RODEOS — Robotic Data Ecosystem Semantic Model: Bridging Siloed Robot Systems

Maximilian Stäbler[1,*,†], Lukas Sohlbach[2,†], Felix Weidinger[2,†], Steffen Turnbull[1,†], Jorge Marx-Goméz[3], Chris Schlueter-Langdon[4] and Frank Köster[1]

[1]German Aerospace Center (DLR), Institute for AI Safety and Security, Wilhelm-Runge-Straße 10, 89081 Ulm, Germany
[2]VDMA Robotics + Automation, Lyoner Straße 18, 60528 Frankfurt am Main, Germany
[3]University of Oldenburg, Department of Business Informatics, Ammerländer Heerstraße 114-118, 26129 Oldenburg, Germany
[4]Drucker School of Management, Claremont Graduate University, 150 E 10th St, Claremont, CA 91711, USA

Abstract

Industry robotics across automotive paint shops, pharmaceutical clean-rooms, and e-commerce warehouses still relies on bespoke data mappings: every new robot arrives with proprietary file formats, capability vocabularies, and safety descriptors, forcing engineering teams into multi-week manual integration cycles. To address this cost and agility gap, we introduce *RODEOS—**RO**botic **D**ata **Ec**O**system** **S**emantic Model*, an emerging, vendor-neutral semantic blueprint co-defined by a consortium of 24 industrial partners. RODEOS extends the W3C DCAT-3 core with robotics-specific classes for raw data, model assets, and executable services, while preserving the lightweight authoring demands voiced in interviews with 17 experts drawn from research institutes, industrial end-users, robotics and automation suppliers, system integrators, IT-infrastructure providers, and the machinery-industry association—thereby capturing a truly holistic cross-domain requirements profile. We conduct a qualitative ablation study comparing JSON schemas generated by an LLM with and without RODEOS schema constraints, finding that schema-guided generation improves precision, coverage, and consistency of the output. These contributions—(i) the RODEOS semantic model, (ii) an LLM-assisted authoring workflow, and (iii) initial empirical validation metrics—aim to accelerate robot-cell integration and lay the groundwork for a community-wide semantic robotics ecosystem.

Keywords

industrial robotics, semantic interoperability, LLM-assisted semantic engineering, vendor-neutral semantics

1. Introduction and Motivation

Industrial robotics today spans highly diverse environments—from automotive paint shops and pharmaceutical clean-rooms to e-commerce warehouses—yet each new robot, gripper, or vision sensor still arrives with proprietary file formats, capability vocabularies, and safety descriptors [1]. Integrators must, therefore, create bespoke data mappings for every installation. This process prolongs ramp-up by an estimated six to eight weeks per cell and drives up engineering costs long before a single product leaves the line [1]. Existing workarounds such as spreadsheet templates and pair-wise converters provide only transient relief because they must be rebuilt whenever a new product variant appears, regulations change, or a different supplier is introduced [2]; worse, these ad-hoc artifacts offer no guarantee of semantic consistency across plants or organizational borders, turning data exchange into an administrative rather than an engineering task. Motivated by this problem landscape, the twenty-four organizations in the RoX [3] consortium—including research institutes, robot and automation vendors, system integrators, industrial end-users, IT infrastructure providers, and the machinery-industry association—jointly articulated a demand for a domain-agnostic yet extensible semantic layer that can be authored and maintained by non-ontologists. We report an ablation analysis highlighting the impact of schema constraints on LLM-generated outputs in Section 2.

ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan
*Corresponding author.
†These authors contributed equally.
maximilian.staebler@dlr.de (M. Stäbler)
0000-0003-1311-3568 (M. Stäbler); 0000-0002-7833-7549 (J. Marx-Goméz)

