15th IFK
Proc. of the 15th International Fluid Power Conference <Symposium/Conference>
<DOI>

Unlocking Semantic Interoperability in Industry with Large Language Models

Barkha Sharma, Malte Becker, Constantin Meier, Katharina Schmitz

RWTH Aachen University, Institute for Fluid Power Drives and Systems (ifas), Campus-Boulevard 30, 52074 Aachen, Germany

Abstract.

The digitalization of industrial processes within Industry 4.0 depends on the reuse of technical product data across systems, organizations, and lifecycle stages. In practice, the main challenge is not often the availability of data, but their heterogeneity, which prevents automated interpretation and semantic interoperability. This paper presents an automated pipeline for incrementally transforming existing heterogeneous technical data into interoperable representations. The approach builds on available industrial data and combines Large Language Models, retrieval-augmented generation, and the Asset Administration Shell to align manufacturer-specific descriptions with standardized vocabularies such as ECLASS. Using fluid power datasheets as an exemplary data source, the pipeline is introduced to demonstrate how heterogeneous engineering data can be made interoperable. The focus is on the pipeline concept itself, which aims to reduce manual mapping effort and support a gradual transition toward Industry 4.0-compliant digital engineering environments, rather than on the processing of datasheets. While demonstrated in the fluid power domain, the approach is transferable to other technical domains facing similar data heterogeneity challenges.

Keywords. Semantic Interoperability, Asset Administration Shell, ECLASS, Large Language Models, Retrieval-Augmented Generation, Digital Twin, Industry 4.0

1. INTRODUCTION

The ongoing digitalization of industrial processes has led to a steadily increasing availability of technical product data. In industrial practice, however, the fundamental challenge not the lack of data, but their significant heterogeneity. Technical data already exists in large volumes, yet it is described, structured, and maintained differently across departments, IT systems, and organizations. As a result, data cannot be automatically interpreted or reused, as a shared semantic basis is missing. Within the context of Industry 4.0, this lack of semantic interoperability limits the integration of technical product data into digital twins, automated engineering workflows, and cross-company data exchange [1].

Within companies, technical product data are typically distributed across ERP systems, PLM systems, engineering tools, spreadsheets, and document-based formats. Different departments often describe identical technical properties using varying terms, units, or levels of detail. Across organizational boundaries, this heterogeneity increases further due to

manufacturer-specific conventions and documentation practices. Although the underlying physical properties are equivalent, digital systems are unable to recognize these relationships [2]. Consequently, engineers are required to manually interpret, compare, and harmonize data, which leads to significant effort, introduces errors, and does not scale with increasing data volumes.

This challenge is particularly evident in the fluid power domain. Technical characteristics such as nominal pressure, maximum operating pressure, or flow rate are described inconsistently by manufacturers and even within individual organizations. For example, the maximum pressure of a hydraulic motor may be specified “maximum pressure”, “maximum operating pressure”, “p_max”, “peak pressure”, or “pressure limit”. While these descriptions refer to the same physical quantity, their semantic equivalence cannot be derived automatically. This lack of semantic interoperability significantly complicates the integration of fluid power components into digital engineering workflows, digital twins, and Industry 4.0 environments [3].

Standardized concepts such as the Asset Administration Shell (AAS) and semantic dictionary like ECLASS define suitable target structures for interoperable technical data. In practice, however, the transformation of existing heterogeneous data into these standardized representations remains largely manual. Industrial organizations cannot replace their existing data sources or enforce new data models across all systems. Instead, approaches are required that build on available data and enable a gradual transition toward semantic interoperability [4], [5], [6]. At the same time, recent advances in artificial intelligence, particularly Large Language Models (LLMs), have demonstrated the ability to capture semantic relationships within heterogeneous and partially unstructured data [7]. These capabilities provide new opportunities to support the semantic alignment of technical data without requiring fundamental changes to existing data sources [8], [9].

Against this background, this paper presents an automated pipeline that acts as an enabler for semantic interoperability. The proposed approach deliberately starts from existing technical data and allows organizations to retain their current data formats and structures. Using fluid power datasheets as a practical entry point, the pipeline combines LLMs, Retrieval-Augmented Generation (RAG) [10], and the AAS to incrementally align manufacturer-specific descriptions with standardized vocabularies such as ECLASS. In doing so, existing technical data can be progressively transformed into interoperable representations that support further digitalization efforts. While the approach is demonstrated in the fluid power domain, it is not limited to this field and can be transferred to other technical domains facing similar challenges related to heterogeneous product data.

The remainder of this paper is structured as follows: Chapter 2 discusses the theoretical foundations of the work. Chapter 3 reviews the state of the art and outlines the motivation for this work. Chapter 4 details the proposed methodology and implementation variants. Chapter 5 presents and discusses the evaluation results. Finally, Chapter 6 summarizes the findings and provides an outlook on future research directions.

2. THEORETICAL BACKGROUND

Achieving semantic interoperability is essential for digital manufacturing systems, allowing consistent data exchange and interpretation across heterogeneous environments. This

chapter introduces the theoretical foundations and core principles that form the basis of the proposed approach.

2.1. Interoperability

Interoperability is a key prerequisite for integrating heterogeneous systems in industrial and digital environments, enabling machines, software, and organizations to exchange data and effectively use the transferred information. In the context of Industry 4.0, interoperability supports collaboration across technical, organizational, and domain boundaries and forms the foundation of data-driven manufacturing [3].

