The Synergy of Large Language Models and Dataspaces: A Functional Exploration

Sebastian Chmielewski[1], Tobias Meisen[1] and André Pomp[1]

[1]Institute for Technologies and Management of Digital Transformation, University of Wuppertal, Wuppertal, Germany

Abstract

Dataspaces provide a decentralized framework for secure and sovereign data exchange, yet ensuring semantic interoperability remains a key challenge. Large Language Models (LLMs) have emerged as powerful tools for enhancing data usability, particularly in metadata enrichment, semantic labeling, and data querying. This paper systematically investigates the role of LLMs in dataspaces through a structured literature review, identifying five core tasks: data querying, visualization, augmentation, cleaning, and metadata enrichment. Our findings highlight that metadata enrichment—specifically semantic labeling and modeling—is a primary area where LLMs can improve interoperability by automatically generating structured and meaningful metadata. However, challenges such as hallucinations, inconsistent labeling, and limited domain adaptation persist, affecting their reliability in real-world applications. We discuss approaches to mitigate these limitations, including the integration of LLMs with knowledge graphs and domain ontologies. By demonstrating how LLMs can contribute to automated metadata enrichment, this study provides a foundational analysis of their role in enabling FAIR (Findable, Accessible, Interoperable, Reusable) data principles in dataspaces.

Keywords

dataspaces, large language models (LLMs), semantic modeling, semantic labeling, semantic interoperability, semantics in dataspaces

This paper addresses issue 6 of the W3C Community Group on dataspaces (Issue #6).

1. Introduction

The increasing heterogeneity of data in the digital age presents a fundamental challenge for efficient data management. Ensuring that diverse datasets remain Findable, Accessible, Interoperable, and Reusable (FAIR) is essential for enabling meaningful data exchange across domains [1]. To address this challenge, dataspaces were introduced as a structured approach for managing interoperability, data sovereignty, and governance. They are becoming increasingly important by proving a decentralized framework for secure data exchange, primarily handling semi-structured and structured data while allowing organizations to retain full control over their assets [2, 3, 4]. Their flexible architecture supports seamless cross-domain data integration, enabling dynamic processing and transparent data governance. A core objective of dataspaces is to facilitate semantic interoperability, ensuring that heterogeneous data sources can be meaningfully linked and processed within a shared ecosystem [1, 5, 6, 7, 8].

However, more recently Large Language Models (LLMs) have emerged as a complement to existing dataspace technologies, offering new ways to enhance data usability. While dataspaces were originally developed to address data integration and interoperability challenges, the emergence of LLMs introduces an additional technological dimension, which could improve the management of these challenges. As pre-trained language models, LLMs excel in handling unstructured and semi-structured data, enabling tasks such as automated metadata enrichment respectively semantic labeling or advanced data querying. By reducing barriers to access and interpretation, LLMs can complement dataspaces by making data more discoverable, structured, and contextually meaningful [9]. While the implementation of dataspaces is often complex and requires specific knowledge, such as how to set up and use connectors, the ability

The Third International Workshop on Semantics in Dataspaces, co-located with the Extended Semantic Web Conference, June 01, 2025, Portorož, Slovenia

chmielewski@uni-wuppertal.de (S. Chmielewski); meisen@uni-wuppertal.de (T. Meisen); pomp@uni-wuppertal.de (A. Pomp)

0000-0002-1234-0000 (S. Chmielewski); 0000-0001-6917-2167 (T. Meisen); 0000-0002-1234-0000 (A. Pomp)

to use a technology such as LLMs allows information to be made easily accessible to end users so that the barrier to entry is as low as possible for people without a background in information technology. [1].

Building on these two promising aspects, this work systematically explores the role of LLMs in dataspaces and assesses their potential to enhance semantic interoperability and metadata enrichment. Specifically, we address the following research question: To what extent are Large Language Models utilized in dataspaces, and how can their integration enhance metadata enrichment?

The key contributions of this paper are:

1. A structured literature review of existing research on the integration of LLMs for task coverage in dataspaces.
2. An in-depth analysis of research in the area of the specific task 'metadata enrichment'.
3. Future research directions, emphasizing the integration of LLMs with domain ontologies and validation approaches to enhance reliability and accuracy.

