

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, Thessaloniki, Greece

Abstract

As the use of sensitive data becomes increasingly prevalent, it is essential to ensure that privacy preserving technologies are effectively utilized. Although relational databases are commonly used for data storage, they may not provide sufficient insights for identifying privacy vulnerabilities. Moreover, the complexity introduced by multiple actors, legal and technical terms poses a challenge in determining the appropriate privacy-preserving configuration for a specific dataset. This paper presents ongoing work 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 stakeholders 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 sensitive information such as medical and financial records in a privacy-preserving manner. Despite relational databases (RDBs) being a crucial component of such information systems, they may not provide sufficient context for determining an appropriate privacy-preserving strategy due to the lack of legal and technical terms and the involvement of multiple actors. Therefore, more sophisticated data models are needed to ensure effective application of such technologies.

Semantic lifting refers to associating the elements of a data source with semantic metadata [1]. 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 common, domain-specific vocabularies.

Despite these advantages, mapping a RDB to a knowledge graph is not a straightforward task due to the inherent structural differences. Additionally, the automated identification of common

KGCW'23: 4th International Workshop on Knowledge Graph Construction, May 28, 2023, Crete, GRE

*Corresponding author.

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

🆔 0000-0002-6451-775X (C. Karalka); 0000-0003-4242-5245 (G. Meditskos); 0000-0001-6035-1038 (N. Bassiliades)



© 2023 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)

concepts is hindered by the anticipated knowledge gap, the use of vague schema annotations and the great number of closely related numerical attributes compared to categorical ones. Furthermore, access to data might be limited during development due to privacy concerns.

This paper presents the SemCrypt framework, an ongoing effort towards introducing a semantic abstraction layer to the underlying data and assisting data stakeholders in assessing the necessary level of privacy in the ENCRYPT platform¹. Specifically, we propose to semantically enrich the schema of a given relational database by mapping it to domain ontologies, while also developing a declarative framework to uncover hidden relationships between entities and the sensitivity of individual attributes.

2. Related work

Annotating tabular data with semantic metadata from existing knowledge graphs (KGs) and ontologies has gained prominence in research, exemplified by the SemTab challenge². Specifically, table cells are identified as KG instances, columns as classes, and column pair relations as properties. Solutions generally follow a standard pipeline that includes data preprocessing, candidate generation and disambiguation [2]. These methods mostly rely on heuristics [3, 4] and their performance is linked with the compatibility of the input data with the KG [5]. Contrarily, learning-based approaches provide more resilience to noise. Deep learning-based systems employ pretrained language models such as Word2Vec [6] and BERT [7, 8]. As word embeddings also capture semantic intricacies, word similarity is better reflected.

Incorporating contextual information improves the accuracy of semantic annotation of table elements. Chen et al.[6] employed convolution networks on the pre-trained word embeddings of cell values for column annotation. To also capture inter-column context, Suhara et al.[7] 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. Instead, rule-based systems are employed in [9, 10] to extract a new "putative" ontology from the schemata and contents of RDBs. Subsequently, ontology alignment is performed to find correspondences with pre-existing ontologies.

Finally, training data scarcity is a common issue in real-world applications. Therefore, instead of leveraging annotated datasets, BERT is fine-tuned in [8] for identifying equivalent classes using 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.

3. Key Concepts and Motivation

Securely exchanging business data while maintaining traceability and sovereignty remains a challenge due to the lack of universally accepted standards, which restricts collaboration and knowledge exchange beyond local boundaries. The Horizon Europe project ENCRYPT aims to develop a scalable, user-centric platform for cross-border, GDPR-compliant³ processing of

