

Semantic Data Management for Managing Heterogeneous Data Sources in Chemistry 4.0

Sayed Hoseini

Department of Electr. Engineering and Computer Science, Niederrhein University of Applied Sciences, Krefeld, Germany

Abstract

Managing large volumes of data poses significant challenges due to the variety of formats, distribution across departments, and different governance structures within organizations. In research and industry environments, this complexity is compounded by the need for streamlined data handling processes to support automated workflows and machine learning (ML) applications. Integrating implicit contextual knowledge alongside data artifacts is critical, especially for non-expert users accessing the data. Data lakes provide a scalable solution by aggregating raw data from disparate sources with minimal upfront integration costs. However, without proper integration, data analysis and interpretation is hindered, rendering the data lake effectively inoperable. This PhD research addresses these challenges by applying semantic data management (SDM) techniques inside a semantic data lake. While initial milestones have been achieved through a systematic literature review and a concrete implementation, further efforts lie ahead. First, the emergence of large language models offers numerous opportunities for automating previously manual processes. Leveraging these models can significantly improve the efficiency of common SDM tasks. Second, extending the application of SDM techniques to data analytics can facilitate the integration of diverse data sources into ML pipelines. Ultimately, we aim to bridge the gap between Big Data and Semantic Web technologies, anticipating the development of advanced semantic data lake solutions in the foreseeable future.



Keywords


Semantic Data Management, Semantic Data Lakes, Semantic Machine Learning


1. Problem statement

Large amounts of data are generated every second to enable the subsequent collection, storage, usage and analysis for various applications. However, managing data can be challenging not only due to the variety of data formats, but additionally, it is often distributed across different departments within a company, under different governance regimes, network topologies and data models. At the Institute for Surface Technology of the Niederrhein University of Applied Sciences (HIT), where this research takes place, similar challenges need to be addressed [1]. There, an automation platform for material development is in operation and various datasets of experiments for paints and coatings need to be captured, cataloged, and made available for ML to be applied in chemistry for experiment suggestions and analysis. Data sources are


Proceedings of the Doctoral Consortium at ISWC 2024, co-located with the 23rd International Semantic Web Conference (ISWC 2024)

* supervised by Christoph Quix  & Stefan Decker .

 sayed.hoseini@hs-niederrhein.de (S. Hoseini)

 <https://www.hs-niederrhein.de/data-science> (S. Hoseini)

 0000-0002-4489-9025 (S. Hoseini)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

very heterogeneous, e.g., streams of machine sensors, interfaces of control software, databases and images of microscopes, and scripts and model checkpoints from ML. As a result, tasks for analyzing data, such as collecting, accessing, searching, understanding and processing data, become very time-consuming. This makes it difficult to realize visions such as *Chemistry 4.0*, which refers to the digital transformation of the chemical industry and emphasizes the integration of data-driven systems for increasing degrees of automation [2]. The centralized management of all (meta-) data with integrated data analytics using a uniform data management system is thus very attractive and actively researched [3, 4, 5].

Data lakes are scalable schema-less repositories to ingest raw data in its original format from heterogeneous data sources. Only minimal effort is required for ingesting data into a data lake making it an efficient tool for collecting, storing, linking, and transforming datasets [4]. However, this approach only postpones the upfront cost of integration, which is why they suffer from the risk of turning into a data swamp [6, 7]. In addition, many existing systems lack matured functions to support data analytics [3, 8]. Furthermore, industrial ML suffers from low transparency of ML towards non-ML experts, poor and non-unified descriptions of ML practices for reviewing or comprehension due to custom-made ad-hoc solutions tailored only to specific applications affecting their re-usability [9].

The main goal of this research is to develop a prototype for the industrial chemistry context of the HIT that not only manages the various (meta-)data assets, but also facilitates data integration, ultimately empowering users unfamiliar with data analytics to derive ML models.

The importance of data integration is rooted in the fact that those users, who ingest the data in the lake and are responsible for the data, may not belong to the group of data scientists, who are going to use the data later on. Likewise, a data scientist crafting a specific model seeks clarity and ease of understanding the detail about the design. Thus, the implicit context knowledge needs to be committed alongside any created artifacts to assist a third party with limited domain knowledge to interpret and use the received assets later on.