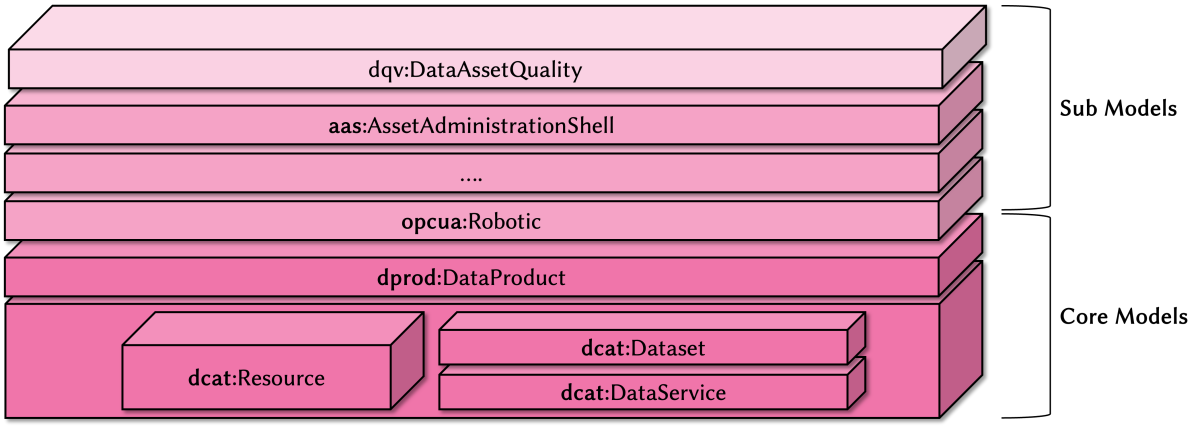

Figure 1: RODEOS stack: a DCAT core (Resource ⟶Dataset, DataService) extended by a DataProduct layer and pluggable robotics sub-models (e.g., Asset-Administration-Shell (AAS)). Full model on GitHub.

2. Survey Insights and RODEOS Blueprint

A structured survey of ten RoX organisations—covering research, automation vendors, system integrators, end-user factories, IT providers, and the machinery association—yielded ten expert interviews, all with individuals in either end-user or system integrator roles. The questionnaire probed data assets, roles, technical readiness, modelling needs, and views on LLMs, with an open slot for additional remarks. Three trained interviewers recorded each session, AI-transcribed the audio and fused transcripts with field notes into concise summaries that the interviewees validated. The interviews identify manual configuration as the main integration bottleneck and highlight automated semantic tooling as the consortium's top priority.

Data-quality safeguards. Potential threats—interviewer bias, transcription errors, and information loss—were mitigated by a common protocol and joint training, manual transcript checks against notes, and member checking of the summaries.

Accepted asset taxonomy. Ninety-five percent of the interviewees endorsed a three-way classification of *assets* into **(raw) data**, **models**, and **services**. *RawData* encompasses sensor logs, trajectory traces, inspection images, and non-technical data (i.e., PDF); *model* refers to kinematic graphs, CAD files, or simulation meshes; *Service* encompasses executable artifacts, including motion skills, perception pipelines, and safety checks. Partners further suggested an optional *HardwareDescriptor* for Automated Guided Vehicles (AGV) and sensors, but agreed that the core taxonomy suffices for an initial release. Most companies focus on internal use of their data, while acknowledging that cross-company exchange remains too manual to scale efficiently.

Requirements distilled from the survey. Interview feedback converged on four design imperatives: (i) an *extensible* semantic kernel aligned with the W3C DCAT-3 catalogue vocabulary [4]; (ii) robotics-specific *submodels* for *Capability*, *Skill*, *Task*, *Risk* and *SafetyPolicy*; (iii) provenance metadata, role-based access control, and deployment variants (cloud, on-prem) to satisfy security and governance constraints; (iv) resilience to change: frequent robot or process upgrades must be accommodated by hot-swapping submodels rather than editing the core ontology.

RODEOS architecture. The resulting *Robotic Data Ecosystem Semantic Model (RODEOS)* therefore adopts a layered approach as shown in Figure 1. A *DCAT-based core model* captures universal metadata—identifier, license, version, provenance—while domain-specific submodels import established standards such as OPC UA information models, Asset Administration Shell (AAS) sub-shells, URDF for kinematics or ISO safety taxonomies. Core and submodels can be combined ad hoc; the complete description is serialized as a self-contained JSON document, ready for exchange between engineering tools, digital twins, and runtime systems. Experts have raised concerns about uncontrolled vocabulary drift and ambiguous references in ad-hoc LLM-generated JSON structures. A formal schema mitigates these risks by enforcing consistent identifiers and enabling governance across evolving systems.

LLM-assisted authoring workflow. The survey revealed that almost all domain and application experts across the consortium had *no prior experience* with ontology engineering; nevertheless, they can describe their use cases, data flows, and system boundaries eloquently in free text. To convert these narratives into formal structure without burdening experts with OWL syntax, we prototype a "semantic assistant'' powered by Retrieval-Augmented Generation (RAG) [5]. Partner documents and the predefined RODEOS schema are indexed. When an expert submits a plain-language prompt, the assistant retrieves relevant fragments aligns them with the *(raw) data*, *model*, or *service* taxonomy, and proposes candidate classes and properties. Regulated sectors can run the workflow on EU-hosted or on-premises open-source models, while others may leverage proprietary LLMs via enterprise licenses. RODEOS, therefore, ships with pluggable language-model backends, providing a scalable, automated path from natural-language descriptions to consistent semantic artifacts.