¹<https://encrypt-project.eu/>

²<https://www.cs.ox.ac.uk/isg/challenges/sem-tab/>

³<https://gdpr-info.eu/>

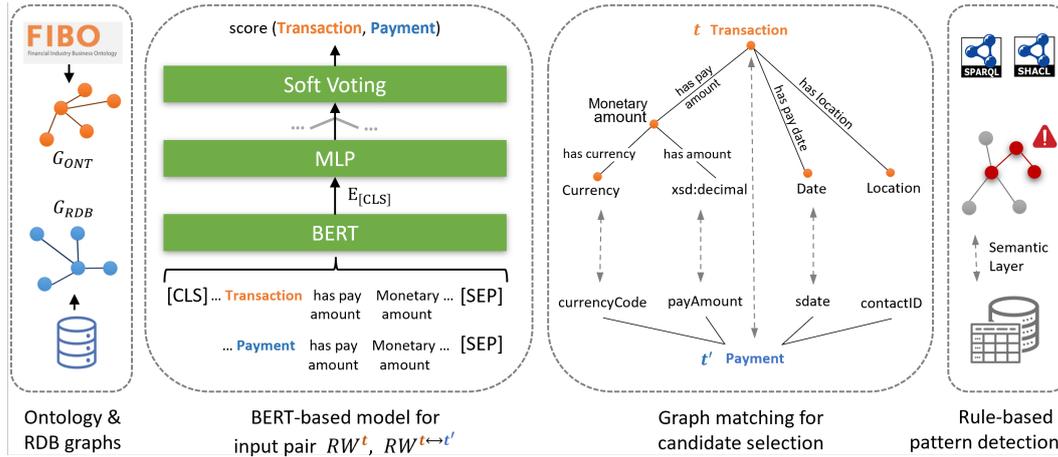


Figure 1: An overview of the components of our proposed conceptual architecture.

privacy-sensitive data in three domains: a) Health (medical records of a cardiology department); b) Cyber threat intelligence (server and database logs); c) Fintech (debt collection services).

SemCrypt strives to enrich ENCRYPT with a semantic annotation layer and facilitate the standardization of data and intelligence exchange across different organizations. Acting as a semantic middleware, it involves the generation of the ENCRYPT KGs by integrating and correlating data at various levels of granularity. The semantic lifting of the input data is achieved by reusing existing domain ontologies whenever possible, such as the Financial Industry Business Ontology (FIBO) in the Fintech case. Additionally, through the development of a declarative framework, privacy vulnerabilities and patterns are identified, so as to foster personalised suggestions on privacy preserving technologies depending on the sensitivity of data.

4. Methodology

4.1. Relational database and ontology graph generation

Semantic correspondences between the input RDB and the domain ontology can be established 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 simplify the structure, the RDB schema is transformed into a putative ontology graph G_{RDB} , following the direct mapping guidelines by W3C [11]. This conversion provides 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 representation permits the inclusion of taxonomical relations derived from the database's contents, if available [10].

Similarly, the domain ontology is expressed as a heterogeneous graph G_{ONT} where entity nodes are associated according to properties, hierarchies and restrictions. Despite the intuitive correlation between SQL constraints and OWL restrictions, such as *NOT NULL* being equivalent

to *owl:minCardinality 0*, such quantifiers are not incorporated into the graphs, as they might lead to confusion during the mapping due to design discrepancies.

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, a term label 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 each item in G_{ONT} we employ random walks, following the RDF2Vec [12] approach for embedding RDF graphs. Specifically, we generate a set of fixed-length sequences of entities and properties, where the sequence length is denoted by l . Setting l to 1 results in single-resource sequences and the task is reduced to synonym pair classification as proposed in [8]. 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 from terminology differences. 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 contextual information from 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' .

We aim to overcome the lack of high-quality training data in real world scenarios by utilizing an extensively pretrained language model. Following [13, 8], finetuning is performed according to synonym and non-synonym pairs derived from external knowledge sources, such as WordNet. Additionally, the widespread use of abbreviations and acronyms in RDBs is addressed by obtaining such training samples from appropriate thesauri [14]. Finally, domain-specific variations of BERT, such as BioBERT for the health use case, will also be evaluated.

4.4. Graph matching for candidate selection

Examining all possible pairs of G_{RDB} and G_{ONT} is computationally expensive. Given the possibility of the matched pairs having no lexical overlap, the search is instead guided by the most central entities in G_{RDB} , usually represented by tables in the initial schema. 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 [15].

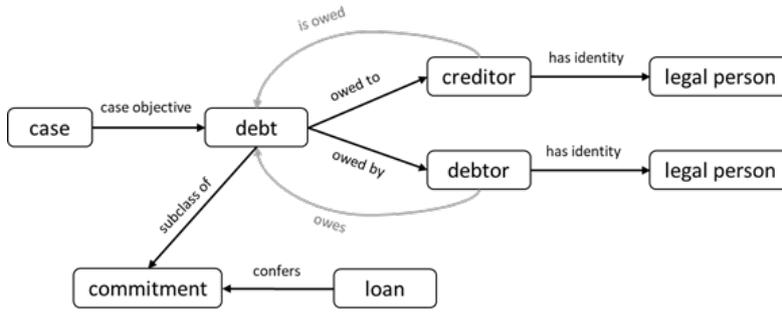


Figure 2: Debtor-creditor pattern in FIBO.

