2025 IMEKO TC-6 International Conference on
Metrology and Digital Transformation - **M4DConf 2025**
Benevento, Italy, September 3-5, 2025

Supporting Medicines Manufacturing through Semantic Technologies

Nina Perić[1], Moulham Alsuleman[2], João Gregório[2], Paul Duncan[2], Michael Chrubasik[2]
[1] *Informatics, Data Science Department, National Physical Laboratory, Teddington, TW11 0LW, London, United Kingdom, nina.peric@npl.co.uk*
[2] *Informatics, Data Science Department, National Physical Laboratory, Glasgow G1 1RD, United Kingdom, moulham.alsuleman@npl.co.uk, joao.gregorio@npl.co.uk, paul.duncan@npl.co.uk, michael.chrubasik@npl.co.uk*

***Abstract* – Pharmaceutical manufacturing involves complex, highly regulated processes that generate large volumes of heterogenous data. While highly valuable, this data is often siloed across systems and not easily interoperable, limiting its useability for provenance, advanced analytics and AI as well as automated decision-making processes. This paper presents a semantic technology driven use case to address these issues through the integration of a pharma-based digital architecture with established domain ontologies to support intuitive, semantically rich queries over manufacturing data. The work presented lays the structural groundwork to integrate sensor and event-based observations, linking them to higher-level domain structures in the future using ontologies such as SSN and the Industrial Ontologies Foundry ontology stack, a key requirement for a fully digitalised, interoperable and linked industrial workflow promoting industry 4.0 principles.**

I. INTRODUCTION

The transition to Industry 4.0 and then Industry 5.0 is largely dependent on the digitalisation of manufacturing processes. Rather than seeing data as a simple byproduct, it is recognised as an essential element of more sustainable, automated, and sophisticated manufacturing processes. This transformation requires that all process data, including basic measurement information, should be contextualised, interoperable, and accessible across large and complex digital architectures (US Food and Drug Administration and others). Still, fragmentation of data and siloing persist in many industries, including pharmaceutical manufacturing, where it is often locked within isolated and proprietary systems or structured inconsistently across abstraction layers (International Society of Automation) (International Society of Automation) [4] [5]. This can prevent the implementation of modern analytics or AI-based workflows, as well as reasoning and traceability across systems intra and inter organisation, meaning that the intended efficiency and productivity gains of digitalisation are not being realised.

The National Physical Laboratory (NPL) has developed a case study in collaboration with the Medicines Manufacturing Innovation Centre (MMIC) operated by the Centre for Process Innovation (CPI). This case study focusses on how semantic technologies, specifically ontologies and knowledge graphs, can bridge the gap between business-level Conceptual Data Models (CDMs) and their technical implementation in Logical Data Models (LDMs) [5] [6], linking data and making knowledge more accessible using domain-specific language. This creates a foundation for data interoperability resulting in democratised data access. While the case study does not initially include specific measurement data, the architecture is deliberately built on modular, ontology-based standards (Basic Formal Ontology (BFO) [7], Industrial Ontologies Foundry Core Ontology (IOF-Core) [6], Supply Chain Reference Ontology (SCRO) [8] [9], Biopharmaceutical manufacturing ontology (NIIMBL) [10]) for enabling future integration of sensor-based [11] or quantitative measurement data [12] [13]. By implementing such a semantic layer, architectures enable the future integration of traceable [14], measurement-driven processes, for more robust analytics, compliance, and automation in line with metrological traceability.

Mapping CDMs to domain model ontologies also enables business professionals and domain experts to query LDMs using familiar language. This improves accessibility for non-technical stakeholders and enhances interoperability between various data sources and domains. The simplified process of linking and combining diverse datasets can potentially reveal new insights, significantly enhancing data management capabilities. This system allows for more efficient use of pharmaceutical manufacturing data across the MMIC and its partners, paving the way for more robust analytics, collaboration, compliance, and automation.

II. METHOD

This work aims to enhance data interoperability in pharmaceutical manufacturing, using a tablet manufacturing process as a case study, focusing on supply chains and internal material movements. By mapping terminology from conceptual data models to ontology

Table 1. Example of competency questions (CQs) and their expected answers.