By addressing these aspects, this paper contributes to the broader discussion on how LLMs can complement dataspaces in achieving FAIR data principles. The remainder of this paper is structured as follows: Section 2 presents the methodology used for the structured literature review, detailing the selection criteria and scope of analyzed publications. Section 3 discusses identified key tasks of LLMs in dataspaces. Additionally, a closer look on the LLM task 'metadata enrichment' is presented. Finally, Section 4 summarizes the findings, discusses implications for future research, and concludes the paper.

2. Methodology of the Literature Review

To address the research question mentioned above, we conducted a structured literature review. The search was performed across the following academic databases: IEEE Xplore, ACM Digital Library (ACM), Elsevier ScienceDirect, EBSCOhost, and Wiley Online Library. We applied a keyword-based search strategy using both English and German search terms in various combinations, as detailed in Table 1. We included German search terms because many dataspace initiatives like the International Data Space (IDS) originate from German-speaking countries.

Table 1
Overview of the used search strings.

„Large Language Model*“ AND „Data Space*“
„LLM“ AND „Data Space*“
„Large Language Model*“ AND „Dataspace*“
„LLM“ AND „Dataspace*“
„Large Language Model“ AND „Datenraum“
„LLM“ AND „Datenraum“
„Large Language Model*“ AND „Datenräume“
„LLM“ AND „Datenräume“

Given the thematic similarity of the selected keywords, we restricted our search to abstracts to ensure that only papers explicitly using both technologies were included, rather than those merely mentioning them in passing. The search was limited to the period of 2020 to 2025, as this time frame aligns with the release of modern language models such as BERT, which introduced capabilities applicable to various dataspace-related tasks, including metadata enrichment. Only papers available in English or German were considered. Our search yielded no relevant results in the ACM Digital Library, ScienceDirect or Wiley Online Library. We found two results in EBSCOhost and five in IEEE Xplore. After removing duplicates, only four distinct papers remained for further analysis. Due to the inclusion criteria, some recently published or yet-to-be-published papers were not available in the selected databases. To ensure a comprehensive review, we also examined publications from last year's ESWC Conference – International Workshop on Semantics in Dataspaces in the relevant fields. We identified that these conference proceedings were not indexed in the above-mentioned search engines yet. This additional search yielded more papers, which were screened according to our predefined criteria and the explicit usage of LLMs in dataspaces. Two additional papers were added to the review that address the application of LLMs in dataspaces using Retrieval Augmented Generation (RAG) in the process. In total, six papers specifically deal with the use of LLMs in dataspaces and form the corpus for our analysis of the LLM driven tasks and metadata enrichment.

3. Analysis of LLM Tasks in Dataspaces

3.1. LLMs in Dataspaces

We first outline the different ways in which LLMs interact with dataspaces, distinguishing between their use with the help of dataspaces and their application in managing dataspaces themselves. We then examine specific tasks that LLMs enable, supported by relevant studies from the examined paper corpus. Finally, we take a closer look at metadata enrichment and summarize key insights regarding metadata enrichment potential and challenges.

In the course of the literature review, it was found that the interaction between dataspaces and LLMs can be divided into two areas – as already stated by Distefano et al. [3]. The authors mention dataspaces as a source of training datasets for LLMs while also highlighting the usage of LLMs within dataspaces. In the following sections, we present and analyze papers which focus on the use of LLMs within dataspaces, while filtering out those in which dataspaces only serve as a data provider for LLMs. We identified the tasks of *data querying*, *data visualization*, *data augmentation*, *data cleaning* and *metadata enrichment* in our paper corpus. As shown in Table 2, the task of metadata enrichment is the main application area for using LLMs within dataspaces. Thus, metadata enrichment is addressed separately in the next section, to take a closer look on opportunities and areas of application.

Table 2
Overview of the different LLM tasks that we identified in dataspaces.

LLM task in dataspaces	Used by
Data cleaning	[3]
Data augmentation	[3]
Data querying	[3, 10]
Data visualization	[3]
Metadata enrichment	[1, 4, 5, 11]

In the following, we present the individual tasks of the LLMs in more detail. An important task is data cleaning, that is presented to improve the quality of the given data in a dataspace. Data cleaning corrects possible errors within the data itself to maintain consistency as well as data accuracy. Identified errors can be addressed using data cleaning rules such as transformation techniques, replacement strategies or filter criteria as shown by Distefano et al. [3].