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

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 attributes using well-defined classes and properties, ontologies can aid the selection of appropriate de-identification techniques. For instance, dates are obscured through noise injection (perturbation), while personal names are completely removed (redaction). Additionally, taxonomy relations are used to increase the abstraction of rarely occurring values (generalization hierarchy) [16].

Privacy vulnerabilities can be identified through the collaboration of automated techniques and domain experts. At the data level, graph analysis techniques could leverage the underlying graph structure of the KG to infer indirect identifiers in the form of outlier combinations. Simultaneously, domain knowledge is encoded in a predetermined rule set with the aim of effectively combining, associating and interpreting the asserted information in the KGs to gain insight into the context and identify privacy-related issues. This 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 debtor could raise privacy concerns when combined with other information and should be reported to the data owner (Figure 2).

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, capitalising on the results of semantic annotation described in the previous sections.

5. Open questions and next steps

This section presents several key challenges and potential future research directions.

Knowledge gap - The concepts defined in distinct data models may overlap but a complete alignment cannot be expected [17]. This is particularly true for domain-specific databases where entities may not have been previously modeled in existing domain ontologies. Leveraging encyclopedic KGs, such as DBpedia⁴, can result in more complete annotations, but their extensive scope also increases the ambiguity and complexity during the mapping process [5].

Data privacy concerns - While the schema of a RDB might be available during development, its contents may not be due to confidentiality. Therefore, instead of also enriching the class taxonomy of the putative ontology with attribute values, the mapping must be performed solely based on the schema [10]. This can limit the disambiguation ability of the framework, thus leading to incorrect mappings. While synthetic data is a possible alternative, the generation of realistic datasets is a time-consuming process.

Annotation of ambiguous data sources - Generic or non-descriptive annotation of schema elements can complicate the mapping. Since no generally accepted standard has been established, different organizations can follow their own preferred naming conventions. However, by matching an ontology term and a database field related to the ambiguous element, a proper annotation may be identified among the entities and properties related to this ontology term. Additionally, as RDB's commonly consist of numerical fields with closely related semantics, mappings that capture their semantic nuances are essential for disambiguation.

Domain independence - In addition to the presented use cases, the proposed mapping framework has potential for extension to other domains where privacy-preserving computations are needed. Achieving domain independence requires a universal representation layer of abstract concepts to facilitate data exchange and decision-making processes across sectors. Interoperability could be enhanced by identifying similar entity roles and patterns across ontologies of different domains.

Acknowledgments

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 101070670.

References

- [1] A. Azzini, C. Braghin, E. Damiani, F. Zavatarelli, et al., Using semantic lifting for improving process mining: a data loss prevention system case study., in: SIMPDA, Citeseer, 2013, pp. 62–73.
- [2] 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.

⁴<https://www.dbpedia.org/>

- [3] 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).
- [4] N. Abdelmageed, S. Schindler, Jentab: A toolkit for semantic table annotations, in: Second International Workshop on Knowledge Graph Construction, 2021.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] J. Chakraborty, H. M. Zahera, M. A. Sherif, S. K. Bansal, Ontoconnect: domain-agnostic ontology alignment using graph embedding with negative sampling, in: *20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2021, pp. 942–945.
- [9] E. Jiménez-Ruiz, E. Kharlamov, D. Zheleznyakov, I. Horrocks, C. Pinkel, M. G. Skjæveland, E. Thorstensen, J. Mora, Bootox: Practical mapping of rdbs to owl 2, in: *The Semantic Web-ISWC 2015: 14th International Semantic Web Conference*, Bethlehem, PA, USA, October 11-15, 2015, *Proceedings, Part II 14*, Springer, 2015, pp. 113–132.
- [10] 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.
- [11] M. e. a. Arenas, A direct mapping of relational data to rdf, <https://www.w3.org/TR/rdb-direct-mapping/>, 2012.
- [12] 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.
- [13] 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.
- [14] Allacronyms, <https://www.allacronyms.com/>, 2005.
- [15] 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.
- [16] E. Purificato, S. Wehnert, E. W. De Luca, Dynamic privacy-preserving recommendations on academic graph data, *Computers* 10 (2021) 107.
- [17] D.-E. Spanos, P. Stavrou, N. Mitrou, Bringing relational databases into the semantic web: A survey, *Semantic Web* 3 (2012) 169–209.