF triple without diluting its semantics. Therefore, the equivalence between a pair of items from G_{RDB} and G_{ONT} should be determined according to the context provided by the ontology graph.

To capture the context of an item in G_{ONT}, we leverage its relations with neighboring entities and properties. Following the RDF2Vec [14] approach for embedding RDF graphs, we propose to employ random walks in G_{ONT}. Specifically, for each item in the graph, we generate a set of fixed-length sequences of entities and properties, where the sequence length is denoted by l. At its simplest form, l can be set to 1, resulting in single-resource sequences where the task is reduced to synonym pair classification as proposed in [7]. Alternatively, l can be set to 3 to extract RDF statements as sequences.

4.3. BERT-based model for semantic annotation in an unsupervised setting

Given a database term t', an ontology term t and a set of random walks RW^t containing t, a set of modified sequences $RW^{t \leftrightarrow t'}$ is generated by replacing t with t'. The semantic similarity between the sentence pairs can then be used to assess if t is a candidate equivalent term for t'. However, a potential challenge arises due to the different naming conventions used by domain ontologies and RDBs. For example, despite the FIBO class "Financial Instrument" being a suitable match for the "Receipt" concept described in the input RDB, there is no lexical similarity between the two terms.

Therefore, the central component of our proposed framework constitutes a BERT-based binary classification model. Specifically, word embeddings are generated for each input pair to capture conceptual similarities between different terms as well as the context provided by the random walk sequences. Subsequently, a downstream MLP module is applied on these representations. Finally, the confidence of the matching can be determined by soft voting considering a set of sequence pairs for t and t'.

Leveraging an extensively pretrained language model facilitates the applicability of our framework in real world scenarios where the availability of high-quality training data is limited. Following [15, 7], our model will be finetuned according to synonym and non-synonym terms derived from external knowledge sources and thesauri, such as WordNet. Additionally, to address the widespread use of abbreviations and acronyms in RDBs, such training samples will also be obtained from appropriate resources [16]. Finally, the performance of domain-specific variations of BERT, such as BioBERT and FinBERT for the health and fintech domains, respectively, will also be examined.

4.4. Graph matching for candidate selection

Examining all possible term pairs of G_{RDB} and G_{ONT} is computationally expensive. Under the assumption that there might not be any lexical overlap between the matched pairs, we instead rely on the most central entities in G_{RDB} to guide the search. Therefore, the main concepts of the RDB, which are usually the entities represented by tables in the schema, are identified. Subsequently, a set of candidate ontology entities with the top-K confidence score is generated using the model, since the optimal match might not always be the most proximate in the embedding space [17].

Having limited the search space in subgraphs around the top-K candidates, the next step is to evaluate the correspondence between related concepts to select the optimal mapping. This can be achieved by employing an inexact graph matching technique to identify matches between entities and relations connected with t and t', that belong to subgraphs $G^{t'}_{RDB}$ and G^{t}_{ONT}, respectively. Random walks containing the candidate t and the term in question can be used to define the modification cost according to the confidence score predicted by the BERT-based model. This process results in aligning the central entities of the database, along with their related concepts, to the highly relevant ontology terms.

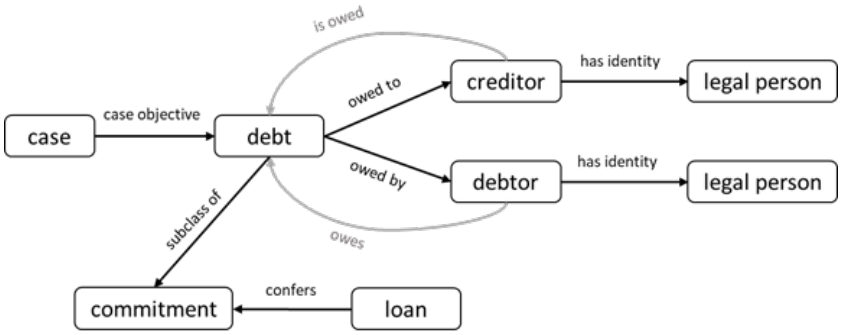

Figure 2: Deptor-creditor pattern in FIBO.

4.5. Enhancing data privacy through semantic intelligence

The semantic interpretation of data sources can assist in the detection of identifying variables and the assessment of privacy risks, ultimately enhancing the robustness of data against re-identification attacks. Specifically, by disambiguating the different types of attributes using well-defined classes and properties, ontologies can aid the selection of appropriate de-identification techniques. For instance, dates can be obscured through noise injection (perturbation), while personal names might be completely removed (redaction). Additionally, taxonomy relations can be used to increase the abstraction of rarely occurring values (generalization hierarchy) [18].

Furthermore, the detection of indirect identifiers is achievable through the collaboration of automated techniques and domain experts. At the level of data, graph analysis techniques could leverage the underlying graph structure of the KG to identify privacy vulnerabilities in the form of outliers, such as a unique edge of a particular type.

Simultaneously, domain knowledge is encoded in a predetermined rule set with the aim of effectively combining, associating and interpreting the asserted information in the graphs to gain insight into the context and identify privacy-related issues. Our knowledge-driven approach enables the system to offer suggestions on the implementation and setup of privacy-preserving technologies according to the type of data they intend to process. For example, the existence of an indirect link between a creditor and a deptor usually raise privacy concerns and should be reported to the data owner.