As illustrated in Figure 1, interoperability can be understood as a layered concept in which higher-level coordination depends on capabilities provided by underlying layers.

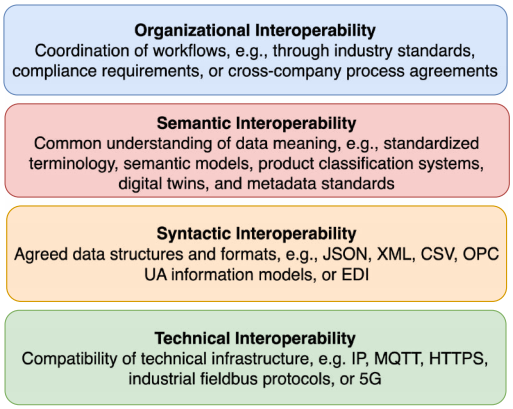


Figure 1. Levels of Interoperability [11]

At the organizational level, interoperability enables coordinated business processes and workflows across company boundaries. This coordination relies on semantic interoperability, which ensures that exchanged data conveys a consistent meaning across heterogeneous systems through shared vocabularies and semantic models [11].

The lower levels, namely technical and syntactic interoperability, provide the prerequisites for connectivity and structured data exchange but do not ensure a shared interpretation of information [11]. Consequently, semantic interoperability represents the critical transition from mere data exchange to meaningful information integration. This work therefore focuses on the semantic layer and explores how LLMs can support semantic alignment between unstructured technical data and standardized vocabularies.

2.2. Asset Administration Shell

The AAS is a core concept of Industry 4.0 that provides a standardized digital representation of physical or virtual assets. It acts as an interface between the physical and digital worlds, enabling interoperable management and exchange of asset-related data across the lifecycle. Within the Reference Architecture Model Industry 4.0 (RAMI 4.0) [12], the AAS constitutes the technical basis for digital twins. Its internal structure is defined by a formal metamodel specified by the International Digital Twin Association (IDTA) [4].

As illustrated in Figure 2, an AAS consists of one or more Submodels (SMs), each describing a specific aspect of an asset. Submodels are composed of SubmodelElements (SMEs), which represent individual data points or relationships. Core SME types include Properties, Files, ReferenceElements, and Ranges, which may be grouped using SubmodelElementCollections to represent complex structures. This hierarchical organization supports modular and scalable modeling of assets, ranging from individual components to complete systems [13], [14]y

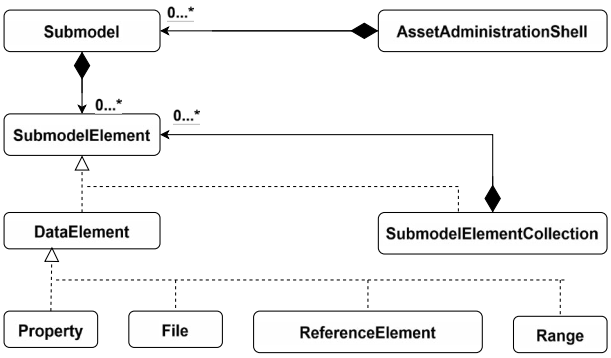


Figure 2. AAS hierarchy according to [5]

Depending on their functionality, AAS implementations can be categorized into three types:

- **Type 1 - Passive AAS:** Provides static information for documentation purposes.
- **Type 2 – Active AAS:** Enables automated data exchange via standardized interfaces.
- **Type 3 – Interactive AAS:** Supports bidirectional communication and autonomous behavior, effectively realizing a digital twin.

With increasing AAS maturity, both the volume and dynamics of exchanged data grow significantly. Consequently, consistent naming conventions and standardized semantic identifiers become essential to ensure reliable interoperability across tools and organizational boundaries, a requirement that is already becoming evident in fluid power research and industrial applications [15].

2.3. ECLASS Semantic Dictionary

Semantic interoperability requires a shared understanding of technical information across systems and organizations. A widely adopted approach to achieve this is ECLASS, a standardized framework for the classification and description of products and services. ECLASS provides language-independent, machine-readable representations of technical and commercial data, ensuring consistent interpretation across heterogeneous systems [15].

As illustrated in Figure 3, ECLASS organizes product information within a hierarchical classification structure, ranging from general product groups to specific classes, such as axial piston motors in the fluid power domain. Each class is associated with standardized properties and identified by International Registration Data Identifiers (IRDIs), enabling persistent and uniquely identifiable referencing of products, characteristics, and values across systems.

This structure supports interoperable and reusable data for applications such as digital twins, automated engineering, and cross-enterprise workflows.