Ablation Study. A stationary industrial robot equipped with a pneumatic gripper performs pick–and–place while a colour camera conducts in-line quality inspection. The cell exposes three digital assets: (i) the raw RGB image stream, (ii) a CNN inference model detecting defects, and (iii) a REST service publishing inferred quality labels.

Method. We queried a GPT-4o class model in two modes: *(a) Prompt-only*, receiving only the natural-language scenario; *(b) Schema-guided*, receiving the same prompt *plus* the RODEOS core and sub-class property list (cf. Fig. 1). Both prompts asked for a JSON description of the cell.

Typical issues in prompt-only output. (i) misuse of `dcterms:title` and `dcterms:identifier`; (ii) hallucinated fields such as `imageResolution`; (iii) omission of mandatory properties (`dcat:contactPoint`); (iv) inconsistent naming (`modelType` vs. `typeOfModel`).

Improvements with schema constraints. The guided run produced a fully RODEOS-compliant graph: every resource inherited the correct `dcat:Resource` properties; the model entity contained `sis:modelType`, `sis:modelParameters`, and `sis:framework`; the service declared `sis:usedModels`, `sis:input`, and `sis:output`; the dataset specified both `sis:dataFormat` and `dprod:informationSensitivityClassification`. No out-of-schema fields appeared, and all cross-references resolved.

Roadmap. A public draft of the core model, along with initial palletizing, pose estimation, and robotic control system submodels, is planned for Q4 2025, followed by validation workshops in the automotive, logistics, and pharmaceutical verticals. Success will be measured, in a laboratory setup provided by the consortium, by the number of hours saved, the reduction of mapping defects, and the number of interface "patches" eliminated. These steps pave the way for a production-ready release that promises to shorten robot-cell integration lead times and unlock new, data-driven business models for the consortium and the wider robotics community.

3. Discussion, Conclusion and Outlook

The cross-domain survey—spanning automotive, logistics, pharma, and research—confirms an industry-wide need for a shareable yet lightweight semantic layer. Although not every conceivable requirement surfaced in the interviews, the aggregated, privacy-protected results (individual transcripts cannot be disclosed) reveal broad agreement on three asset types—*(raw) data*, *models*, and *services*—as the cornerstone of such a layer. RODEOS addresses this need with a DCAT-aligned core, domain-specific submodels, and an LLM-assisted authoring workflow that non-ontologists can extend and maintain. While the approach has been validated in a single-domain pilot, its generalizability to other domains remains to be evaluated. Cross-factory tests are planned to assess robustness and adaptability across heterogeneous industrial settings. Looking ahead, two research challenges remain: (i) defining governance patterns that satisfy high-security domains (defense, pharma) without stifling open innovation; and (ii) benchmarking LLM-generated artifacts against expert-crafted baselines. In parallel, each submodel must be rigorously validated against the relevant domain and application standards to ensure interoperability. By tackling these hurdles, RODEOS paves the way for faster integration, reduced engineering effort, and new data-driven business models—benefits that extend to the wider community.

Declaration of GenAI

During the preparation of this work, the author(s) used Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] D. Tola, E. Madsen, C. Gomes, L. Esterle, C. Schlette, C. Hansen, P. G. Larsen, Towards Easy Robot System Integration: Challenges and Future Directions, in: 2022 IEEE/SICE International Symposium on System Integration (SII), IEEE, Narvik, Norway, 2022, pp. 77–82. doi:10.1109/SII52469.2022.9708846.

[2] M. Noura, M. Atiquzzaman, M. Gaedke, Interoperability in Internet of Things: Taxonomies and Open Challenges, Mobile Networks and Applications 24 (2019) 796–809. doi:10.1007/s11036-018-1089-9.

[3] VDMA e.V. Robotics + Automation, RoX Enabling AI Robotics, https://www.project-rox.ai/, 2025.

[4] Riccardo Albertoni, David Browning, Simon J D Cox, Alejandra Gonzalez Beltran, Andrea Perego, Peter Winstanley, Data Catalog Vocabulary (DCAT) - Version 3, 2024.

[5] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, Q. Li, A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM, Barcelona Spain, 2024, pp. 6491–6501. doi:10.1145/3637528.3671470.