The problem statement can be formulated mathematically. Let:

- $D = \{d_1, d_2, \dots, d_n\}$: the set of heterogeneous data sources,
- $A = \{a_1, a_2, \dots, a_k\}$: the set of analytical models applied to D for generating insights,
- $M = \{m_1, m_2, \dots, m_m\}$: the set of metadata artefacts that describe and link data items

managed and stored by a data lake. The objective is to minimize the **human effort** required to prepare and integrate heterogeneous data sources through metadata, leveraging the capability of the lake to derive insights from ML with maximum automation and smart assistance.

$$\text{Minimize: } E_{\text{total}}(D, A, M) = E_{\text{prep}}(D, M) + E_{\text{use}}(D, A, M) + E_{\text{meta}}(M)$$

- $E_{\text{prep}}(D, M)$: Effort required to harmonize, transform, and integrate heterogeneous data sources D using the available metadata M .
- $E_{\text{use}}(D, A, M)$: Effort required for users to interpret, and utilize D , A , and M for deriving insights and crafting ML models.
- $E_{\text{meta}}(M)$: Effort required to create and maintain metadata M .

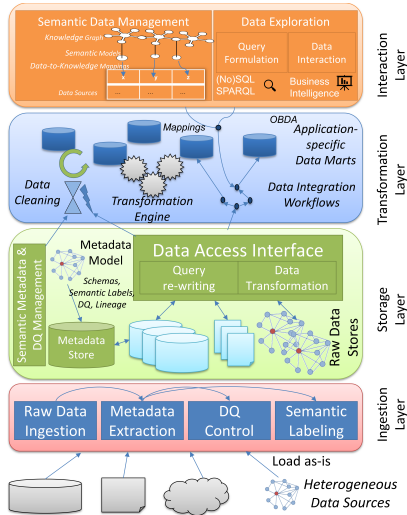


Figure 1: Semantic data lake architecture from [10]

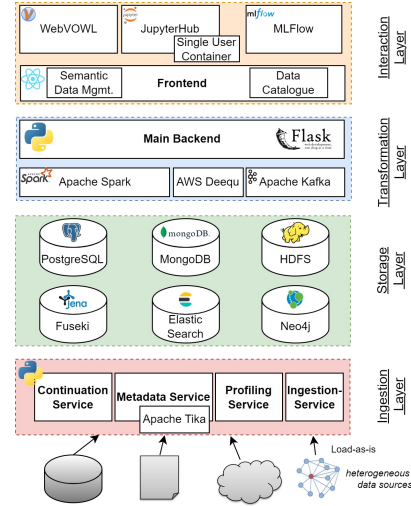


Figure 2: SEDAR architecture from [13]

2. Related work

Semantic data management (SDM) is one way of formalizing the context and domain knowledge of data sources [10]. SDM proposes the linkage of metadata to knowledge graphs (KG) based on the Linked Data principles [11] to provide more meaning to the data in the lake by establishing an additional semantic layer between the data and the knowledge layer [12]. A semantic layer can be used not only for data management but also to address the challenge of integrating data from heterogeneous sources [13].

Semantic data lakes store and manage this serialized semantics between data sources. They are a specific form of traditional data lakes that extend the capabilities through a semantic layer that enriches and connects the stored data semantically. The semantic data lake explicitly integrates semantic descriptions into its data management and governance capability [14, 15], where an ontology or KG serves as a universal data model, offering a conceptual representation of an organization’s data assets. In Figure 1, we propose a four-layered data lake architecture in [16], where especially metadata-related functions are enriched with semantics. For example, a semantic labeling component in the ingestion layer adds semantic labels to the extracted metadata elements. The semantic information (labels, models, KG, etc.) is managed in the storage layer in an extended semantic metadata repository. To facilitate the usage and interpretation of data the interaction layer has several additional components, e.g., for browsing the KG and semantic models and editors for refining the semantic mappings and models. Figure 2 represents a particular instantiation of this architecture (see C2) illustrating the various utilized technologies along the four layers.