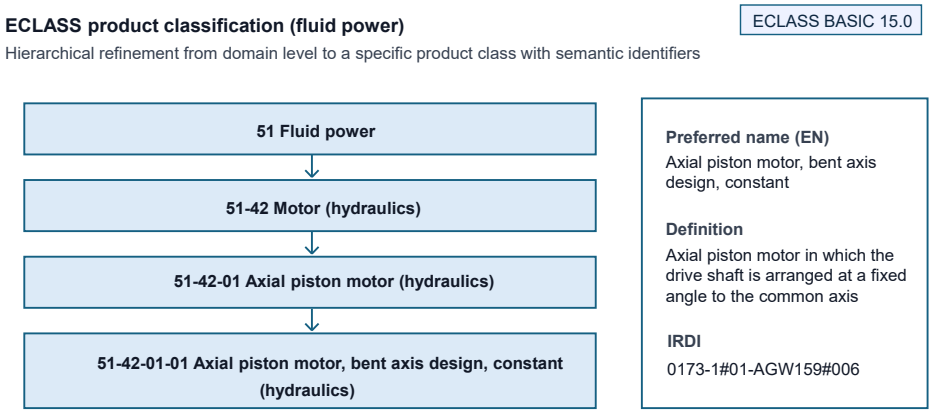


Figure 3. Example of hierarchical structure within the ECLASS classification [15]

Compared to conceptual standards such as IEC 61360 [16], which define general principles of data modelling, ECLASS provides a practical, domain-specific implementation that operationalizes these concepts for industrial use. Owing to its maturity and established adoption in domains such as fluid power, ECLASS serves as the semantic backbone for the semantic mapping approach presented in this work, particularly in the context of AAS. Further details on the internal ECLASS data model are outside the scope of this paper and are documented in the official ECLASS specifications [15].

2.4. Large Language Models

LLMs are based on the transformer architecture, which was introduced by Vaswani et al. [17]. The core element of this technology is the self-attention mechanism, in which each word is related to all other words in a text. As a result, it is possible to identify not only long dependencies but also complex semantic relationships [8], [10].

The development of LLMs is characterized by various fundamental learning approaches [18]:

- **Supervised learning:** Learning from labelled examples (e.g., a sentence is annotated with a category).
- **Unsupervised learning:** Learning from unstructured data without labels, only through pattern recognition.
- **Self-supervised learning:** The model generates the learning tasks itself, for example by predicting the next word or filling in missing text passages.

The development of LLMs typically follows three consecutive learning stages. First, during pretraining, the model learns general linguistic structures and statistical patterns from large volumes of text data using self-supervised objectives such as next-token prediction. Second, in the fine-tuning stage, the pretrained model is adapted to specific domains or tasks through supervised learning with curated datasets. Finally, in the alignment stage, instruction tuning or reinforcement learning from human feedback (RLHF) is applied to optimize the model's

responses for usefulness, safety, and factual accuracy. Together, these stages enable LLMs to combine broad language understanding with domain-specific reasoning capabilities [7], [9].

Although LLMs develop a broad understanding of language, they remain limited in two respects: Firstly, their knowledge is frozen at the level of the training data, and secondly, they may hallucinate information when context is insufficient. This is particularly problematic in industrial scenarios, such as when assigning technical data to standards, where precision and timeliness are crucial.

2.5. Retrieval-Augmented Generation

RAG is used to overcome these limitations. It extends a LLM with an external knowledge source that is consulted for every query, ensuring access to up-to-date and verifiable information. The RAG process follows three sequential steps [2], [19]:

1. **Retrieval:** In the first step, the user query is converted into a vector representation. This vector is compared against a knowledge base, such as technical documentation, standards, or AAS data sheets, to identify the most semantically similar text passages [19].
2. **Augmentation:** In the second step, the retrieved document excerpts are provided to the LLM as additional context. This augmented input serves as a factual reference for the model, anchoring its subsequent reasoning [19].
3. **Generation:** In the final step, the LLM produces an answer by combining its internal language knowledge with the externally retrieved information [19].

In this way, RAG combines two strengths: the language processing and generalization capabilities of LLMs with the timeliness and traceability of external knowledge sources. Transparency is particularly important in an industrial context as RAG allows answers to be provided with source references, which facilitates verification.

3. RELATED WORK

Achieving semantic interoperability remains a central challenge for realizing the Industry 4.0 vision. Reference architectures such as RAMI 4.0 and concepts like the AAS establish structured frameworks for linking physical assets with their digital representations. While these frameworks define harmonized data models and interfaces, the transformation of manufacturer-specific documentation into standardized vocabularies such as ECLASS or IEC remains largely manual in practice [12], [15]

In the fluid power domain, this challenge is particularly pronounced due to the high number of suppliers and the strong reliance on manufacturer-specific documentation practices. Hydraulic and pneumatic components are described using diverse terminologies, parameter definitions, and documentation formats. Although standards such as ISO 18582 and ECLASS define structured property sets for pumps, valves, and actuators, their consistent application across vendors is limited. As a result, semantic inconsistencies persist, hindering automated data integration, digital twin creation, and cross-company data exchange, which are key objectives of Industry 4.0 [20].

Recent research has explored the use of artificial intelligence to reduce manual effort in semantic mapping and interoperability tasks. LLMs, often combined with RAG, have

demonstrated the ability to extract, interpret, and structure data from unstructured technical documents [21], [22]. Xia et al. proposed an LLM-based approach for generating AAS from unstructured asset descriptions, showing that technical attributes can be identified and standardized with reduced manual intervention [14]. De Bellis investigated how semantic relations are encoded within LLM representations and how these can support ontology alignment tasks [23]. Wang et al. introduced Unify, a system for deriving structured semantics from heterogeneous industrial documents using LLMs [24]. In infrastructure engineering, Hoeltgen et al. demonstrated the enrichment of technical condition data through LLM-based semantic processing, illustrating the applicability of these methods beyond manufacturing domains [8].