Furthermore, LLMs can be used for the augmentation of datasets. Distefano et al. [3] improve datasets using special rules to ensure domain-specific integrity. The authors initially use the LLM to examine the existing structure of the data and use the identified structure to define the rules for the data augmentation. According to the authors, these rules enable the augmentation of realistic data.

In the context of dataspaces, another task of LLMs is *data querying*. This includes querying data sources, filtering and aggregating data as well as making it available to users. For instance, Hoseini et al. [5] implement a number of these LLM tasks in a workflow that interacts with a medical dataspace. In this study, the authors use Chat-GPT 4o to search, filter and summarize the desired data within the dataspace. The LLM then generates visual representations and charts (*data visualization* [5]).

Hermsen et al. [10] consider different variants of using a RAG in connection with a dataspace, whereby the task of data querying is used. In this case, the data provider independently indexes its own data and constructs a dedicated vector database for the RAG. The data consumer can then send queries that are processed by the provider. At the user's end, the information received is processed and displayed by an LLM. In another approach, the provider does not share any information with the consumer, as the RAG process takes place in the dataspace. In this case, a federator takes over the semantic search and the processes of the LLM [10].

Data visualization is another task of LLMs in dataspaces. The LLM's task is to visualize data extracted from the dataspace, while suggesting different styles to present the data. In the experiments carried out by Distefano et al. [3], the LLM is suggesting suitable visualization techniques in form of different charts and plots as well as presenting the different visualizations [3].

3.2. Metadata enrichment with LLMs in dataspaces

Metadata enrichment is an additional task, which is especially important for the interoperability in dataspaces. Metadata enrichment annotates the data with information and assigning an abstract formalization that is, for example, machine-readable. This includes, for instance, semantic labeling and semantic modeling. Because tasks such as semantic labeling and semantic modeling are time-consuming and complex tasks that require a high level of specialist knowledge, the LLM approach proves to be advantageous. This supports the interoperability approach within dataspaces [1, 4, 5, 11]. The task of LLM-based metadata enrichment is intended to make these processes more efficient.

In their publication, Hoseini et al. [5] investigate the usage of LLMs in the field of semantics, especially for data management in dataspaces as well as tasks of the creation of semantic models. They implement different variants of LLM-supported metadata enrichment in their experiments. In the first experiment, the LLM maps dataset labels to the VC-SLAM ontology [12] to assess its ability to recognize semantic types. The model is provided only with the dataset labels and must identify the most suitable ontology concepts. In the second experiment, the LLM receives additional textual documentation to evaluate whether supplementary information improves its semantic classification accuracy. The third experiment explores the adaptability of ChatGPT-4.0 by omitting the VC-SLAM ontology and instead leveraging various pre-trained ontologies, such as schema.org. The fourth and last experiment investigates the impact of ontology complexity by using a simplified version of the VC-SLAM ontology to determine how model performance varies with ontology granularity. These experiments confirm that LLMs can feasibly perform semantic type detection. However, they also reveal a persistent challenge: LLMs still exhibit hallucinations in certain cases. As a result, the authors recommend integrating knowledge graphs into dataspaces to enhance reliability.

Martorana et al. [11] also investigate LLM-supported semantic enrichment of metadata. In their experiments with ChatGPT-3.5, Google Bard and Google Gemini, the column headers of the datasets are automatically classified using a zero-shot method to enrich the metadata. The importance of the investigations carried out in the study is justified by the significance of metadata and semantic descriptions in the context of the FAIR principles. Further steps, such as an investigation with an extended dataset or checking whether LLMs can recognize a semantic similarity between the columns, are mentioned. A RAG approach is also recommended for further research.

Another study by Arnold et al. [1] has shown that the thematized LLM task could be used for example for extending the metadata, like domain specific properties by prompting the steps required to achieve the desired data enrichment. In conclusion the authors mention that it is possible for LLMs to prepare the data for use in dataspaces which could use the FAIR principles.

Besides these concrete approaches, the vision paper on the future of dataspaces discussed by Deshmukh et al. also describes that the use of LLMs for metadata enrichment would be useful. For example, they mention automate data mapping and semantic enrichment of data [4].