Data Management for ML has been well-researched for at least ten years [17] and one subfield is also known as MLOps [18]. Hai et al. [3] underline the importance of ML-driven metadata management and in-lake ML which means supporting the training and inference process directly inside the data lake platform. Zhao et al. and Schlegel et al. [8, 19] present

metadata models for data lakes to capture not only descriptive but also analytical info about datasets and performed analyses. *MLSea* [20] is a resource consisting of *MLSO*, an ontology to model ML pipelines, *MLST*, a collection of taxonomies of ML-related concepts, and *MLSea-KG*, a KG containing ML datasets, pipelines and scientific works from diverse sources. By leveraging semantic technologies *MLSea* integrates ML datasets, experiments, software and scientific works for improving the search, explainability and reproducibility.

Large Language Models (LLMs) are expected to have a major impact on the landscape of data utilization and exchange. LLMs have demonstrated remarkable capabilities in understanding, generating, and processing vast amounts of textual data [21, 22, 23]. Promising fields of LLM application are the integration of heterogeneous data sources in the sense of SDM [24, 25] and automated machine learning (AutoML) [26, 27].

3. Research Questions

Closer collaboration between human-machine and machine-machine systems has revolutionized the current industrial landscape, leading to Industry 4.0 [28]. Here, challenges in terms of data management are to be addressed [5]. The advantage of employing a data lake system lies in the centralized management of (meta-) data and analytics. Thus, all model artifacts and their associated datasets, are accessible, registered, documented, and understandable by both humans and machines. The main goal of this research is to install such a prototype in the industrial chemistry context of the HIT leading to the following research questions and related hypotheses:

RQ1: *What role does SDM play in improving the integration and usability of heterogeneous data generated in an industrial context, particularly facilitated within a semantic data lake?*

H1: *SDM facilitates the integration of heterogeneous data sources and enhances data usability by providing a unified structure and enabling interoperability based on Linked Data principles.*

To manage heterogeneous data it is important to have a clear and logical structure when presenting this information. This demands a common understanding across the data landscape, i.e., *a lingua franca for data moderation* [29] based on the Linked Data principles.

RQ2: *How can LLMs be utilized to identify and formalize the context of given datasets, creating a full semantic model?*

H2: *LLMs automate substeps in semantic model creation, in particular semantic labeling.*

Automating the semantic modeling task is complex, because creating semantic models entails deciphering the existing data source and establishing connections between data attributes and concepts drawn from a KG. Open questions remain on how to utilize the LLM for individual tasks along a pipeline or instead prepare the LLM for the entire task.

RQ3: *How can semantic descriptions of data sources, ML pipelines, and their context be used to enhance data analytics within the data lake?*

H3: *Structured semantic knowledge about ML pipelines improves accuracy and efficiency of contemporary methods for automating ML workflows.*

While the demand is increasing, ML models are still often manually created by humans, because

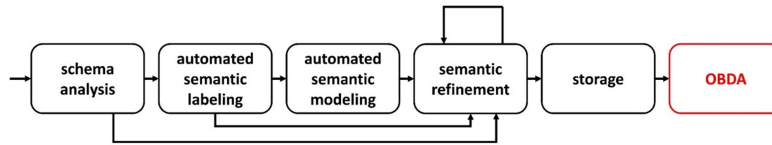


Figure 3: SDM pipeline inspired by [32] and extended by Ontology-based data access (OBDA)

the need for statistical and technical knowledge pose significant challenges for non-technical users [30]. Current methods are only capable of assisting in the substep of model creation [30], but data integration is a major obstacle [31].

4. Research Methods

Contribution 1 (C1): Systematic Literature Review: We systematically reviewed the literature of the last 20 years of research in the field of SDM related to semantic data lakes in particular [10]. The article gives an overview of semantics-based methods for data management, access, and integration and related those findings to current semantic data lake proposals. Furthermore, we identified a gap in today’s landscape between present data lakes, semantic technologies for data accessing, and the semantic modeling of heterogeneous datasets.