Our solution is implemented using SPARQL and executes a set of CONSTRUCT graph patterns to detect problematic situations. To facilitate interoperability, the SPARQL graph patterns are defined as SHACL Rules on top of domain ontologies, such as in FIBO (Figure 2), capitalising on the results of semantic annotation described in the previous sections.

Finally, the use of encrypted data can pose significant challenges for precise and effective predictions [18], particularly in fields such as healthcare where accuracy is crucial. Semantic annotation can assist in balancing privacy maintenance and accuracy, by categorizing and identifying the criticality of each attribute in sensitive domains.

5. Conclusions and Next Steps

Being able to semantically understand the structure of data is significant for many tasks, such as knowledge discovery and integration. In this paper we presented preliminary work on the development of a knowledge-driven framework aiming at building semantically rich and interlinked knowledge graphs to capture the semantics of structured data. We presented the motivation of our work, along with the key challenges that we aim to address, and we elaborated on the key research directions we are currently investigating. The solution is combined with a declarative framework that identifies patterns in the graphs relevant to privacy-preserving requirements, so as to assist data owners in taking decisions about the confidential processing of privacy-sensitive data. Next steps involve the finalisation of the implementation and the experimental evaluation of the solution.

Acknowledgments

This research has received funding from the European Union (HORIZON-RIA ENCRYPT - GA No. 101070670).

References

[1] A. Dimou, D. Chaves-Fraga, Declarative description of knowledge graphs construction automation: Status & challenges, in: Proceedings of the 3rd International Workshop on Knowledge Graph Construction (KGCW 2022) co-located with 19th Extended Semantic Web Conference (ESWC 2022), volume 3141, 2022.

[2] X. Li, S. Wang, W. Zhou, G. Zhang, C. Jiang, T. Hong, P. Wang, Kgcode-tab results for semtab 2022, Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab), CEUR-WS. org (2022).

[3] N. Abdelmageed, S. Schindler, Jentab: A toolkit for semantic table annotations, in: Second International Workshop on Knowledge Graph Construction, 2021.

[4] J. Liu, Y. Chabot, R. Troncy, V.-P. Huynh, T. Labbé, P. Monnin, From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods, Journal of Web Semantics (2022) 100761.

[5] J. Chen, E. Jiménez-Ruiz, I. Horrocks, C. Sutton, Colnet: Embedding the semantics of web tables for column type prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 29–36.

[6] Y. Suhara, J. Li, Y. Li, D. Zhang, Ç. Demiralp, C. Chen, W.-C. Tan, Annotating columns with pre-trained language models, in: Proceedings of the 2022 International Conference on Management of Data, 2022, pp. 1493–1503.

[7] J. Chakraborty, H. M. Zahera, M. A. Sherif, S. K. Bansal, Ontoconnect: domain-agnostic ontology alignment using graph embedding with negative sampling, in: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2021, pp. 942–945.

[8] A. Thawani, M. Hu, E. Hu, H. Zafar, N. T. Divvala, A. Singh, E. Qasemi, P. A. Szekely,

J. Pujara, Entity linking to knowledge graphs to infer column types and properties., SemTab@ ISWC 2019 (2019) 25–32.

[9] J. Chakraborty, H. M. Zahera, M. A. Sherif, S. K. Bansal, Ontoconnect: domain-agnostic ontology alignment using graph embedding with negative sampling, in: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2021, pp. 942–945.

[10] D.-E. Spanos, P. Stavrou, N. Mitrou, Bringing relational databases into the semantic web: A survey, Semantic Web 3 (2012) 169–209.

[11] Á. Sicilia Gómez, et al., Supporting Tools for Automated Generation and Visual Editing of Relational-to-Ontology Mappings, Ph.D. thesis, Universitat Ramon Llull, 2016.

[12] J. Chen, P. Hu, E. Jimenez-Ruiz, O. M. Holter, D. Antonyrajah, I. Horrocks, Owl2vec*: Embedding of owl ontologies, Machine Learning 110 (2021) 1813–1845.

[13] W. W. W. C. (W3C), A direct mapping of relational data to rdf, https://www.w3.org/TR/rdb-direct-mapping/, 2012.

[14] P. Ristoski, H. Paulheim, Rdf2vec: Rdf graph embeddings for data mining, in: The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part I 15, Springer, 2016, pp. 498–514.

[15] P. Kolyvakis, A. Kalousis, D. Kiritsis, Deepalignment: Unsupervised ontology matching with refined word vectors, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 787–798.

[16] Allacronyms, https://www.allacronyms.com/, 2005.

[17] V. Mijalcheva, A. Davcheva, S. Gramatikov, M. Jovanovik, D. Trajanov, R. Stojanov, Learning robust food ontology alignment, in: 2022 IEEE International Conference on Big Data (Big Data), IEEE, 2022, pp. 4097–4104.

[18] E. Purificato, S. Wehnert, E. W. De Luca, Dynamic privacy-preserving recommendations on academic graph data, Computers 10 (2021) 107.