3.3. Metadata enrichment with LLMs across other Fields of Application

Since metadata enrichment is a task that does not necessarily require dataspaces as a basis, we have additionally investigated which other scientific works deal with the use of LLMs for metadata enrichment, especially semantic labeling and modeling. Here, available literature shows a number of studies on the combined use of LLMs and semantic labeling and semantic modeling, which underlines the relevance of this topic. We discuss a selection of these approaches whereby we especially focus on the datasets that the authors use for evaluating their approach. The work of Hoseini et al. [5] has shown that the combination of the used dataset and its corresponding ontology play a crucial role in evaluating the performance of LLMs for the task of semantic labeling and modeling. Table 3 gives an overview of the approaches and their used datasets.

Table 3
Overview of metadata enrichment approaches with their used datasets and applied LLMs.

Semantic Labeling	Semantic Modeling	Authors	Datasets	(Large) Language Models
-	x	Hou et al., 2024 [13]	- Regional Digital Control Center - Electrical and Mechanical Service Department Hongkong	- Open-source pre-trained LLM from Hugging Face
-	x	Ding, Du & Feng, 2025 [14]	- Data from various American art museums - Museum data from European Data Model - Football dataset	- GPT-4o
-	x	Mulayim et al., 2024 [15]	- Real facility data	- Claude 3.5 Sonnet - DeepSeek V2.5 - GPT-4 Turbo
x	-	Trabelsi, Cao & Heflin, 2020 [16]	- WebTables - Log Tables	- BERT
x	-	Guan, Chen & Koudas, 2023 [17]	- Yelp, YouTube, SMS, IMDB - Agnews, ArxivAbs, MedsAbs - TREC, CDR, Spouse, ChemPlot - SemEval	- BERT - GPT-3.5-turbo - GPT-4 - Llama2-Chat-70b - Llama2-Chat-7b - Llama2-Chat-13b
x	-	Li, Zhang & Wang, 2024 [18]	- Viznet - WikiTable	- GPT-4
x	-	Burgdorf et al., 2022 [19]	- VC-SLAM	- BERT

In the area of semantic modeling, we identified three papers focusing on LLMs and semantic modeling [13, 14, 15]. The study of Hou et al. [13] is investigating the increase in efficiency of smart building management systems through the integration of Artificial Intelligence (AI) and semantic modeling. Therefore they developed an AI driven knowledge base with a multi-agent architecture and LLMs and enrich the LLM with semantic models. The developed system is tested and evaluated in a case study. In one office Building they evaluate the effectiveness of the system and the implementation time. In addition to brick schema ontologies, they use the following two datasets in their study: Regional Digital Control Center (RDCC) and Electrical and Mechanical Service Department Hongkong (EMSD).

Mulayim et al. [15] discuss the use of semantic models, such as Brick Schema, in the building domain. Even if there are positive effects, these systems present challenges due to their steep learning curve and complexity, which can often only be mastered by employees with specialized knowledge. In the study, the authors analyze the use of LLM to meet these challenges. The LLM is to be used to create and query semantic models. The study describes requirements and metrics for evaluating the scalability and effectiveness of LLM-based tools using real building data and the brick schema ontologies [15].

In addition to semantic modeling, other authors use LLMs for semantic labeling. In a study from 2025, Ding, Du and Feng [14] use three LLMs for the tasks of semantic modeling and semantic labeling. The LLMs ChatGPT-4 Turbo, Claude 3.5 Sonnet and DeepSeek-V2.5 are used. In the investigations carried out, the LLMs receive three datasets consisting of structured data and ontologies. The datasets consist of data from various American art museums from the CIDOC Conceptual Reference Model, museum data from the European Data Model and football data set.

In the area of semantic labeling, we identified four different papers focusing on LLMs and semantic labeling [16, 17, 18, 19]. The study of Trabelsi, Cao and Heflin [16] introduces an approach for semantic labeling, utilizing Bert as a pre-trained language model. By analyzing both the data values and their surrounding context, this method enhances the accuracy of assigning semantic labels. The study primarily utilizes the datasets WebTables and Log Tables, large collections of structured tables from the web, for training and evaluation.