In parallel, research on knowledge graphs and linked data emphasizes structured semantic representations as a foundation for interoperability across engineering systems [11]. While these approaches provide robust mechanisms for representing standardized relationships, they typically rely on manually curated data and predefined schemas, limiting their scalability and adaptability to heterogeneous legacy documentation.

Despite these advances, most existing approaches remain largely domain-agnostic and do not sufficiently address the specific terminology, parameter sensitivity, and contextual dependencies inherent to fluid power systems. Given the domain's reliance on precise physical parameters and cross-vendor component integration, automated semantic alignment tailored to fluid power documentation remains underexplored.

This work addresses this gap by proposing a domain-specific methodology that builds on existing heterogeneous technical data and supports their gradual transformation toward semantic interoperability. By combining retrieval-based context, LLM-supported reasoning, and established standards such as ECLASS and the AAS, the approach enables manufacturers and system integrators to retain their current data sources while systematically aligning fluid power documentation with standardized semantic models. In doing so, the methodology reduces manual mapping effort and supports digital twin creation and lifecycle-oriented data integration in accordance with Industry 4.0 principles.

4. METHODOLOGY AND RESULTS

This chapter introduces the proposed workflow for achieving semantic interoperability between unstructured industrial datasheets and standardized digital representations. The methodology is executed in parallel for a cloud-based and a fully local pipeline, both following the same processing stages and differing only in the execution environment.

The pipeline was implemented in Python due to its extensive ecosystem for document parsing, vector computation, and model integration. Its modular design supports flexible deployment in scalable cloud environments as well as local setups where data confidentiality is required. Figure 4 summarizes the shared workflow, consisting of five main stages described in the following subsections.

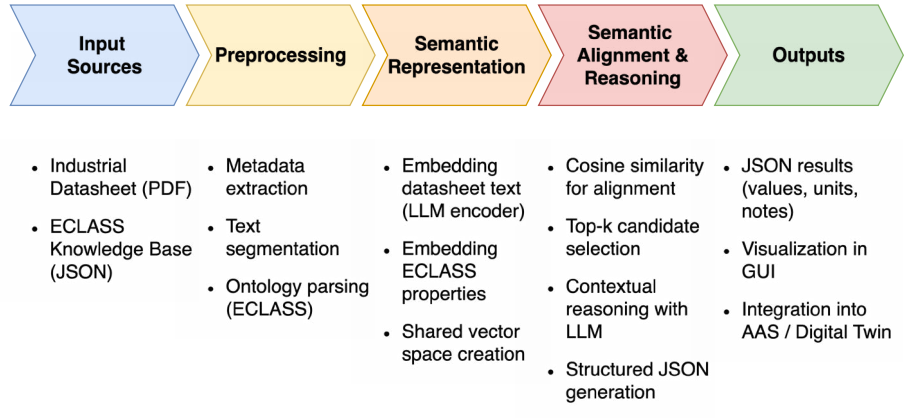


Figure 4. Workflow for achieving semantic interoperability

4.1. Input Sources

The workflow integrates two complementary data sources:

- **Technical datasheets (PDF):** These documents contain manufacturer-specific data on hydraulic and pneumatic products, such as nominal pressure or flow rate, but differ strongly in terminology, and structure. This heterogeneity makes automated processing difficult and motivates the need for a domain-independent mapping method.
- **ECLASS semantic knowledge base (JSON):** Using ECLASS ensures that the extracted data can be semantically aligned with established industrial standards and reused across systems, e.g., in AAS Submodels.

Integrating both sources forms the foundation for semantic interoperability. The datasheet delivers real-world technical content, while ECLASS defines the target semantic schema that standardizes its meaning.

4.2. Preprocessing

The preprocessing stage converts both the unstructured datasheets and the structured ECLASS ontology into machine-readable form. The workflow begins with data acquisition, where technical datasheets, usually PDF files containing textual descriptions, tables, and diagrams, are processed page by page. Text and metadata such as document titles or product identifiers are extracted using the Python library pdfplumber.

In the cloud-based setup, the same operation is performed through LangChain's PyPDFLoader, which supports API-driven processing. All extracted content is saved in JSON format, ensuring that each segment can be traced back to its original page and reused in later stages without reprocessing.

In parallel, the ECLASS ontology is loaded as a JSON file, providing the hierarchical structure of classes and standardized property definitions that serve as the reference vocabulary for later semantic alignment. Following extraction, the textual data undergoes normalization to correct common formatting issues in industrial PDFs, such as broken lines, hyphenations,

and layout artifacts. The cleaned text is then divided into overlapping sections using LangChain's RecursiveCharacterTextSplitter.

Unlike rule-based segmentation by headings or tables, this character-based approach is robust to inconsistent layouts and ensures that related data, such as a parameter and its corresponding unit, remains within a single segment. By the end of this stage, both sources are harmonized and structured, providing a consistent foundation for the embedding and semantic mapping steps.

4.3. Semantic Representation

In the third stage, the workflow converts textual and structured data into numerical vector representations, known as embeddings. This step allows semantic comparison between expressions even when they differ linguistically.

