Navigating the Data Model Divide in Smart Manufacturing: An Empirical Investigation for Enhanced AI Integration

István Koren[1], Matthias Jarke[2], Judith Michael[3], Malte Heithoff[3], Leah Tacke Genannt Unterberg[1], Max Stachon[3], Bernhard Rumpe[3], and Wil M.P. van der Aalst[1]

[1] Process and Data Science, RWTH Aachen University, Aachen, Germany
{koren,leah.tgu,wvdaalst}@pads.rwth-aachen.de
[2] Information Systems and Databases, RWTH Aachen University, Aachen, Germany
jarke@dbis.rwth-aachen.de
[3] Software Engineering, RWTH Aachen University, Aachen, Germany
{michael,heithoff,stachon,rumpe}@se-rwth.de

Abstract. In engineering informatics, the myriad data types, formats, streaming and storage technologies pose significant challenges in managing data effectively. The problem grows, as new analytics perspectives are emerging from a totally different AI-based tradition. This divide often necessitates the development of custom solutions that link specific data capture methods to particular AI algorithms. Encouraged by the success of object-centric mining models for discrete processes, we look for large clusters of data management practices where novel bridging data models can help navigate the data model divide. We address this question in a two-cycle design science approach. In a first cycle, over 80 actual data model practices from a wide variety of engineering disciplines were analyzed, leading to four candidate fields. In a second cycle, an initial bridging data model for one of these fields was developed and validated wrt some of the found practices. Our findings offer the prospect of significantly streamlining data pipelines, paving the way for enriched AI integration in production engineering, and consequently, a more robust, data-driven manufacturing paradigm.

Keywords: Industry 4.0 · Manufacturing Data Model · Empirical Study · AI Integration · Digital Shadow.

1 Introduction

The Industrial Internet of Things (IIoT) and Industry 4.0 have ushered in a new era of opportunities for the manufacturing industry. They promise enhanced operational efficiency, increased productivity, and the potential for innovation in product design and manufacturing processes. Central to realizing these opportunities is the integration of Artificial Intelligence (AI) tools which can provide intelligent analytics, predictive maintenance, and autonomous decision-making,

among other benefits. However, the implementation and optimization of AI in manufacturing hinges on the effective management and integration of vast and varied data generated across the production lifecycle [15]. A predominant challenge in leveraging this data effectively is the heterogeneous nature of data models and pipelines across different use cases in manufacturing. Current common practice involves custom solutions for data management and analytics for each application, owing to the lack of better, standardized approaches. This practice, while solving immediate challenges, consumes significant resources and obstructs cross-domain interoperability and knowledge transfer.

This paper explores innovative solutions to this problem at the data model level. After a review of data management research and practice focused on Digital Twins (DT) in manufacturing, it pursues a two-cycle design science approach [16], contributing to these research questions:

RQ1: Do data model practices within and across engineering disciplines expose sufficient similarity to make the existence of useful standardized data models plausible? Answering such a question is not easy due to the reluctance of many companies to share their practices, let alone the cross-validation of claimed practices by looking at actual data. Fortunately, the research cluster Internet of Production (IoP) at RWTH Aachen University [9] with over 25 different engineering and related natural science disciplines—all actively involved in application-oriented research and practice—offers a unique alternative setting for such a study. In Section 3, we report on design, results, and implications of a structured analysis of over 80 such data model practices at both a conceptual and data-example level. The identification of at least four broadly observed candidate clusters of practices indicates an affirmative answer to RQ1.

RQ2: How can a "bridging" data model be derived for such a cluster, and how can it be validated from formal and practice perspectives? Formal requirements for the models were initially derived from both general design principles in database and knowledge graph research, and the insights of pioneering work in object-centric process mining [3]. Subsequently, a bridging model was developed, addressing the commonly observed need to integrate measurements of continuous processes and discrete events with high-level analytics. This model underwent both a formal evaluation and a practical assessment. The practical assessment involved quantifying the effort necessary to generate data from five distinct sensor data practices identified in RQ1, and examining the utility of this data in various analytical processes. The results indicate potential for enhanced analytics and data sharing despite reduced effort, but also show limitations and needs for further research.