How LLMs can automatically create labeling functions, minimizing the reliance on manually annotated training data is investigated by the authors Guan, Chen and Koudas [17]. Using different prompting strategies, the study shows that LLMs can produce precise and varied labeling functions, which in turn enhance the overall quality of semantic labeling in datasets. The study evaluates its approach on multiple unstructured text classification and entity recognition datasets, like Yelp, Spouse, YouTube, and News Aggregator datasets, which contain labeled text samples for different semantic categories.

Another approach is introduced by Li et al. [18]. They propose leveraging LLMs to automatically generate labeling functions through prompt engineering, aiming to reduce the manual effort required in labeling training data for semantic type detection. The Viznet and WikiTable datasets are being used here [18]. In addition to LLMs, there are also studies that implement metadata enrichment using conventional language models rather than LLMs. In the study presented by Burgdorf et al., RoBERTa is used to label data from the VC-SLAM dataset [19].

In summary, the analysis of the examined studies in this chapter and subsection 3.2 shows that metadata enrichment is a task of LLMs in dataspaces and contributes to the implementation of the FAIR principles, especially interoperability. Through semantic labeling and semantic modeling, LLMs can contribute to enriching structured and unstructured data with metadata, thereby improving its usability and findability. Various studies show, for example, that LLMs can automatically generate metadata and augment additional data. At the same time, it becomes clear that the combination of LLMs with textual documentation can further increase the potential of metadata enrichment [5]. The broad spectrum of approaches, language models and datasets shown in Table 3 used in the context of metadata enrichment may indicate that LLM-based metadata enrichment could be successfully applied in a wide range of dataspace situations. Despite the approaches presented, various authors mention challenges, such as the occurrence of hallucinations or the inadequate ability to acquire domain-specific knowledge, which need to be investigated further in the future.

4. Discussion and Conclusion

The limited number of publications on the synergy between LLMs and dataspaces is surprising. Although several relevant papers were initially identified, only six explicitly focus on LLM applications in this context. Given the widespread attention on LLMs and their potential for dataspaces, this scarcity was unexpected. One possible reason is concerns within the research community about their practical implementation. Hoseini et al. [5] demonstrated that while LLMs perform well with established ontologies, they struggle with specialized ones, as often used in dataspaces, leading to inaccuracies and hallucinations.

The literature review highlights the diverse potential of integrating LLMs into dataspaces. LLMs can improve usability through tasks such as data querying, visualization, and cleaning, while also supporting metadata enrichment, particularly in semantic labeling and modeling. Metadata enrichment is crucial for semantic interoperability. It facilitates efficient data exchange and strengthens other LLM-driven tasks. These applications align with the FAIR principles, improving data findability, accessibility, interoperability, and reusability.

Despite these advantages, challenges remain, particularly the risk of hallucinations in LLM-generated metadata, raising concerns about reliability and consistency. Addressing these issues is a key avenue for future research, particularly in enhancing the accuracy of LLMs in semantic modeling and labeling. A critical aspect of this research should be the measurability and comparability of LLM applications in dataspaces. Additionally, incorporating historical data alongside textual documentation and predefined ontologies may offer promising directions for improving LLM performance in this domain.

The combination of LLMs and dataspaces opens up various additional technical possibilities. For example, LLMs enable data queries with user input in natural language. Furthermore, they support context-based search processes in RAG architectures as well as automated metadata enrichment. The previously presented research shows that visualization methods have been successfully proposed and realized. Accordingly, the potential of data visualization using LLMs can be pointed out.

As part of the conducted review, it was determined that a number of studies on LLM-based semantic labeling and semantic modeling have been accomplished. These studies were performed with a large number of datasets in a wide variety of domains, which made it possible to explore a diverse range of applications for this use. These studies emphasize the importance behind the use of LLMs for metadata enrichment, suggesting a high potential of this application area. A promising approach could be to consider the findings from the various domains in the research efforts in the area of dataspaces in combination with LLMs. We are aware that this study is only a first step in investigating the use of LLMs in dataspaces and metadata enrichment. In order to investigate the further possibilities of LLM-based metadata enrichment in dataspaces in more detail, the research results of other studies from other fields of application should be examined and analyzed for use in dataspaces. In the extended literature search, different synonyms of the relevant technical terms and a broad range of search engines should be selected. These investigations will be addressed in a comprehensive survey. In addition, the area of knowledge graph mapping with LLMs should be examined more closely. The methodologies used should be analyzed and examined how they can be implemented in the area of LLM-based semantic modeling. To this end, a modular pipeline is to be set up in which these steps are applied, examined and adapted for application in dataspaces according to the results achieved.