Each text segment t_i from the datasheet and each property s_j from ECLASS is mapped into an n-dimensional vector space through the embedding function:

$$E_T: T \to \mathbb{R}^n, E_S: S \to \mathbb{R}^n$$

where $E_T$and $E_S$represent the encoder functions for the datasheet and the semantic model, respectively.

Two embedding methods were implemented: The cloud-based pipeline employs the OpenAI model "text-embedding-3-large", which offers high semantic precision and context sensitivity. The local variant uses the Sentence Transformers library, specifically the model "all-mpnet-base-v2", which generates 768-dimensional embeddings and runs entirely offline. Both approaches use cosine similarity as their underlying measure of semantic proximity. Embeddings were selected over keyword-based methods because they capture conceptual meaning rather than literal matching. This enables the system to identify equivalent terms such as "pmax", "maximum pressure", and "max. operating pressure", even when phrased differently across manufacturers.

4.4. Semantic Alignment and Reasoning

In this stage, the embedded datasheet segments and ECLASS properties are compared within the shared vector space to identify semantic correspondences.
The similarity between a text segment $t_i$and an ECLASS property $s_j$is computed using the cosine similarity:

$$\text{sim}(a, b) = \frac{a \cdot b}{\| a \| \| b \|}$$

where $a = E_T(t_i)$ is the embedding vector of the text segment, and $b = E_S(s_j)$ is the embedding vector of the standardized property.
A higher value of $\text{sim}(a, b)$indicates a stronger semantic relation between both expressions. For each property, the top-k most similar text segments (typically $k = 3$-5) are selected as candidates for interpretation. Cosine similarity was chosen because it is scale-invariant and well suited for comparing high-dimensional sentence embeddings, providing more robust matches than keyword-based methods. The selected text segments are then processed by a LLM to extract structured data such as values, units, and contextual notes. This reasoning

step converts semantically aligned evidence into standardized outputs. In the cloud-based setup, reasoning is performed by OpenAI's gpt-4o-mini, while the local variant uses Meta-Llama-3-8B-Instruct with quantization for efficiency. Prompts follow a fixed JSON schema, ensuring syntactic consistency and easy post-processing. Together, the embedding and reasoning stages combine semantic precision with contextual understanding, enabling accurate and verifiable mappings between unstructured datasheets and standardized interoperability models.

4.5. Output

The workflow produces structured, machine-readable outputs that capture both the extracted data and its traceability. All results are stored as JSON files, each entry containing the standardized property, extracted value, physical unit, contextual notes, and a link to the corresponding page in the original datasheet. This design ensures that every mapping remains verifiable and can be reused for further processing.

4.6. Case Study

To evaluate the proposed framework, a case study was conducted using several publicly available datasheets from the fluid power domain. The documents originate from different manufacturers and describe hydraulic components using heterogeneous terminologies and specification formats. All datasheets were processed using the workflow described in Sections 4.1- 4.5. While multiple datasheets were used to assess general robustness, the quantitative evaluation focuses on one representative reference datasheet for which a fully manually curated ground truth mapping to ECLASS properties was available. This reference document comprises 114 technical features and enables a transparent and reproducible assessment of mapping quality.

Table 1 presents correctly mapped examples extracted from this reference datasheet, illustrating successful semantic alignment of heterogeneous fluid power specifications, where the data extraction and mapping were performed using an OpenAI-based variant (gpt-4o-mini).

Extracted Feature	**Assigned ECLASS Feature**	**ECLASS ID**	**Similarity Score**
Maximum operating pressure: 250 bar	Maximum operating pressure	0173-1#02-BAF016#005	0.90
Displacement volume: 45 cm^3/rev	Geometric displacement	0173-1#02-AAO677#004	0.85
Minimum ambient temperature: −20 °C	Minimum ambient temperature	0173-1#02-AAZ778#003	0.87

Table 1. Selected examples illustrating correct semantic alignment of heterogeneous fluid power datasheet descriptions with standardized ECLASS properties.

During preprocessing, textual content was extracted and normalized. Semantic embeddings were generated and compared with ECLASS feature embeddings using cosine similarity, and selected candidates were structured into feature–value representations by the reasoning

stage. A qualitative comparison with the expert reference confirmed that most mappings were technically plausible.

In addition to the qualitative assessment, the similarity scores obtained during the retrieval stage were analyzed in relation to the manually curated ground truth. Across all 114 mapped features, the average cosine similarity was 0.72. It was observed that correctly mapped features generally exhibited higher similarity values, while incorrect mappings occurred predominantly at lower similarity levels. Mappings with similarity scores above approximately 0.85 were found to correspond almost exclusively to correct semantic alignments with the expert reference. This value was not predefined as a hard decision threshold but emerged empirically from the evaluation as a high-confidence region, indicating stable semantic correspondence between heterogeneous datasheet descriptions and standardized ECLASS properties.

Despite the overall high mapping quality, mismatches and omissions were observed in specific cases. These can be attributed to the following factors:

- **Complexity of data extraction from industrial datasheets:** Fluid power datasheets often contain complex layouts, diagrams, characteristic curves, and hydraulic schematics that are not reliably captured by text-based extraction. Relevant features may therefore be partially extracted or missed before semantic mapping.