ID	Competency	Competency Questions	Purpose of the question
1	Material Flow	For a process, which materials are used and in which quantities? Where are the final locations of the materials?	If two or more processes require the same material, can that material be stored in more than one location?
2	Scheduling	For all processes, what are the start and end times? What are the events associated with each process?	Processes may have multiple events or just one. How does this affect the start and end times of the process?
3	Material Availability	For all processes, which materials are required (specifically, which material lots)? What are the storage locations of specific material lots?	If a material is required in two processes, how do we check its availability? How do we monitor material availability if it is stored in multiple locations?

classes, the semantic layer acts as a translator between domain experts and technical infrastructure. This allows non-technical users to frame queries using familiar business terms without needing to understand the underlying data scheme. For example, a process engineer can retrieve information about material lots or job steps using domain-relevant language.

Example competency questions and their answers can be found in Table 1. Example of competency questions (CQs) and their expected answers. Competency questions were chosen based on discussions with domain experts, focusing on key tasks such as material tracking and scheduling. These discussions reflected real information needs and helped guide ontology development. Each question was tested using SPARQL queries based on ontology terms, to check that relevant and accurate answers could be retrieved—validating both the ontology structure and its practical value. To ensure the accuracy and reliability of the ontology, we generated a test dataset with predefined and controlled scenarios. This allowed us to validate the outputs of SPARQL queries against known expectations, thereby confirming the ontology's structure and its practical utility. The result was a clear list of connections that allowed logical data fields to be understood as formal ontology elements, making it easier to share information and ask questions in domain-specific language.

A. Ontology Review and Selection

A detailed examination of openly available and specialised ontologies was completed to find frameworks that effectively represent entities and relationships in structured data. The criteria for choosing ontologies included

- their importance to manufacturing (generally and pharmaceutical more specifically), logistics, and quality testing
- how well they fit with existing Conceptual and Logical Data Models
- their ability to support inferencing and SPARQL-based semantic querying
- their agreement with well-known upper ontologies—high-level, abstract frameworks that provide general concepts and relationships applicable across various domains, serving as a foundation for more specific domain ontologies to ensure interoperability and semantic consistency—such as BFO.

The selected ontologies, IOF-Core, SCRO, and NIIMBL ontology, help follow industry standards and improve how different systems understand each other.

B. Ontology Mapping

The mapping process involved aligning elements from the LDM's with classes and properties defined in the selected ontologies, exemplified in Tab. 2. This task required a detailed examination of schema definitions and ontology structures to ensure semantic consistency and coverage. System Architecture

To help combine ontologies with MMIC's systems, a thorough study of the existing underlying digital architecture was completed. This architecture comprises two layers: conceptual and physical. The physical layer provides technical specifications of data classes, attributes, relations, and constraints, which are essential for database design and deployment. Key ideas of the architecture were recognised to help connect the ontology layer, which acts as a bridge between the conceptual and physical layers.

The system architecture consists of three main components:

- *PostgreSQL Relational Database:* Hosted on Azure to store MMIC's structured data aligned with the LDM.
- *GraphDB Triplestore:* Hosted on NPL's servers to enable SPARQL-based semantic querying.
- *Data Connector:* Uses the Ontop virtualisation engine and R2RML mappings to convert relational data into RDF triples that follow the structure of the ontology.

The Ontotop virtualisation engine is a system that exposes relational database content as virtual knowledge graphs, while R2RML mappings are a W3C standard language used to define customised mappings from relational databases to RDF datasets. This architecture enables querying over semantically enriched data without duplication, preserving alignment with the source data while allowing flexible, ontology-driven exploration.

C. Data Interfacing

Data interfacing connects systems to facilitate

Tab. 2. Mapping table. Relationship between SQL (structured) classes and corresponding ontology classes used in knowledge graph development. PK (Primary Key) depicts a unique, non-null identifier for each record in a database class.

SQL Table	Ontology Class	Superclass	PK Column	Identifier Property
process. job	:Job	iof-core: PlannedProcess	jobid	:jobIdentifier
process . isa95joborder	:JobOrder	iof-core: DirectiveInformationContentEntity	joborderid	:jobOrderIdentifier
process. isa95jobresponse	:JobResponse	iof-core: DescriptiveInformationContentEntity	jobresponseid	:jobResponseIdentifier
process. materiallot	:MaterialLot	scro: MaterialEntity	materiallotid	:designatedByLotNumber

information exchange. The semantic data interface removes the need for manual porting and streamlines the conversion from relational databases into a semantically connected fabric. Two options were considered for the development of the data connector:

- *Virtualised Instantiation:* Using R2RML to dynamically pull data from a relational database and map it into a graph within a triple store.
- *Python-based ETL Process:* Hardcoding an ETL processing and mapping layer through Python, providing materialised triples for loading into a triple store.

