A Semantic Interoperability Toolkit for Sharing Energy Data and Models in a Manufacturing Data Space

Tharindu Ranathunga
Munster Technological University, Ireland
tharindu.ranathunga@mtu.ie

Sourabh Bharti
Munster Technological University, Ireland
sourabh.bharti@mtu.ie

Alan McGibney
Munster Technological University, Ireland
alan.mcgibney@mtu.ie

ABSTRACT

This position paper discusses the key issues and requirements for achieving semantic interoperability in a manufacturing data space, focused on the seamless sharing of energy-related data and machine learning (ML) models across sites to support energy efficiency in manufacturing systems. Several publicly available ontologies exist for sharing energy-related data and ML models. However, there is a requirement to extend these ontologies to support the ML model lifecycle in the realm of data spaces. This paper proposes an extension to existing ontologies and introduces a software toolkit designed to utilize these semantic models throughout the lifecycle of ML model development. This covers interoperability and integration across local data silos (energy and performance data) to the sharing of ML models across sites/organisations using International Data Spaces (IDSA) connectors. Both the semantic models and developed toolkit are publicly released[1] to support and benefit the broader research community in this space.

1 INTRODUCTION

Advanced digital technologies such as Machine Learning (ML) are a key enabler to allow companies uncover energy saving opportunities through evidence-based analysis of asset and performance data. However, access to data remains a significant barrier to the adoption of AI approaches. This is due to siloed infrastructure, lack of interoperability, low volumes and poor-quality data and complexity of solutions. This is further exacerbated due to vendor lock-in, privacy and security concerns. Open data spaces are emerging as key drivers to overcome these barriers. This position paper is a research outcome of the CORDS project [1] that leverages the concept of data spaces to accelerate energy-efficient opportunities in the manufacturing domain by moving beyond siloed intelligence towards shared intelligence built on an open innovation ecosystem.

As part of the CORDS project several stakeholder engagements and co-creation activities were conducted to capture the interoperability requirements for energy-related data sharing in the manufacturing domain. During the stakeholder engagement, the following key questions were explored with industrial stakeholders to frame the challenges in adopting data-driven energy-efficient solutions:

- What are the barriers to integrate and access energy related data stored at different sources e.g., edge, fog, and cloud?
- What are the barriers to sharing energy-saving insights or ML models with others?

Responses indicated several key challenges (1) *an interoperability issue exists between equipment, site and building energy-related data*; (2) *there is no Common Information Model (CIM) in place to unify the data representation across the manufacturing site*; and (3) *no CIM is developed or shared across sites to facilitate the seamless sharing of ML models*. CORDS addresses these challenges by designing CIMs for both energy-related data and ML models and by building sophisticated software tools for various actors, from database administrators to ML engineers, to utilize these CIMs in real-world scenarios.

Specifically, the primary objective of this position paper is to present these CIMs (Section 3) designed to be shared across manufacturing sites and companies to foster semantic interoperability and a collaborative ecosystem. Additionally, the paper introduces a toolkit (Section 4) that supports the use of the defined semantic models throughout the life cycle of energy prediction ML models. Finally, Section 5 concludes the paper.

2 BACKGROUND & USE CASES

Within manufacturing sites, energy consumption data is typically collected with the aid of sensors and energy meters. Such energy meters can be mounted on various manufacturing machines, devices, processes and auxiliary services (e.g. Heating, Ventilation, and Air Conditioning (HVAC)). Every machine/HVAC system can have multiple metering points measuring different energy types such as gas, electricity, thermal etc. The energy data is generally gathered and stored in a distributed manner across multiple systems i.e., Building Management System (BMS), and Energy Management System (EMS). Other data sources such as weather data, energy costs, and production planning can also impact the overall energy consumption and as such should also be considered. However, due to multiple representations of the energy data, there can be a lack of interoperability between different energy management systems in the organization. Implementing CIMs that can integrate the energy-related data distributed across multiple systems can be a crucial step towards interoperability across organizations. CORDS adapts and extends existing semantic models such as SAREF [2] and SEAS to enable energy-related data sharing at a local level.