This paper is structured as follows. The next section presents background and related work. Subsequently, Section 3 describes the empirical study of data model practices. Section 4 introduces and discusses the *Measurement and Event Data* format as the first example of a bridging model. Finally, Section 5 concludes this paper.

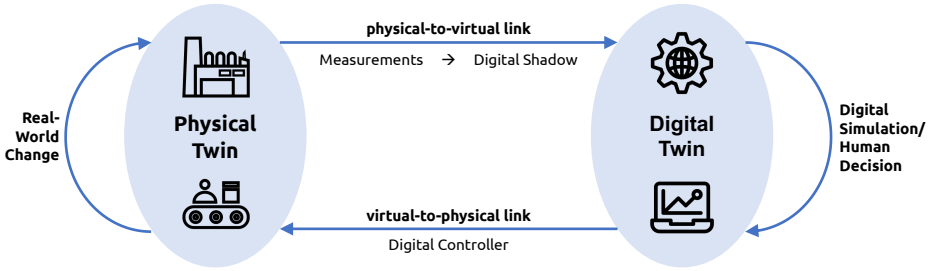

Fig. 1. Digital Twin showing the data model gap between sensor-based metrology and AI-based digital shadows.

2 Background and Related Work

Persistent problems of data availability in manufacturing engineering, operations, and usage [6] as well as in AI generally [15] are well-known. Partial solution proposals come from many areas of Computer Science [8]. Examples include: requirements engineering in manufacturing cases [23]; data lake-based layered metamodel for Computer-Aided Engineering [33]; optimization and security of physical dataflows in the edge-cloud spectrum [28,27].

In industry, leading cloud providers offer their own standard data models on open platforms, such as the Open Manufacturing Platform (OMP) on top of Microsoft's Azure IIoT cloud (cf. `https://azure.microsoft.com/solutions/industrial-iot`). Several initiatives are standardizing approaches to reduce reliance on vendor-specific solutions and domain-centric modeling languages. AutomationML [1], an XML-based, object-oriented data modeling language, supports the creation, storage, and exchange of engineering models. It serves as a neutral format for data exchange across diverse manufacturing scenarios. The OPC Unified Architecture (OPC UA) offers standardized information models with associated guidelines and best practices, including standard APIs for novel specialized services such as data access or alarms and conditions [26]. In the domain of standards and reference models that enrich the solution space through ontologies, notable examples include the Smart Appliances REFerence (SAREF) ontology [19] and the framework provided by the International Data Spaces Association [4].

The growing complexity of manufacturing systems with multiple conflicting goals, frequently changing boundary conditions and strategies have led to the conclusion that any solution concept must take the essentially decentralized and modular, yet interoperable nature of manufacturing data management into account. Interacting DTs have emerged as a widely accepted abstraction paradigm, often inspired by experiences from multi-agent systems [29]. Recently, also the IIoT community, like Industrie 4.0, re-interpreted their idea of Asset Administration Shell (AAS) as enablers for DTs [22].

Each DT accompanies the life of some real-world object, process, or aggregate Cyber-Physical Production System (CPPS) in a so-called twinning cycle,

as illustrated in Figure 1. This twinning requires a bi-directional connection between the real world and the DT, such that real-world changes and digitally found decisions are reflected transparently with well-defined frequency and faithfulness [13].

The importance of data in DT architectures was already recognized a decade ago [12], recently also in civil engineering [24]. A Digital Shadow (DS), in this context, refers to a digital representation of a physical asset or process, which is essential for data-driven decision-making and analytics. Organized around this DS concept, data management must support two core tasks in a DT-based infrastructure. It focuses on the creation and maintenance of DSs by a wide range of intelligent analytics combining model-based and AI approaches [11,10]. But DTs are also active cooperating or competing agents that sovereignly share DSs in data spaces [17,32]. From a conceptual modeling perspective, DSs have recently been characterized as materialized views and as shareable, even tradeable data assets [21], but also as software engineering artifacts with a real-world grounding and well-defined provenance information [25].