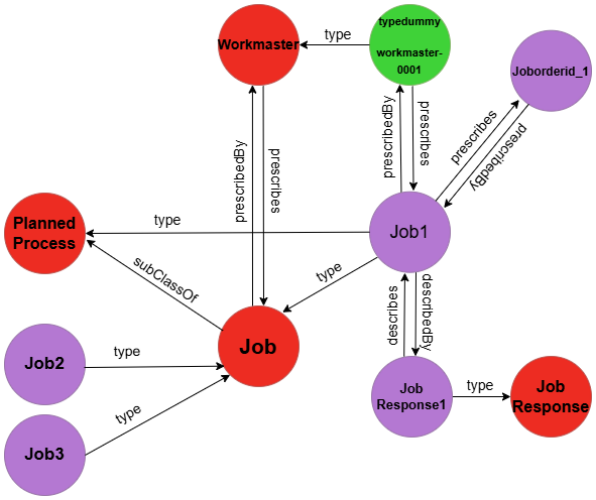

Figure 1. Knowledge graph visualisation illustrating a manufacutring process. Red nodes represent classes, green nodes represent individual instances and purple nodes denote instance-level data. data. Solid arrows represent explicit relationships such as type, subClassOf, prescribes and describes.

R2RML, along with RML, provides a strong and adaptable system that supports both relational and different types of data sources (RML), making sure that mappings are consistent and simple to access. An Extract-Load-Transform (ELT) approach was not considered due to factors like the limited scope of the project, the nature of the available data sources as well as available infrastructure.

III. RESULTS

To enable interoperability of structured data, key entities and attributes were mapped to an LDM using ontology classes. The process of mapping the structured data to an ontology involved several key steps:

- *Identifying Relevant Entities and Attributes:* Key components within the LDM were identified.
- *Aligning with Ontology Classes:* These elements were aligned with ontology classes to ensure semantic compatibility. An application ontology was generated as a subclass of the domain ontology, allowing for different levels of abstraction and future growth.
- *Validating Mappings:* Mappings were validated to confirm that the structured representation accurately reflects domain requirements.

This customised ontology, with relevant entities and attributes, extended standard classes to meet domain-specific requirements. It takes organised data from the relational database and represents by using ontologically enriched RDF triples. This graph structure allows for consistent representation of entities such as materials, manufacturing steps, and material lots. The data was instantiated using this ontology and SPARQL queries were ran to test the system's capabilities. These queries were developed based on the competency questions to validate the system's functionality and demonstrate its capability to answer stakeholder-driven questions.

Evaluation was conducted by translating each competency question into one or more SPARQL queries a validating the returned results against known records from the LDM. For instance, to test material availability (CQ3), queries checked for the correct retrieval of lot identifiers and their associated storage locations. Success was defined as accurate retrieval of all expected links between materials, locations, and process requirements. Partial

failures highlighted ontology gaps, which are discussed in Section IV.

The result of this methodology is a knowledge graph that integrates data from a relational database as RDF triples, semantically enriching information on materials, processes, and production steps. A snippet of this graph is shown in **Error! Reference source not found.**. This structure supports consistent, machine-operable links across the manufacturing lifecycle.

IV. DISCUSSION

A. Data Integration and Interoperability

This work demonstrates how semantic technologies can address key data integration and interoperability challenges in pharmaceutical manufacturing. By aligning structured data with a domain-specific ontology, we enabled richer, more contextualised representations of manufacturing processes. The knowledge graph helps access and interpret data more consistently as well as offers a flexible way to add new features in the future, such as connecting real-time process data or including external standards [4].

B. Stakeholder Engagement and Accessibility

One of the main benefits of the knowledge graph is its ability to enable non-technical users to explore data using domain-specific language and concepts. For example, a process engineer can issue a query like "which material lots are stored at Facility A?" without needing to know SQL schema. The semantic layer acts as a translator between user-friendly terms in the conceptual model and the underlying RDF structure. This significantly reduces the learning curve for domain users and encourages wider adoption across teams.

C. Scalability and Applicability

While the current demonstrator is specific to pharmaceutical manufacturing (tablet production), the approach is broadly applicable across other manufacturing contexts. The scalable framework provided by the knowledge graph allows for future extensions, including the integration of real-time process data and external standards. Ongoing work includes extending the ontology, improving automation of data mapping, and evaluating impact in operational settings.