- **Variability in LLM reasoning performance:** Closely related technical concepts may be misinterpreted when contextual data is weak. A recurring example observed during evaluation is the confusion between nominal pressure and maximum operating pressure. Such cases were identified during qualitative comparison with the reference mapping and are not reflected in the correctly mapped examples shown in Table 1.

Overall, the case study demonstrates that automated semantic alignment of heterogeneous fluid power datasheets is feasible using the proposed workflow, while highlighting data extraction and reasoning robustness as key areas for further improvement.

5. DISCUSSION

This chapter discusses the main outcomes of the proposed approach, emphasizing its advantages, limitations, and validation perspective in the context of industrial data interoperability.

5.1. Advantages

The proposed workflow demonstrates that semantic data alignment in industrial domains can be automated largely without sacrificing interpretability. By embedding both datasheet content and ECLASS descriptors into a shared semantic space, the system can automatically recognize and assign technical data such as pressure, flow rate, or mounting type to standardized properties. This capability replaces a traditionally manual, time-intensive process with a consistent and reproducible mapping strategy. As a result, human involvement shifts from repetitive data entry to higher-level validation and quality assurance, which substantially increases efficiency and reduces the likelihood of human error.

Another notable strength lies in the method's scalability and robustness. Because the matching process operates on semantic similarity rather than keyword overlap, it generalizes well across heterogeneous datasheet formats, varying manufacturer terminology, and even multilingual documentation. The integration of ECLASS as a reference ontology ensures that the resulting data is compatible with AAS and digital twin implementations, thereby supporting interoperable data exchange across different systems and organizations.

Equally important is the flexible system architecture, which supports both cloud-based and local execution. The OpenAI-based configuration provides high inference speed and strong contextual understanding, making it suitable for rapid prototyping or large-scale processing. In contrast, the local pipeline based on Meta-Llama-3-8B-Instruct enables confidential, fully offline deployment, which is essential in regulated or security-critical industrial settings. Together, these two configurations demonstrate the approach's adaptability to different operational environments and data governance requirements.

5.2. Challenges

Despite the advantages of the proposed workflow, several challenges remain that affect its reliability. The most critical limitation is not the semantic mapping with LLMs itself, but the extraction of technical data from industrial datasheets. In practice, datasheets often contain complex layouts, scanned content, diagrams, characteristic curves, and hydraulic schematics that convey essential data but are not directly accessible through text-based extraction methods. As a result, relevant data may be partially extracted or missed entirely, which directly impacts all downstream processing steps.

Although preprocessing techniques such as OCR, normalization, and recursive text segmentation alleviate some of these issues, non-textual and weakly structured content remains a major source of errors. Improving document understanding at this stage therefore represents a key prerequisite before further improvements through larger or more advanced reasoning models can be expected.

Beyond data extraction, interpretive behaviour of LLMs constitutes an additional challenge. While larger, higher-parameter models generally improve contextual understanding and reasoning stability, they do not fully eliminate misinterpretations when technical descriptions are ambiguous. Moreover, increased model size introduces trade-offs in inference latency, computational cost, and deployment feasibility, particularly for fully local industrial deployments.

Finally, limitations in the coverage of existing interoperability schemas such as ECLASS can lead to unmapped attributes when manufacturer-specific or highly specialized parameters are encountered, highlighting the need for flexible hybrid approaches and future schema extensions.

5.3. Comparison and Validation Perspective

When compared to traditional manual annotation, the proposed workflow improves scalability and reduces processing effort. While manual mapping can achieve high precision for small datasets, it does not scale to large component portfolios or frequently updated documentation. The presented approach enables automated mappings that can be reviewed by domain experts in a human-in-the-loop setup, providing a practical balance between automation, reliability, and reduced manual effort.

Two deployment variants were evaluated: a cloud-based implementation using the OpenAI model gpt-4o-mini and a fully local open-source implementation based on Meta-Llama-3-8B-Instruct. The quantitative evaluation focuses on one representative reference datasheet with a complete manually curated ground truth comprising 114 technical features. Table 2 summarizes the quantitative comparison of both variants, including feature coverage, semantic similarity, inference time, and deployment-related aspects relevant for industrial application.

Criterion	OpenAI-based Variant (gpt-4o-mini)	Local Open-Source Variant (Meta-Llama-3-8B-Instruct)
Evaluated datasheet	1 reference datasheet	1 reference datasheet
Mapped features	114/114	114/114
Inference time per attribute	2.95s	5-10s
Total inference time per datasheet	5min 36s	10-19 min
Average Semantic Similarity	0.72	0.66-0.69
Data sovereignty	Requires external data exchange	Fully offline and confidential
Scalability	Easily scalable via cloud deployment	Limited by on-premise capacity

Table 2. Quantitative comparison of cloud-based and local deployment variants. Inference times reflect end-to-end execution and depend on deployment constraints.

Both deployment variants successfully mapped all 114 features contained in the reference datasheet. Mapping quality was assessed using three metrics. First, feature coverage was measured as the number of successfully detected and mapped features, resulting in full coverage for both variants. Second, semantic similarity was evaluated using cosine similarity during the retrieval stage. The cloud-based implementation achieved an average similarity score of 0.72, while the local implementation achieved values between 0.66 and 0.69. The observed differences in average semantic similarity between the cloud-based and local deployment variants are primarily caused by differences in the employed embedding models and execution configurations. The cloud-based variant utilizes higher-capacity embedding models with greater contextual sensitivity, while the local variant relies on smaller, quantized models optimized for offline execution. This results in slightly lower average similarity scores without fundamentally affecting mapping correctness for well-defined features.

