An Evolving Data Space: A framework for unifying domain knowledge and data

Pedro Guimarães[1,3*], António C. Vieira[2,3] and Maribel Y. Santos[2,3]

[1*]CCG/ZGDV, Portugal.
[2]Information Systems Department, University of Minho, Portugal.
[3]ALGORITMI Center, University of Minho, Portugal.

*Corresponding author(s). E-mail(s): pedro.guimaraes@ccg.pt;
Contributing authors: avieira@dsi.uminho.pt; maribel@dsi.uminho.pt;

Abstract

Semantic integration across diverse data sources is a major challenge. This paper proposes a framework leveraging Data Spaces and International Data Spaces (IDS) connectors to develop an evolving domain reference model. Our approach proposes the use of a continuously enriched knowledge graph (KG), combined with intelligent unification techniques like knowledge completion and alignment. The framework integrates heterogeneous data using knowledge and data wrappers, extractors, and mappings, resulting in an updated domain reference model that adapts to new information. This approach not only supports the semantic integration process but also enhances the discovery of hidden relationships and insights within the knowledge and data, contributing to a comprehensive and evolving understanding of domain-specific knowledge.

Keywords: Data Spaces, Knowledge Graphs, Semantic Integration

1 Introduction

Reference models, often referred to as "universal models, generic models, or model patterns" [1], are crucial for understanding domain-specific structures. Gray and Rumpe [2] expanded this concept by introducing the importance of the expert and their domain knowledge, defining a reference model as a domain-specific structure that clearly expresses the expertise produced by an expert or body of experts.

In the context of Data Spaces and IDS connectors, these reference models may support the semantic integration of knowledge and data. Data Spaces emphasize the sovereign and sharing of data across various domains, requiring robust semantic integration solutions. A domain reference model represents a set of models within the same domain and its multi-relational contexts, making a graph-structured data model suitable for representing domain knowledge.

Due to this, KGs are, as defined by many authors, data models that can represent and extract knowledge using deductive and inductive techniques, integrating information from diverse, dynamic, and large-scale collections of data [3–5]. Experts may model the same domain differently, reflecting their unique perspectives and technical understanding. Combining these varying perspectives to create a comprehensive domain reference model is a complex task, often referred to as knowledge fusion or knowledge integration. Our proposal aims to address this challenge by providing a mechanism for the continuous evolution and validation of the domain reference model.

This article proposes a novel framework that leverages evolving KGs to address the semantic integration challenges inherent in Data Spaces. By accommodating the dynamic nature of both data and domain expertise, our approach ensures that the domain reference model remains current and accurate, thereby enhancing data interoperability.

2 Background

The knowledge fusion problem was described in the study of Dong et al. [6] as the process of employing multiple knowledge extractors to extract values from each data source, and then deciding the degree of correctness of the extracted knowledge. The both definitions of knowledge fusion problem are strongly aligned with the Gray and Rumpe [2] definition of domain reference model.

A significant challenge in updating a domain reference model is the requirement to reevaluate the knowledge and data each time new information is added [7]. This is because new knowledge and data can potentially change the context or validity of previous integrations (become outdated [7, 8]) in earlier integrations. As a result, not only must the new knowledge and data be integrated, but the entire model must also be reassessed to ensure that all integrations remain coherent and accurate.

KGs, due to their flexible structure, offer a promising approach to this challenge. However, traditional KGs do not incorporate evolving features, where knowledge dynamically grows and changes over time to include continuously emerging new facts [9]. In the process of incremental construction of KGs, the input encompasses not only the new data intended for addition, but also the existing version of the KG [10]. The research of Liu et al. [9] identifies key differences between traditional KGs and evolving KGs. The authors argue that evolving KGs depict interactions across different generation times, adding a temporal dimension that traditional KGs cannot model due to their inherent incompleteness.