With data structured in a manner that is usable within the organisation, it is then possible to explore the opportunities to maximise its value. For example, the development of ML models on the locally integrated energy data can bring new optimisation opportunities to reduce cost and resource usage. This can be taken a step further and involves sharing such models globally in a privacy-preserving manner across sites. This emerges as another driver for leveraging a CIM for ML models. To this end, CORDS utilises and enhances existing semantic models such as DCAT [3] and ML Schema [4] and develops a CIM that aligns with the IDSA [5] technical components to enable interoperability across data space actors involved in ML model exchange. As developing energy-prediction models and

[1] Released under Apache Version 2.0 License

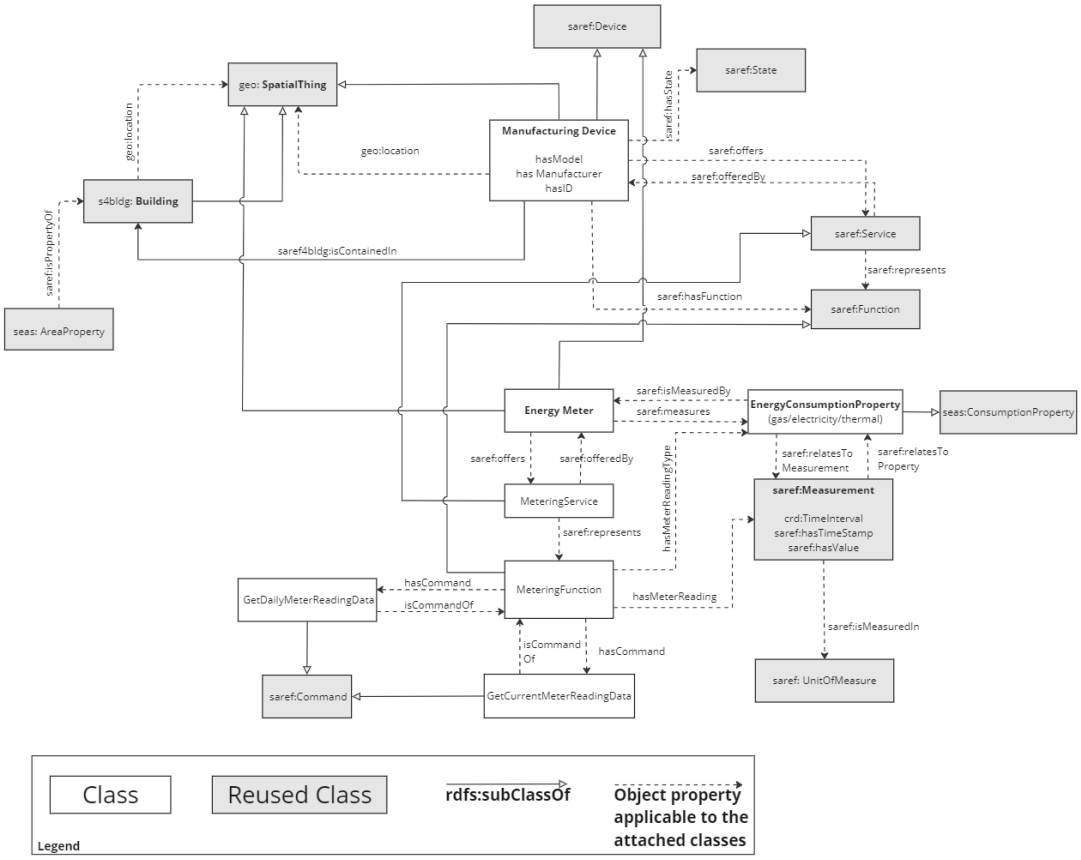

Figure 1: CIM for Energy Data

sharing them across a network of data space connectors involves semantic interoperability at both local and global levels, it is equally important to provide a set of tools to support practitioners in their practical use.

The methodology for adopting/extending existing ontologies involved several key steps, including use-case description data gathering from stakeholders, identifying key concepts (classes), properties and instances and defining relationships between them. This was followed by identifying and re-using similar/related elements in existing ontologies and finally extending them to support CORDS-related elements. The methodology for developing the toolkit is explained in Section 4.

3 CORDS COMMON INFORMATION MODELS

3.1 CIM for Energy Data

There are several existing semantic models related to energy data in the manufacturing domain, such as SAREF, CIM, and SEAS. SAREF provides a standardized semantic framework for interoperability among smart devices and appliances. CIM offers a comprehensive schema for representing and managing data in the electric power industry. SEAS focuses on enhancing energy efficiency and sustainability through semantic descriptions. However, some concepts overlap or are missing across these models.

