Automated Construction and Semantic Interoperability for Digital Twins: Integrating Heterogeneous Data with Large Language Models

Leonardo Pilarski[1], Luiz E. Luiz[1,2], Gonçalo S. Gomes[1], Tiago Pinto[1,3],
Vitor M. Filipe[1,3], João Barroso[1,3] and Gonçalo Rijo[4]
[1]*School of Science and Technology, University of Trás-os-Montes and Alto Douro*, 5000-801, Vila Real, Portugal
[2]*CeDRI, SusTEC, Instituto Politécnico de Bragança*, 5300-253 Bragança, Portugal
[3]*INESC-TEC, UTAD Pole*, 5000-801, Vila Real, Portugal
[4]*Neoception, Unipessoal, Lda.*, 5000-801, Vila Real, Portugal
{pilarski, tiagopinto, vfilipe, jbarroso}@utad.pt, al2024157754@alunos.utad.pt, luiz.luiz@ipb.pt, grijo@neoception.com

2025 IEEE Conference on Artificial Intelligence (CAI) | 979-8-3315-2400-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/CAI64502.2025.00014

***Abstract*—Digital twins are increasingly used, as they allow the creation of detailed virtual representations of physical products and systems. They face, however, significant challenges such as heterogeneous data integration and high costs. This article presents an innovative methodology that uses Large Language Models to unify information and automate the generation of Digital Twin models. The proposal comprises several modules, covering the stages of data collection, semantic processing, modular construction and validation of the Digital Twin. In this way, the proposed model guarantees interoperability, efficiency and scalability for various domains.**

***Index Terms*—Asset Administration Shell, Semantic standardization, Industrial Automation and Artificial Intelligence.**

I. Introduction

The concept of Digital Twin (DT) stands out in Industry 4.0, allowing the creation of detailed virtual representations of products and physical systems for monitoring, simulation and decision-making support [1], [2]. These models can integrate data from sensors, technical drawings and other sources in structured or unstructured formats, allowing efficient operations [5]. Despite this potential, the structuring of DT faces significant challenges. Integrating data from multiple domains and in heterogeneous formats is a complex process that requires high costs and manual effort, limiting scalability and applicability in diverse environments [3], [5].

Large Language Models (LLMs), such as OpenAI ChatGPT, offer an approach to overcoming traditional limitations by understanding data through vector numerical representations, using embeddings and transformers to convert textual information into numbers. These models stand out for their ability to process a variety of data, such as images, texts and unstructured data, including texts, informal records and others, and to extract complex semantic relationships, understanding both the literal meaning and the context of words [3], [5]. LLMs can identify patterns, infer connections between data from different sources and formats, and correlate visually distinct but semantically similar information, such as sensor data, operational reports and technical documents for use in DTs [3], [5]. In addition, LLMs provide gains in efficiency, automation and scalability, adapting to the needs of different industrial sectors and promoting more intelligent and fluid data integration [6].

This article proposes a novel methodology based on LLM to automate the structuring of DT. The approach emphasizes the ability of these models to identify and infer relationships between distinct datasets [6] and integrate information in multiple formats, creating interoperable digital representations [3]. By addressing identified literature gaps, this proposal seeks to broaden the scope and applicability of DT, promoting its adoption in different industrial and technological contexts.

II. State of the Art and Beyond

Both et al. [4], there are difficulties in scalability and automated protocols. At the same time, [3], [5] highlights using Asset Administration Shells (AAS) and LLMs for automation and industrial organization, although they still face barriers with unstructured data and creating efficient interfaces. In [1], [2], the authors reinforce gaps in standards and implementation costs, indicating the need for scalable and interoperable solutions. Despite advances with AAS in modular production, challenges remain in modularity and multi-sector integration [5], [6].

The proposed solution combines LLMs and AAS to create an automated pipeline for generating DT. The approach uses LLMs to interpret heterogeneous information and generate standardized semantic models, reducing manual effort and increasing scalability. Integration with AAS guarantees interoperability between industrial systems, while LLM's continuous learning capability allows them to adapt in real-time.

III. Methodology

The proposed solution, described in the flowchart in Fig. 1, addresses the gaps identified using a methodology presented in the subsections on III-A, III-B, III-C and III-D. The aim is to create an automated tool that presents the DT, adapting

979-8-3315-2400-5/25/$31.00 ©2025 IEEE
DOI 10.1109/CAI64502.2025.00014

it according to the information provided and ensuring heterogeneous data integration, automation, and interoperability. LLM is fundamental to the intelligence and automation of the process.