Third, all generated mappings were compared against the expert reference. For clearly defined features, both implementations produced equivalent mappings, whereas differences primarily occurred for weakly specified or context-dependent attributes.

Correct mappings in both deployment variants were consistently associated with higher similarity values, while misassignments predominantly occurred in intermediate similarity ranges, particularly for closely related technical concepts such as nominal and maximum operating pressure. For the OpenAI-based configuration, similarity scores above approximately 0.85 emerged as an empirically observed high-confidence region for the evaluated

reference datasheet. This observation is model- and dataset-dependent and should not be interpreted as a strict or generally valid threshold. Future work will extend the evaluation to larger and more diverse datasets to further analyze similarity distributions and derive more robust confidence estimates.

Differences in inference time are primarily attributable to the execution environment rather than the semantic mapping methodology itself. The cloud-based variant benefits from highly optimized infrastructure, parallelized inference, and optimized model serving, whereas the local open-source variant prioritizes offline capability and data sovereignty and was executed on on-premise hardware with limited computational resources. Consequently, longer processing times represent an expected trade-off rather than a methodological limitation.

Overall, both deployment variants achieve comparable mapping quality for well-defined features. Differences are mainly observed in processing time and sensitivity to weakly specified input, indicating that stable semantic alignment depends not on model size alone but on the combination of structured preprocessing, retrieval-based context, constrained model interaction, and expert validation.

6. Conclusion and Outlook

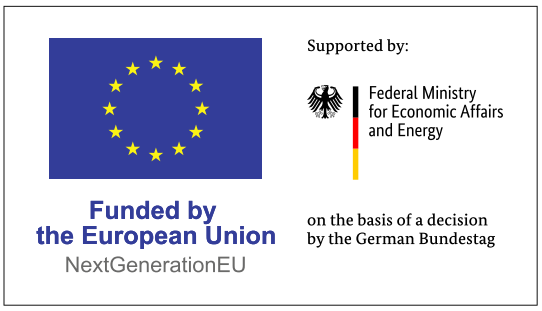
This paper demonstrated that LLMs can support semantic interoperability in the fluid power domain, where heterogeneous datasheets and inconsistent terminology continue to limit automated data integration. By combining embedding-based similarity search with structured reasoning and standardized vocabularies, the proposed workflow enables the transformation of existing manufacturer documentation into machine-readable representations suitable for digital engineering applications. The modular architecture allows the approach to be applied in different deployment scenarios. A cloud-based implementation supports scalable processing with low inference times, while a fully local configuration enables offline operation and compliance with data protection requirements, which are common constraints in industrial environments.

At the same time, the evaluation highlights that the main limitation is not the semantic mapping itself, but the extraction of technical data from industrial datasheets. Complex layouts, diagrams, characteristic curves, and schematic representations remain challenging and directly affect the quality of downstream semantic alignment. Future work should therefore prioritize improved data extraction and document understanding, including multimodal approaches, before extending the reasoning models or scaling the framework further. Additional integration with established interoperability frameworks such as the AAS may further support practical adoption.

Overall, the presented approach provides a feasible step toward improving semantic interoperability in fluid power domain by systematically bridging unstructured technical documentation and standardized digital representations, contributing to more consistent and reusable industrial data in Industry 4.0 contexts.

7. ACKNOWLEDGEMENT

This work has been funded by the European Union (EU) and supported by the German Federal Ministry for Economic Affairs and Energy (BMWE) in the project Fluid4.0 – Revolutionizing Fluid Engineering: Implementation of digitalization for fluid power 4.0 in the cross-industry and cross-manufacturer data space using Asset Administration Shells, submodels and demonstrators. (project number 13IK039L). The views and opinions expressed are solely those of the authors and do not necessarily reflect the views of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.



8. REFERENCES

[1] Bauernhansl, T. 'Handbuch Industrie 4.0', Springer Berlin Heidelberg, Berlin, Heidelberg, 2023. DOI: 10.1007/978-3-662-58532-0.

[2] Cartus, A. 'Interoperability of semantically heterogeneous digital twins through Natural Language Processing methods', https://proceedings.open.tudelft.nl/clima2022/article/view/143. DOI: 10.34641/clima.2022.143.

[3] Murrenhoff, H. '11th International Fluid Power Conference - 19th-21st of March 2018, Aachen, Germany', RWTH Aachen University, http://www.ifk2018.com/frontend/index.php_folder_id=470.html, Aachen, 2018.

[4] Rahal, J. R., Schwarz, A. et al. 'The asset administration shell as enabler for predictive maintenance: a review', Journal of Intelligent Manufacturing, Vol. 36, Nr. 1, pp. 19–33, 2025. DOI: 10.1007/s10845-023-02236-8.

[5] Zhang, J., Ellwein, C. et al. 'Digital twin and the asset administration shell', Software and Systems Modeling, Vol. 24, Nr. 3, pp. 771–793, 2025. DOI: 10.1007/s10270-024-01255-0.