Altogether, the role of LLMs in dataspaces is to improve data handling in this environment and make it more user friendly by implementing the above tasks through LLMs. Previous solutions show a promising entry into this research area. In summary, it can be seen that the use of LLMs in dataspaces has only been implemented in a few projects up until now. The results indicate that LLM tasks such as metadata enrichment are useful to promote semantic interoperability and work to a certain extent, but should be further developed and optimized, e.g., to reduce barriers such as hallucinations and to improve the overall quality of the results.

Declaration on Generative AI

During the writing of this paper, the author(s) used DeepL and GPT-4o in order to: Grammar, translation and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

Acknowledgements

This work has been supported as part of the research project *DigiTal Zwilling* in collaboration with the city of Wuppertal, funded by the Federal Ministry of Housing, Urban Development and Building (BMWSB) and the Reconstruction Loan Corporation (KfW) through the funding program “Modellprojekte Smart Cities: Stadtentwicklung und Digitalisierung” (grant number 19454890).

References

[1] B. Arnold, J. Theissen-Lipp, D. Collarana, et al., Towards enabling FAIR dataspaces using large language models, 2024.

[2] G. Rehm, S. Piperidis, K. Choukri, et al., Common european language data space, 2024.

[3] S. Distefano, Y. Yifru, Exploring the interplay between dataspaces and large language models, 2024.

[4] R. Deshmukh, D. Collarana, J. Gelhaar, et al., Challenges and opportunities for enabling the next generation of cross-domain dataspaces, 2024.

[5] S. Hoseini, A. Burgdorf, A. Paulus, T. Meisen, C. Quix, A. Pomp, Towards LLM-augmented creation of semantic models for dataspaces, 2024.
[6] A. Braud, G. Fromentoux, B. Radier, O. Le Grand, The road to european digital sovereignty with Gaia-X and IDSA, IEEE Network 35 (2021) 4–5.
[7] A. Hutterer, B. Krumay, Integrating heterogeneous data in dataspaces - a systematic mapping study, 2022.
[8] A. Gieß, T. Schoormann, F. Möller, I. Gür, Discovering data spaces: A classification of design options, 2025.
[9] W. Zhao, K. Zhou, J. Li, et al., A survey of large language models, 2023.
[10] F. Hermsen, L. Nitz, M. Akbari Gurabi, R. Matzutt, A. Mandal, On data spaces for retrieval augmented generation, Gesellschaft für Informatik e.V, 2024.
[11] M. Martorana, T. Kuhn, L. Stork, J. van Ossenbruggen, Zero-shot topic classification of column headers: Leveraging LLMs for metadata enrichment, 2024.
[12] A. Burgdorf, A. Paulus, A. Pomp, T. Meisen, DocSemMap: Leveraging textual data documentations for mapping structured data sets into knowledge graphs, 2022.
[13] Y. Hou, K. Leung, P. So, Z. Fan, R. Lekan, Enhancing building services management system with AI and semantic model: A case study on improving system efficiency through an AI-based knowledge library, Osaka, Japan, 2024.
[14] N. Ding, J. Du, Z. Feng, Knowledge prompt chaining for semantic modeling, 2025.
[15] O. Mulayim, L. Paul, M. Pritoni, A. Prakash, M. Sudarshan, G. Fierro, Large language models for the creation and use of semantic ontologies in buildings: Requirements and challenges, in: Large Language Models for the Creation and Use of Semantic Ontologies in Buildings: Requirements and Challenges, ACM, New York, NY, USA, 2024, pp. 312–317.
[16] M. Trabelsi, J. Cao, J. Heflin, Semantic labeling using a deep contextualized language model, 2020.
[17] N. Guan, K. Chen, N. Koudas, Can large language models design accurate label functions?, 2023.
[18] C. Li, D. Zhang, J. Wang, LLM-assisted labeling function generation for semantic type detection, 2024.
[19] A. Burgdorf, A. Paulus, A. Pomp, T. Meisen, VC-SLAM—a handcrafted data corpus for the construction of semantic models, 2022.