D. Automation and LDM Alignment

Several challenges were encountered during the automation and alignment of the LDM with the ontology. The physical data model was auto-generated from an enterprise system to be regenerated on NPL servers, but due to using different database systems, the auto-generated PostgreSQL did not consider dependencies in the order of generation, and some constraints and foreign keys were missing. Manual adjustments were made to generate the model on the NPL server, but differences between the LDM and the implemented model remain. Future work will focus on aligning both models and generating synthetic data to cover all tables and test cases.

E. Ontology Maturity and Scope Limitations

While IOF-Core, NIIMBL, and SCRO provided a robust starting point, these ontologies required adaptation to reflect the realities of pharmaceutical manufacturing. Several domain-specific concepts were not supported, limiting immediate coverage of some user requirements. This challenge underscores the importance of treating semantic modeling as an iterative process, continuously refining domain ontologies as the data landscape and use cases evolve.

To ensure semantic consistency across data and knowledge layers, we introduced the ontology classes JOB and JOB STEP as bridging constructs between the architecture (LDM, CDM) and the ontology layer. While the LDM and CDM provided structural definitions of operations, they lacked explicit semantic representation of process hierarchies and temporal dependencies. By modeling JOB as a subclass of `PlannedProcess` and JOB STEP as a granular temporal part thereof, we were able to represent domain-specific workflows such as granulation and blending more precisely. This design also supports alignment with definitions found in the ISA-95 and MMIC data models, where a job is commonly understood as a unit of scheduled work composed of ordered steps. Mapping this to the ontology enabled more meaningful query results and improved traceability across digital models.

F. Alignment Between LDMs and Ontologies

Mapping technical attributes from specific logical data models to high-level ontology concepts presented difficulties. Not all terms aligned clearly, and some attributes had to be reinterpreted or abstracted during the mapping process. Close collaboration between domain experts and data scientists was essential to maintaining semantic integrity without distorting the original meaning of the data. However, due to time constraints, not all modelling was performed according to BFO and IOF standard patterns, particularly in the representation of temporal aspects.

G. Virtualised Queries

Using R2RML virtualisation to avoid data duplication was a practical choice, but it introduced trade-offs. Virtualisation supports live-access data but can introduce lag and increased query time, reducing the speed and availability of data through advanced inferencing. Some advanced SPARQL query features are not yet implemented, such as asking whether an entity "exists" in the virtualised graph. Future improvements will focus on optimising query performance and queryability.

V. CONCLUSION

This project has demonstrated the value and feasibility of integrating semantic technologies within the MMIC digital architecture and, more generalised, pharmaceutical manufacturing. By developing a domain-specific ontology and applying it to real-world manufacturing data, we created a functional knowledge graph that supports more informed decision-making. The approach is flexible, extensible, and offers a foundation for broader digital transformation in the sector. The lessons learned from this pilot will inform the next generation of data-driven tools in medicine manufacturing. With continued collaboration, semantic technologies have the potential to drive efficiency, interoperability, insight generation, and digital maturity across the pharmaceutical ecosystem.

VI. FUTURE WORK

Future work will expand the use case to implement actual sensor and process events, demonstrating full-scale implementation. This includes modelling units [12] and sensors [11] with appropriate ontologies to support provenance and traceability [14]. Expanding the ontology to include testing materials and incorporating quality assurance, formulation processes, deviation handling, regulatory documentation, and batch-specific testing protocols will deepen the semantic fidelity of the graph.

Improving user access with graphical query builders, dashboards, and natural language interfaces will demonstrate the value of natural language to SPARQL to knowledge graph for data democratisation. Building tutorials and methods for ontology mapping and integration across the architecture will facilitate adoption.

Improvements in graph query performance are needed for production-level deployment, including materialising selected RDF graphs rather than relying on R2RML or applying query rewriting and indexing strategies within GraphDB. A more scalable and enterprise-ready implementation of this architecture could be applied to numerous other pharmaceutical manufacturing domains, such as environmental monitoring and compliance, production scheduling and capacity planning, as well as formulation management and change control.

ACKNOWLEDGMENTS

This work was funded by the Department for Science, Innovation and Technology through the UK's National Measurement System. The authors would further like to thank our collaboration partners at the Medicines Manufacturing Innovation Centre consisting of our partners at CPI, University of Strathclyde, UK Research & Innovation, Scottish Enterprise and founding industry partners, AstraZeneca and GSK as well as the Industrial Ontologies Foundry, specifically members of the Supply Chain Reference Ontology working group for the discussions and suggestions.