3 Proposal

Semantic integration across multiple data sources and domains is a significant challenge in today's data-driven world. Our proposal. depicted in Figure 1, leverages Data Spaces and IDS connectors to create an evolving domain reference model that tackles this challenge effectively. The accompanying diagram illustrates our approach, which centers on continuously enhancing and updating a domain-specific KG. The starting point of our approach is the **Domain Reference Model Knowledge Graph**. At this stage, the KG encapsulates the current state of domain-specific knowledge, including entities and their relationships. This graph serves as a foundation upon which we build and refine our understanding of the domain. To enrich this model, we employ **Knowledge and Data Intelligent Unifiers**.

These methods perform essential tasks such as blocking, linking, knowledge completion, and alignment. By doing so, they ensure that the KG is accurate, complete, and well-aligned with various data sources. Next, **Knowledge and Data Wrappers** encapsulate the extracted knowledge and data, preparing it for integration into the domain reference model. This step is crucial for maintaining the integrity and usability of the integrated information. An **Assessment** phase follows, where the quality and relevance of the integrated knowledge and data are evaluated. This step ensures that only relevant information is incorporated into the evolving model. T

The **Knowledge and Data Extractors and Mappings** components are responsible for extracting knowledge and data from a variety of external sources, including schemas, ontologies, domain vocabularies, and data through the IDS connector. The extracted information is then mapped onto the existing domain reference model, facilitating seamless integration. This integrated knowledge and data lead to the creation of an updated **Domain Reference Model (t+1)**. This new version represents the next stage in the evolution of the domain model, enriched with new knowledge and data. Our framework can operate within the broader ecosystem of **Data Spaces (k+1)**. Data Spaces represent the diverse and dynamic data environments from which new knowledge and data are continuously sourced. The iterative nature of this process ensures that our domain reference model remains relevant and up-to-date.

Our approach addresses the semantic integration problem by creating an evolving domain reference model that adapts to new knowledge and data. This model benefits from the robustness of KGs and the standardized data exchange provided by IDS connectors. The process involves:

- **Incremental Evolution**: Continuously updating the domain reference model with new knowledge and data.
- **Intelligent Unification**: Ensuring the integrated knowledge is accurate, complete, and aligned.
- **Dynamic Adaptation**: Iteratively enhancing the reference model to keep pace with evolving domains and data sources.