The CORDS use case specifically focuses on consolidating multiple energy data sources into a unified view. Smart meters provide real-time data on energy consumption and demand, while HVAC sensors monitor climate control. Building management systems (BMS) offer an aggregated view of the plant's overall energy consumption, integrating data from multiple sources. In CORDS, data from these various sources should be consolidated to derive insights that can improve energy efficiency. To achieve this, a Common Information Model (CIM) is needed to provide the required semantic interoperability. While existing semantic models like SAREF [6], CIM [7], and SEAS [2] represent some concepts related to CORDS, there were minor extensions were done as part of ontology analysis. For instance, the Weather concept in SEAS was extended by adding another property named as 'Humidity' and the Device concept in SAREF is extended by adding another property named as 'ID'.

Fig 1 shows the semantic data model for which we defined our use case-specific concepts related to Energy Management. CORDS adopts the concept definitions already defined in benchmark ontologies such as SAREF and SEAS. The building concept is re-used in

Figure 2: Contract class added to MLS Ontology

Figure 3: Query class added to MLS Ontology

S4bldg: Building ontology. The concept of 'Manufacturing Device' extends the class SAREF: Device and adds another unique property as hasID to have unique identification (eg: meter id). Every Manufacturing Device offers a SAREF: Service representedexecuted by SAREF: Function. The concept of 'Energy Meter' also extends the class SAREF: Device. Every 'Energy Meter' offers a MeteringService represented by MeteringFunctions executed on SAREF:Commands and measures an EnergyConsumptionProperty extending the SEAS: ConsumptionProperty. Many customised MeteringFunctions can be defined such as GetDailyMeterReadingData, GetCurrentMeterReadingData as shown in Fig 1.

3.2 CIM for Machine Learning

When sharing energy prediction models, it is essential to include a semantic description of the model training process to facilitate easier reuse or retraining by the consumer. To achieve this, semantic modelling is used to represent the concept of ML model training. Existing ontologies such as ML Schema [4] and DCAT [3] were examined to identify reusable elements. The ML Schema ontology was particularly useful in defining the relationships and properties involved in the training process. This schema includes components like algorithms, implementations, datasets, and evaluation measures, all of which are critical for providing a comprehensive description of the model training lifecycle. By adopting and extending these existing ontologies, the CORDS ML ontology was developed to align with the specific use case requirements. Fig 2 and 3 show the IDS-related classes added by CORDS to the existing ML Schema ontology. The full current version of CORDS ML ontology is available online [8].

4 TOOLKIT

Semantic models can exist independently of their practical applications; however, the crucial factor is to facilitate their use in real-world scenarios. In organizational settings, various roles, from database administrators to ML engineers, work at different stages of data analytics and ML life cycles. To effectively utilize semantic models, there should be a comprehensive set of tools that assist these stakeholders in integrating and applying the models. Specifically, when preparing these resources to be shared using the IDSA protocol, these tools should enable the seamless extraction and publication of semantic descriptions as metadata in the data space ecosystem. We have implemented a set of tools that support the use of defined semantic models during the life cycle of developing energy prediction models.

Figure 4 illustrates how the developed tools integrate with the end-to-end workflow from data acquisition and ML model training until a model is shared in the data space. It facilitates the extraction of metadata from various stages and prepare them in a IDSA compatible semantic interoperable way. The components that play a significant role in terms of semantic interoperability are discussed here.

4.1 Semantic Engine

The CORDS semantic engine empowers data administrators to create a unified view of energy data from multiple sources. It provides an API that allows data engineers to integrate existing local data sources into the CORDS CIM. The Data Source Manager enables the integration of different data source connectors into the pipeline with minimal configuration. A key feature of this system is its ability to let data engineers describe the content of the integrated data sources using the provided semantic model. For instance, it allows the addition of the domain-specific ontology, such as one that was developed in Section 3.1 (Ontology for Energy Data), to the semantic engine. Subsequently, the API can be used to describe