Contribution 2 (C2): The Semantic Data Reservoir (SEDAR) [13] is an implementation to bridge this gap. *SEDAR* is a prototype (see Figure 2) of a semantic data lake built on existing open-source technologies in the area of big data management. For the implementation of *SEDAR* we were inspired by the SDM pipeline (see Figure 3). The pipeline is designed for modeling data at the schema level and the first phase after extracting those schemas is automated semantic labeling, because semantic labels are a prerequisite for deriving a full semantic model automatically followed by semantic refinement, i.e. manual oversight to verify the automated outcomes. We then extended the pipeline and reinterpreted the storage phase conceptually, in the sense that we convert the resulting semantic model into RML mappings [33] to be used for Ontology-based data access (OBDA) [34]. OBDA allows for on-demand translation of queries against heterogeneous data sources directly in their original form without having to know how the data is organized physically, which is particular attractive in data lake environments. Thus, *SEDAR* implements a polystore with semantic query processing engine grounded on semantic models. The synergy between the automation platform at the HIT and *SEDAR* has been utilized in production and presented as original research at the *ICPS’24* conference [35].

Contribution 3 (C3): Automated Semantic Labeling using LLMs. In a publication for the *ESWC* conference [25] we conduct experiments demonstrating the applicability of LLM for semantic labeling and propose directions to address discovered challenges.

Contribution 4 (C4): Standardizing ML pipelines. Recently we have continued progressing *SEDAR* towards the support of standard ML pipelines with higher degrees of automation [36].

5. Evaluation

This PhD is already in a later stage, hence some research questions can already be addressed. Through **C1**, we have illuminated how SDM can help with the management of heterogeneous process data and gained knowledge of the current state of the field to understand how other researchers aim to implement particular SDM techniques. Open questions remain on how to convert these formal ideas into a particular implementation. Through **C2**, we proved to a wider audience how semantic processing can meet modern big data requirements. Therefore, we accept **H1** by providing a comprehensive field survey and demonstrating practically how the semantic layer of *SEDAR* enables more expressive data management, integration, and access. Through **C3** we address the applicability of LLMs for the first steps in the semantic model creation process. The experiments demonstrate the feasibility of utilizing LLMs for semantic type detection with a fixed or limited set of labels derived from legacy KGs. The findings further suggest that LLMs can effectively engage in semantic type detection tasks even when presented with new, unfamiliar, or arbitrary domain ontologies, by leveraging their inherent knowledge and understanding of language and as well as additional contextual information that is possibly provided alongside the ontology. Therefore, we accept the premise of **H2**. Through **C4**, we have been progressing towards standardizing ML pipelines. In the future, we plan to research how to perform a fusion between the SDM techniques and the existing works towards automating ML. To this end, we want to propose a software system that allows to reuse and generalize data analytics for arbitrary use cases. The goal is to answer **RQ3** by incorporating structured semantic knowledge about previously conducted ML experiments, such as the *MLSea KG* [20] to improve the efficiency and accuracy of current automated ML methods. By addressing the more challenging preceding phases of any ML project, i.e. business & data understanding, and especially **data preparation & integration** [37], this research agenda will advance the SOTA.

6. Conclusion and Future Work

This doctoral research addresses the challenge of managing diverse data sources and their integration into common ML pipelines semantically. So far, to face this issue, we first conducted a systematic literature review, then presented *SEDAR*, an open-source data management platform. We then proceeded to investigate the applicability of LLMs for semantic labeling and to enhance *SEDAR* to standardize ML pipelines by integrating principles from AutoML and MLOps. As this Ph.D. is already in a later stage, through these contributions we were able to answer the two out of three research questions. The remaining phase will focus on integrating semantically standardized ML pipelines to improve the efficiency of automated ML methods.

Acknowledgments

The author thanks Maribel Acosta and Christoph Quix for reviewing this article. This work has been sponsored by the German Federal Ministry of Education and Research in the funding program “Forschung an Fachhochschulen”, project *i²DACH* (grant no. 13FH557KX0).