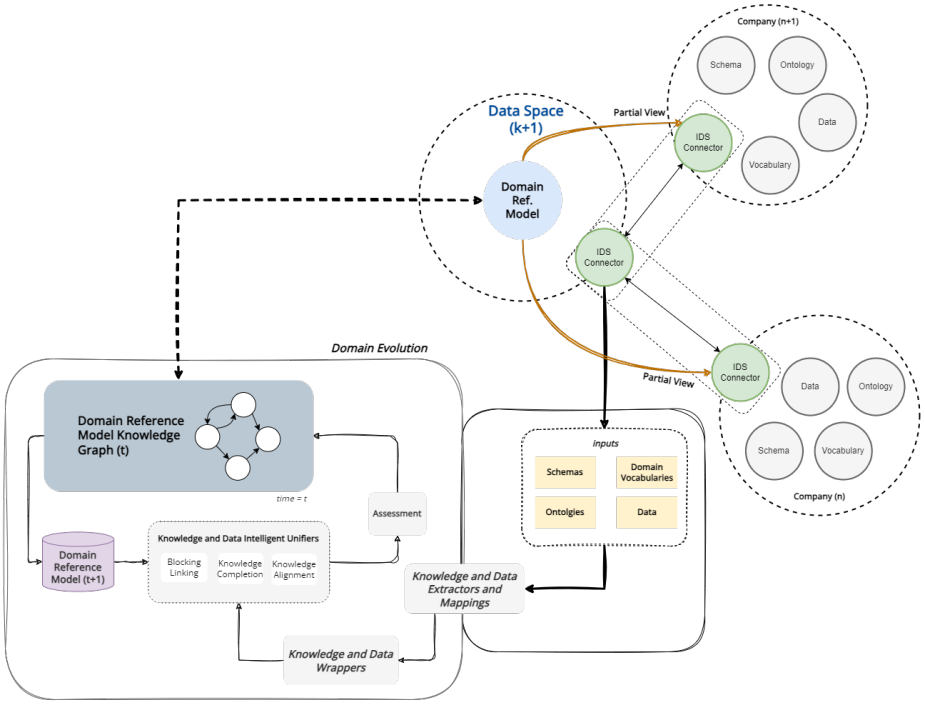

Fig. 1 Proposal of an Evolving Domain Reference Model framework to a Data Space

4 Conclusion

This paper introduces a novel framework designed to address the complexities of semantic integration within Data Spaces. By leveraging evolving KGs and intelligent unification techniques, our approach enables the creation of a dynamic and adaptable domain reference model.

A key challenge that remains unexplored is the assessment of the validity and the relevance of knowledge after new integrations. Furthermore, ensuring that the information is not outdated and remains accurate is still an ongoing concern. Therefore, the dynamic process of alignment must be continuously improved to handle the growing and changing nature of KGs. This includes developing artifacts (methods, models, frameworks, etc.) that can adapt to new knowledge and data, ensuring seamless integration and accurate alignment.

Future work will focus on refining the knowledge and data extraction and mapping processes, as well as exploring advanced techniques for knowledge completion and alignment. Additionally, we aim to evaluate the framework's effectiveness in real-world Data Space scenarios, further solidifying its potential to improve semantic integration in the ever-evolving landscape of data-driven technologies.

References

[1] Fettke, P., Loos, P.: Classification of reference models - a methodology and its application. Inf. Syst. E-Business Management **1**, 35–53 (2003) https://doi.org/10.1007/BF02683509

[2] Gray, J., Rumpe, B.: Reference models: how can we leverage them? **20**(6) (2021) https://doi.org/10.1007/s10270-021-00948-0

[3] Sjarov, M., Franke, J.: Towards knowledge graphs for industrial end-to-end data integration: Technologies, architectures and potentials. Lecture Notes in Production Engineering **Part F1160**, 545–553 (2022) https://doi.org/10.1007/978-3-030-78424-9_60

[4] Hogan, A., Blomqvist, E., Cochez, M., D'Amato, C., Melo, G.D., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S., Ngomo, A.-C.N., Polleres, A., Rashid, S.M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., Zimmermann, A.: Knowledge graphs. ACM Computing Surveys **54**(4) (2021) https://doi.org/10.1145/3447772

[5] Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., Taylor, J.: Industry-scale knowledge graphs: Lessons and challenges. Communications of the ACM **62**(8), 36–43 (2019) https://doi.org/10.1145/3331166

[6] Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., Zhang, W.: Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion, pp. 601–610. Association for Computing Machinery, ??? (2014). https://doi.org/10.1145/2623330.2623623

[7] Tu, H., Yu, S., Saikrishna, V., Xia, F., Verspoor, K.: Deep Outdated Fact Detection in Knowledge Graphs, pp. 1443–1452 (2023). https://doi.org/10.1109/ICDMW60847.2023.00184

[8] Ilyas, I.F., Rekatsinas, T., Konda, V., Pound, J., Qi, X., Soliman, M.: Saga: A Platform for Continuous Construction and Serving of Knowledge at Scale, pp. 2259–2272 (2022). https://doi.org/10.1145/3514221.3526049

[9] Liu, J., Zhang, Q., Fu, L., Wang, X., Lu, S.: Evolving knowledge graphs. In: IEEE INFOCOM 2019 - IEEE Conference on Computer Communications, pp. 2260–2268 (2019). https://doi.org/10.1109/INFOCOM.2019.8737547

[10] Hofer, M., Obraczka, D., Saeedi, A., Köpcke, H., Rahm, E.: Construction of Knowledge Graphs: State and Challenges (2023)