[6] Salari, A. 'Details of the Asset Administration Shell - Part 2', https://digitalcollection.zhaw.ch/server/api/core/bitstreams/35301a19-fade-444f-bb40-8975bad47f48/content, 2021.

[7] Zhao, W. X., Zhou, K. et al. 'A Survey of Large Language Models', arXiv, 2023. DOI: 10.48550/arXiv.2303.18223.

[8] Höltgen, L., Zentgraf, S. et al. 'Utilizing large language models for semantic enrichment of infrastructure condition data: a comparative study of GPT and Llama

models', AI in Civil Engineering, Vol. 4, Nr. 1, 2025. DOI: 10.1007/s43503-025-00055-9.
[9] Minaee, S., Mikolov, T. et al. 'Large Language Models: A Survey', arXiv, 2024. DOI: 10.48550/arXiv.2402.06196.
[10] Zhao, P., Zhang, H. et al. 'Retrieval-Augmented Generation for AI-Generated Content: A Survey', arXiv, 2024. DOI: 10.48550/arXiv.2402.19473.
[11] Both, M. 'Semantische Interoperabilität digitaler Zwillinge als Basis für automatisierte Erkundungsmechanismen', 2024. DOI: 10.25673/116680.
[12] Technische Regel: 'DIN SPEC 91345:2016-04, Referenzarchitekturmodell Industrie 4.0 (RAMI4.0)', DIN Media GmbH, Berlin. DOI: 10.31030/2436156.
[13] Wei, K., Sun, J. Z., Liu, R. J. 'A Review of Asset Administration Shell', 2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), pp. 1460–1465, 2019. DOI: 10.1109/IEEM44572.2019.8978536.
[14] Xia, Y., Xiao, Z. et al. 'Generation of Asset Administration Shell With Large Language Model Agents: Toward Semantic Interoperability in Digital Twins in the Context of Industry 4.0', IEEE Access, Vol. 12, pp. 84863–84877, 2024. DOI: 10.1109/ACCESS.2024.3415470.
[15] ECLASS: 'Fluid power', https://eclass.eu/eclass-standard/content-suche/, 2026.
[16] IEC 61360: 'Datenbank für elektrische/elektronische Bauteile', https://www.vde-verlag.de/iec-normen/iec-datenbanken/iec-61360-elekronische-bauteile.html, 2026.
[17] Vaswani, A., Shazeer, N. et al. 'Attention Is All You Need', arXiv, 2017. DOI: 10.48550/arXiv.1706.03762.
[18] Naveed, H., Khan, A. U. et al. 'A Comprehensive Overview of Large Language Models', arXiv, 2023. DOI: 10.48550/arXiv.2307.06435.
[19] Gao, Y., Xiong, Y. et al. 'Retrieval-Augmented Generation for Large Language Models: A Survey', arXiv, 2023. DOI: 10.48550/arXiv.2312.10997.
[20] Alt, R., Schmitz, K. 'Basic requirements for Plug-and-Produce of 14.0 fluid power systems', 2019 IEEE 8th International Conference on Fluid Power and Mechatronics (FPM), pp. 1440–1448, 2019. DOI: 10.1109/FPM45753.2019.9035813.
[21] Vaswani, A., Shazeer, N. et al. 'Attention Is All You Need', arXiv, 2017. DOI: 10.48550/arXiv.1706.03762.
[22] Zeid, A., Sundaram, S. et al. 'Interoperability in Smart Manufacturing: Research Challenges', Machines, Vol. 7, Nr. 2, pp. 21, 2019. DOI: 10.3390/machines7020021.
[23] Bellis, A. de: 'Structuring the unstructured: an LLM-guided transition', https://iswc2023.semanticweb.org/wp-content/uploads/2023/11/ISWC2023_paper_376.pdf.
[24] Wang, J., Li, Y. et al. 'Unify: A System For Unstructured Data Analytics', Proceedings of the VLDB Endowment, Vol. 18, Nr. 12, pp. 5287–5290, 2025. DOI: 10.14778/3750601.3750653.

Biographies

Barkha Sharma received a bachelor's degree in computer science from the University of Stuttgart in 2021 and a master's degree in computer science from the University of Stuttgart in 2025. She is a Research Associate in the Smart Systems group at the Institute for Fluid Power Drives and Systems (ifas) at RWTH Aachen University. Her research areas include machine learning, especially Large Language Models and AI planning, as well as data analysis.

Malte Becker obtained his Master of Science degree in Mechanical Engineering from RWTH Aachen University. Since 2021, he has been working as a research associate at the Institute for Fluid Power Drives and Systems (ifas) at RWTH Aachen University. His research focuses on the Industrial Internet of Things (IIoT) and Systems Engineering. Since 2025, he has been serving as Chief Engineer at ifas.

Prof. Katharina Schmitz studied mechanical and chemical engineering at RWTH Aachen University and Carnegie Mellon University, Pittsburgh (USA) and graduated 2015 as Dr.-Ing. at RWTH Aachen University. Since 2018, she is full professor at RWTH Aachen University and director of the Institute for Fluid Power Drives and Systems (ifas). In addition, she is Vice Dean of the Faculty for Mechanical Engineering at RWTH Aachen, a position she holds since 2020. Prof. Schmitz's awards and honors include several best paper awards and 2023 IMechE Joseph Bramah Medal award.