References

- [1] D. A. C. Beck, J. M. Carothers, V. R. Subramanian, J. Pfaendtner, Data science: Accelerating innovation and discovery in chemical engineering, *AIChE Journal* 62 (2016). doi:<http://dx.doi.org/10.1002/aic.15192>.
- [2] B. Strehmel, C. Schmitz, K. Cremanns, J. Göttert, Photochemistry with cyanines in the near infrared: A step to chemistry 4.0 technologies, *Chemistry–A European Journal* 25 (2019) 12855–12864.
- [3] R. Hai, C. Koutras, C. Quix, M. Jarke, Data lakes: A survey of functions and systems, *IEEE TKDE* 35 (2023) 12571–12590.
- [4] F. Nargesian, E. Zhu, R. J. Miller, K. Q. Pu, P. C. Arocena, Data lake management: challenges and opportunities, *Proc. VLDB Endow.* 12 (2019) 1986–1989.
- [5] T. P. Raptis, A. Passarella, M. Conti, Data management in industry 4.0: State of the art and open challenges, *IEEE Access* 7 (2019) 97052–97093. doi:10.1109/ACCESS.2019.2929296.
- [6] W. Brackenbury, R. Liu, M. Mondal, A. J. Elmore, B. Ur, K. Chard, M. J. Franklin, Draining the data swamp: A similarity-based approach, in: *Proceedings of the Workshop on Human-In-the-Loop Data Analytics, HILDA '18*, Association for Computing Machinery, New York, NY, USA, 2018. doi:10.1145/3209900.3209911.
- [7] P. Sawadogo, J. Darmont, On data lake architectures and metadata management, *JJIS* (2021).
- [8] Y. Zhao, et al., Analysis-oriented metadata for data lakes, in: *Proceedings of the 25th IDEAS*, ACM, 2021, p. 194–203.
- [9] Z. Zheng, B. Zhou, D. Zhou, X. Zheng, G. Cheng, A. Soyly, E. Kharlamov, Executable knowledge graphs for machine learning: a bosch case of welding monitoring, in: *International Semantic Web Conference*, Springer, 2022, pp. 791–809.
- [10] S. Hoseini, J. Theissen-Lipp, C. Quix, A survey on semantic data management as intersection of ontology-based data access, semantic modeling and data lakes, *Journal of Web Semantics* 81 (2024) 100819. doi:<https://doi.org/10.1016/j.websem.2024.100819>.
- [11] C. Bizer, T. Heath, T. Berners-Lee, Linked data: The story so far, in: *Semantic services, interoperability and web applications: emerging concepts*, IGI global, 2011, pp. 205–227.
- [12] A. Pomp, A. Paulus, A. Kirmse, V. Kraus, T. Meisen, Applying semantics to reduce the time to analytics within complex heterogeneous infrastructures, *Technologies* 6 (2018) 86.
- [13] S. Hoseini, A. Ali, H. Shaker, C. Quix, SEDAR: A semantic data reservoir for heterogeneous datasets, in: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, Birmingham, UK, October 21-25, 2023, ACM, 2023, pp. 5056–5060.
- [14] H. Dibowski, S. Schmid, Y. Svetashova, C. Henson, T. Tran, Using semantic technologies to manage a data lake: Data catalog, provenance and access control, in: *Proc. Scalable Semantic Web Knowledge Base Systems Workshop*, volume 2757 of *CEUR WS*, 2020, pp. 65–80.
- [15] A. Usmani, M. J. Khan, J. G. Breslin, E. Curry, Towards multimodal knowledge graphs for data spaces, in: *Companion Proceedings of the ACM Web Conference 2023, WWW '23 Companion*, Association for Computing Machinery, New York, NY, USA, 2023, p.