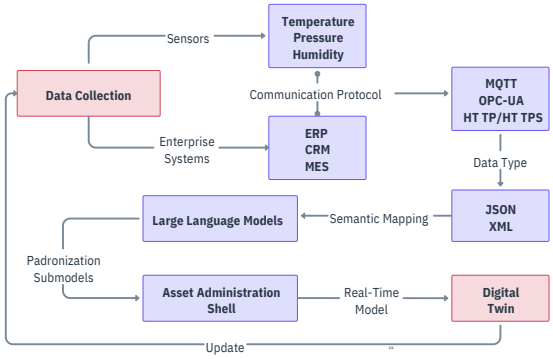

Fig. 1. Flowchart of DT deployment life cycle, integrating LLM and AAS.

A. Data Collection

Creating a digital twin begins with collecting data from various sources, such as sensors and business systems. The data is transmitted using various communication protocols, ensuring a continuous data flow. Some of the main protocols are MQTT, which collects data from sensors in real-time, such as temperature, pressure and humidity; OPC-UA, which integrates industrial systems, such as machines, and HTTP/HTTPS, which connects data from corporate systems, such as ERP (Enterprise Resource Planning), CRM (Customer Relationship Management) and MES (Manufacturing Execution System), providing information on production, stock and customer interactions. This data can include physical variables (such as temperature) and operational information (such as production status), forming the basis for creating an accurate and functional digital replica.

B. Semantic Standardization of Data

After data has been collected, it needs to be standardized to be processed efficiently and integrated into Digital Twin. During this phase, data collected from various sources may have different formats, such as JSON or XML, and needs to be converted into a single, consistent format. This is where Large Scale Language Models (LLMs) come into play, performing the semantic mapping between data from different sources. LLMs can interpret unstructured data, such as reports or texts, and transform it into structured and consistent formats, ensuring that different terms (e.g. “temp” and “temperature”) are correctly mapped and that units of measurement are normalized.

C. Submodels and Asset Administration Shell

Since the data is standardized, it is organized into submodels, where the Asset Administration Shell (AAS) acts as a centralized and standardized repository for storing information about the assets and their operating conditions. The AAS integrates different data types, such as sensor information, operational status and maintenance history, into a modular and interoperable model, allowing Digital Twin to be built scalable and efficiently. Each sub-model represents a specific aspect of the asset, such as sensor data (temperature, pressure, humidity), operational status (on/off, under maintenance) and maintenance history (interventions carried out, parts replaced), ensuring that the digital replica is accurate, flexible and capable of evolving in line with the needs of the physical system.

D. Digital Twin

Finally, the organized data and submodels are integrated to create the Digital Twin, a real-time digital replica of the physical asset or system. The Digital Twin is continuously fed with real-time data from the AAS, ensuring that the digital model is always up to date.

IV. Final Remarks

The integration of LLM with AAS offers a transformative approach to automating the structuring of DT, addressing challenges such as data heterogeneity and interoperability. By taking advantage of the semantic capabilities of LLM and the modular flexibility of AAS, this methodology may enhance efficiency, scalability, and adaptability for various industrial contexts.

Acknowledgment

The study was developed under the project A-MoVeR – “Mobilizing Agenda for the Development of Products & Systems towards an Intelligent and Green Mobility”, operation n.º 02/C05-i01.01/2022.PC646908627-00000069, approved under the terms of the call n.º 02/C05-i01/2022 – Mobilizing Agendas for Business Innovation, financed by European funds provided to Portugal by the Recovery and Resilience Plan (RRP), in the scope of the European Recovery and Resilience Facility (RRF), framed in the Next Generation UE, for the period from 2021 -2026.

References

[1] M. Segovia and J. Garcia-Alfaro, “Design, modeling and implementation of digital twins,” Sensors, vol. 22, no. 14, p. 5396, 2022.
[2] D. M. Botín-Sanabria, A.-S. Mihaita, R. E. Peimbert-García, M. A. Ramírez-Moreno, R. A. Ramírez-Mendoza, and J. Lozoya-Santos, “Digital twin technology challenges and applications: A comprehensive review,” Remote Sensing, vol. 14, no. 6, p. 1335, 2022.
[3] Y. Xia, Z. Xiao, N. Jazdi, and M. Weyrich, “Generation of Asset Administration Shell with Large Language Model Agents: Towards Semantic Interoperability in Digital Twins in the Context of Industry 4.0,” IEEE Access, 2024.
[4] M. Both, B. Kämper, A. Cartus, J. Beermann, T. Fessler, J. Müller, and C. Diedrich, “Automated monitoring applications for existing buildings through natural language processing based semantic mapping of operational data and creation of digital twins,” Energy and Buildings, vol. 300, p. 113635, 2023.
[5] Y. Xia, N. Jazdi, and M. Weyrich, “Automated generation of Asset Administration Shell: a transfer learning approach with neural language model and semantic fingerprints,” in *Proc. 2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA)*, pp. 1–4, 2022.
[6] Y. Xia, M. Shenoy, N. Jazdi, and M. Weyrich, “Towards autonomous system: flexible modular production system enhanced with large language model agents,” in *Proc. 2023 IEEE 28th International Conference on Emerging Technologies and Factory Automation (ETFA)*, pp. 1–8, 2023.