REFERENCES

[1] US Food and Drug Administration, "Guidance for Industry: Process Validation: General Principles and Practices," Center for Biologics Evaluation and Research. US Department of Health and Human Services, Rockville, MD, Jan. 2011.

[2] International Society of Automation, "ISA95, Enterprise-Control System Integration," 2000. [Online]. Available: https://www.isa.org/standards-and-publications/isa-standards/isa-standards-committees/isa95. Accessed: 2 May 2025.

[3] International Society of Automation, "ISA88: Batch Control," 2010. [Online]. Available: https://www.isa.org/standards-and-publications/isa-standards/isa-standards-committees/isa88. Accessed: 2 May 2025.

[4] Y. Chen, C. Sampat, Y-S. Huang, S. Ganesh, R. Singh, R. Ramachandran, G. V. Reklaitis, M. Ierapetritou, "An integrated data management and informatics framework for continuous drug product manufacturing processes: A case study on two pilot plants," International Journal of Pharmaceutics, vol. 642, 2023, pp. 123086. doi:10.1016/j.ijpharm.2023.123086

[5] H. Cao, S. Mushnoori, B. Higgins, C. Kollipara, A. Fermier, D. Hausner, S. Jha, R. Singh, M. Ierapetritou, R. Ramachandran, "A Systematic Framework for Data Management and Integration in a Continuous Pharmaceutical Manufacturing Processing Line," Processes, vol. 6, No. 5, 2018, pp. 53. doi:10.3390/pr6050053

[6] B. Kulvatunyou, M. Drobnjakovic, F. Ameri, C. Will, B. Smith, "The Industrial Ontologies Foundry (IOF) Core Ontology," in Formal Ontologies Meet Industry (FOMI) 2022, Tarbes, FR, 2022. doi:10.1007/978-3-031-13889-8_10

[7] R. Arp, B. Smith, A. D. Spear, "Building Ontologies with Basic Formal Ontology," MIT Press, Cambridge, MA, USA, 2015.

[8] Semantic Computing Lab Arizona State University, "Supply Chain Reference Ontology (SCRO)," Nov. 18, 2022. [Online]. Available: https://spec.industrialontologies.org/iof/ontology/supplychain/SupplyChainReferenceOntology/?version=release/202301. Accessed: 2 May 2025.

[9] F. Ameri, E. K. Wallace, B. Kulvatanyou, "Towards a Reference Ontology for Supply Chain Management," in I-ESA Workshops, 2020. doi:10.1007/978-3-030-68462-4_2

[10] The National Institute for Innovation in Manufacturing Biopharmaceuticals, "Open-sourced Biopharmaceutical Manufacturing Ontology," 2024. [Online]. Available: https://www.niimbl.org/projects/opensourced-

biopharmaceutical-manufacturing-ontology/
[11] M. Compton, P. Barnaghi, L. Bermudez, R. García-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog, V. Huang, K. Janowicz, W. Kelsey, D. Le Phuoc, L. Lefort, M. Leggieri, H. Neuhaus, A. Nikolov, K. Page, A. Passant, A. Sheth, and K. Taylor, “The SSN Ontology of the W3C Semantic Sensor Network Incubator Group,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 17, pp. 25–32, Dec. 2012.
[12] FAIRsharing.org: QUDT; Quantities, Units, Dimensions and Types, DOI: 10.25504/FAIRsharing.d3pqw7, Accessed:22 May 2025.
[13] R. Albertoni, D. Browning, S. Cox, A. N. Gonzalez-Beltran, A. Perego, P. Winstanley, "The W3C Data Catalog Vocabulary, Version 2: Rationale, Design Principles, and Uptake," Data Intelligence, vol. 6, no. 2, 2024, pp. 457–487. doi:10.1162/dint_a_00241
[14] L. Moreau, P. Groth, J. Cheney, T. Lebo, S. Miles, "The rationale of PROV," *Journal of Web Semantics*, vol. 35, Part 4, 2015, pp. 235-257. doi:10.1016/j.websem.2015.04.001. Available: https://www.w3.org/TR/prov-o/
[15] J. Gregório, M. Alsuleman, M. Chrubasik, P. Duncan, G. Bisland, "A competency question driven approach to conceptual data model design for digital verification and validation," in Advanced Mathematical and Computational Tools in Metrology and Testing XIII, 2025, pp. 71-82.