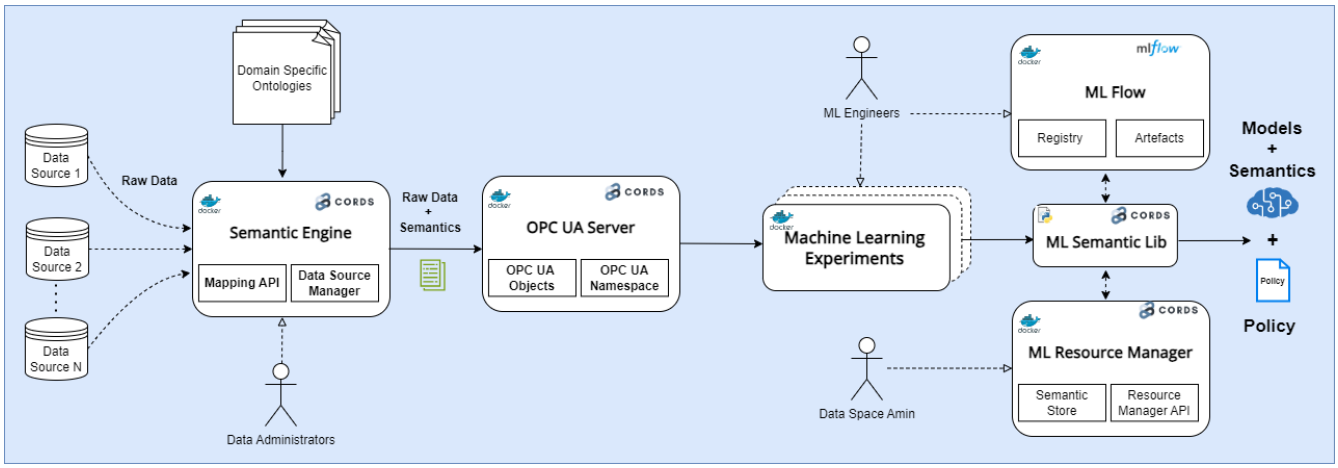

Figure 4: CORDS Semantic Toolkit

custom scenarios using the terminology defined by the ontology as shown in Fig 5. This mechanism can be further automated using more intelligence tagging mechanisms [9].

4.2 OPC UA Server

CORDS utilizes OPC UA [10] as the common interface for data access, a design choice made carefully considering the nature of time series energy data generated in manufacturing processes and the required level of communication interoperability. To achieve semantic interoperability for the CORDS UA Server, the information provided by the Semantic Engine is leveraged. The *type space* of the server can be generated based on the ontology, and the address space of the OPC UA is generated dynamically based on the mapping available within the Semantic Engine. An OPC UA address space is a collection of Nodes arranged in a hierarchical manner, whereas the definition of such Nodes can be taken from the benchmark ontologies. Figure 6. shows such a mapping where an OPC UA address space is mapped to a Type Space and their associated ontological definitions. Finally, the data pushed into the OPU UA server can be accessed by data analysts or ML Engineers, and the semantic model can be utilized to interpret data and understand the use-case specific contexts.

4.3 Semantic Library for ML

The data from the OPC UA server is fed into ML experiments to train energy prediction models. As with any ML training workflow, this process involves several steps, such as pre-processing and transformation, model training, testing and validation, and model selection. The semantic model presented in Section 3.2 comprehensively describes what occurs during these phases. However, the main challenge is enabling ML engineers to articulate the processes using the provided ontology. To address this, a Python-based semantic library [11] was developed that integrates with various tools and libraries used by ML engineers. This library provides the following three key features for ML Engineers: (i) Artefacts Tagging: Provides CORDS ML vocab (based on the ML semantic model) as a set of Tags that can be used to describe metadata in ML experiments; (ii) Extract Meta Data: Provide an interface to extract these from ML experiments (iii) Serialize the semantic description using the extracted tags into a RDF file.

For example, ML engineers often use tools like MLflow[12] to track and manage their ML experiments and the artefacts, such as ML models, developed during these experiments. Our ML semantic library is integrated with MLflow. The set of CORDS ML vocab tags can be used to describe ML Flow experiments. This includes details on the nature of transformations applied, the ML algorithms used, and the model evaluation metrics. This integration allows ML engineers to utilize these tags within MLflow's visualization features. Additionally, using the CORDS ML semantic library, metadata from experiments can be extracted as an RDF definition, making it compatible with IDS resource descriptions. These resource descriptions can then be registered in an IDSA ecosystem using the ML Resource Manager component. This component provides IDSA-specific resource preparation functionalities, such as defining usage policies, registering on a Connector, or passing resource meta-data to a Broker. This enables seamless semantic interoperable resource management within the IDSA ecosystem.

5 CONCLUSION

Two robust semantic models were developed to ensure interoperability during the process of energy data collection and prediction model training in a manufacturing context. The toolkit implemented aims to streamline the entire workflow, from data collection to ML model training, enabling seamless semantic representation of activities, algorithms used, and expected performance outcomes. Additionally, this toolkit integrates with the IDSA data space components, allowing for the extraction and advertisement of metadata in the data space using the IDSA information model. These advancements facilitate a more efficient and transparent approach to energy management, promoting the adoption of ML-driven energy-efficient solutions across the manufacturing sector. As the next step, this toolkit will be integrated into a Minimum Viable IDSA Data Space to demonstrate how self-sovereign, privacy-preserving