- 1494–1499. doi:10.1145/3543873.3587665.
- [16] C. Quix, R. Hai, Data lake, in: *Encyclopedia of Big Data Technologies*, Springer, 2019. doi:10.1007/978-3-319-63962-8_{7}{-}{1}.
- [17] C. Chai, J. Wang, Y. Luo, Z. Niu, G. Li, Data management for machine learning: A survey, *IEEE TKDE* 35 (2023) 4646–4667.
- [18] S. Alla, S. K. Adari, *What Is MLOps?*, Apress, Berkeley, CA, 2021.
- [19] M. Schlegel, K. Sattler, Extracting provenance of machine learning experiment pipeline artifacts, in: *27th ADBIS Conference*, Barcelona, Spain, volume 13985 of *LNCS*, Springer, 2023, pp. 238–251.
- [20] I. Dasoulas, D. Yang, A. Dimou, Mlsea: A semantic layer for discoverable machine learning, in: A. Meroño-Peñuela, A. Dimou, R. Troncy, O. Hartig, M. Acosta, M. Alam, H. Paulheim, P. Lisena (Eds.), *The Semantic Web - 21st International Conference, ESWC 2024*, Hersonissos, Crete, Greece, May 26–30, 2024, *Proceedings, Part II*, volume 14665 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 178–198. URL: https://doi.org/10.1007/978-3-031-60635-9_11. doi:10.1007/978-3-031-60635-9_11.
- [21] H. Liu, R. Ning, Z. Teng, J. Liu, Q. Zhou, Y. Zhang, Evaluating the logical reasoning ability of chatgpt and gpt-4, *arXiv preprint arXiv:2304.03439* (2023).
- [22] T. Krüger, M. Gref, Performance of large language models in a computer science degree program, in: *Artificial Intelligence. ECAI 2023 International Workshops*, Springer Nature Switzerland, Cham, 2024, pp. 409–424.
- [23] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models, *ACM Trans. Intell. Syst. Technol.* (2024). doi:10.1145/3641289, just Accepted.
- [24] K. Korini, C. Bizer, Column type annotation using chatgpt, *arXiv preprint arXiv:2306.00745* (2023).
- [25] S. Hoseini, A. Burgdorf, A. Paulus, T. Meisen, C. Quix, A. Pomp, Towards llm-augmented creation of semantic models for dataspace, in: J. Theissen-Lipp, P. Colpaert, S. K. Sowe, E. Curry, S. Decker (Eds.), *Proceedings of the Second International Workshop on Semantics in Dataspace (SDS 2024) co-located with the 21st Extended Semantic Web Conference (ESWC 2024)*, Hersonissos, Greece, May 26, 2024, volume 3705 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024.
- [26] M. M. Hassan, et al., Chatgpt as your personal data scientist, 2023. *arXiv:2305.13657*.
- [27] N. Hollmann, et al., Large language models for automated data science: Introducing caafe for context-aware automated feature engineering, *Advances in Neural Information Processing Systems* 36 (2024).
- [28] A. Ustundag, E. Cevikcan, *Industry 4.0: Managing the Digital Transformation*, Springer, 2018.
- [29] S. Auer, *Semantic Integration and Interoperability*, Springer International Publishing, Cham, 2022, pp. 195–210. doi:10.1007/978-3-030-93975-5_12.
- [30] S. K. Karmaker, et al., Automl to date and beyond: Challenges and opportunities, *ACM Comput. Surv.* 54 (2021).
- [31] Z. Li, W. Sun, D. Zhan, Y. Kang, L. Chen, A. Bozzon, R. Hai, Amalur: Data integration meets machine learning, *IEEE Transactions on Knowledge and Data Engineering* (2024).

- [32] A. Paulus, A. Burgdorf, A. Pomp, T. Meisen, Recent advances and future challenges of semantic modeling, in: 2021 IEEE 15th International Conference on Semantic Computing (ICSC), IEEE, 2021, pp. 70–75.
- [33] A. Dimou, M. Vander Sande, P. Colpaert, R. Verborgh, E. Mannens, R. Van de Walle, Rml: A generic language for integrated rdf mappings of heterogeneous data., Ldow 1184 (2014).
- [34] G. Xiao, D. Calvanese, R. Kontchakov, D. Lembo, A. Poggi, R. Rosati, M. Zakharyashev, Ontology-based data access: A survey, International Joint Conferences on Artificial Intelligence, 2018.
- [35] S. Hoseini, et. al., Coatings intelligence: Data-driven automation for chemistry 4.0, in: 2024 IEEE 7th (ICPS), 2024, pp. 1–8. In-press.
- [36] S. Hoseini, M. Ibbels, C. Quix, Enhancing machine learning capabilities in data lakes with AutoML and LLMs, in: European Conference on Advances in Databases and Information Systems, Springer, 2024. Accepted.
- [37] R. Wirth, J. Hipp, Crisp-dm: Towards a standard process model for data mining, in: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, volume 1, Manchester, 2000, pp. 29–39.