Figure 1 implicitly showcases a "data model gap". This gap is not merely about the physical and digital representations but also about how data is modeled, structured, and utilized in these two realms. In the physical world, data capture methods are often heterogeneous, reflecting the complex reality of physical processes. Conversely, the digital world, particularly within AI algorithms, requires data to be structured in a highly standardized format for efficient processing and analysis. This discrepancy between the physical "as-is" and the digital "to-be" structured data leads to a data model divide.

Contrary to the often complex semantic structures emphasized in the discussed standards and models, our approach portrayed in this paper aligns more closely with the methodology observed in general AI libraries. As input, these libraries rely on a limited number of standardized data formats, such as CSV or other forms of tabular data, which serve as the basis for parameterizing algorithms and frameworks. By adopting parameterizable data models, we facilitate an amalgamation of both schema and instance data, simplifying the data model complexity while maintaining versatility and effectiveness in the AI-driven analysis and decision-making processes.

3 Stage 1: Empirical Study of Data Model Practices

Our approach is informed by observations in Figure 1. In their survey of DT approaches, Jones et al. [18] emphasize that the activities involved in the physical2virtual link span two largely disjoint communities of research and practice. The long established engineering theories of measurement (metrology) with the related sensor management IT community (e.g., [31]) must somehow be matched to the explosively growing field of model-driven analytics and data-driven AI for the creation, optimization, visualization, and sharing [20] of purpose-oriented DSs. The challenge arises in managing the multitude of potential $m \times n$ mappings between these two parts. The claim pursued in this paper is that a few

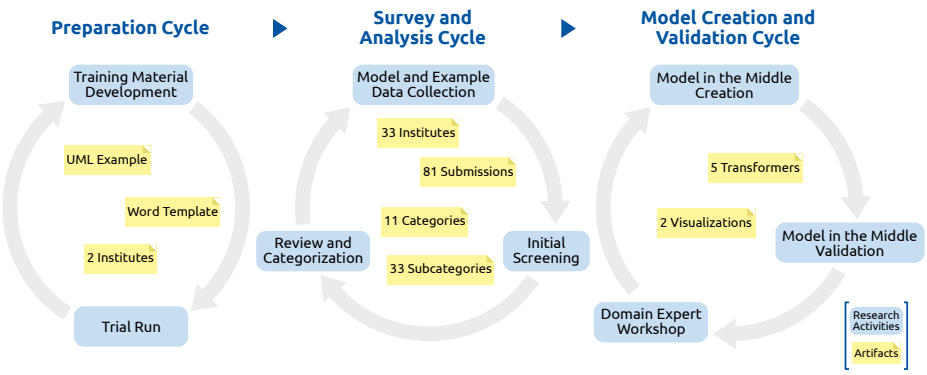

Fig. 2. Research Cycles of the Empirical Study.

(say: b) standardized "models in the middle" could reduce the mapping complexity to $b \times (m+n)$. This could offer scalability, reuse, and cross-enterprise sharing, with the potential for significant reduction in effort, and increased utility.

We conducted an empirical study to investigate how data models are employed across various use cases within the manufacturing domain. Our approach aligns with the Design Science Research Methodology (DSRM) [16], which we adapted to include a survey to gather empirical data and validate our research artifacts. The overall design science process is portrayed in Figure 2.

3.1 Research Design

Adhering to the DSRM, our initial effort was aimed at understanding the problem domain and the requirements for a potential solution. We selected the highly interdisciplinary engineering team from the IoP research cluster as study participants to obtain meaningful information, considering their diverse backgrounds.

Preparation Cycle: Previous modeling experiences in the research cluster indicated familiarity with modeling languages such as UML, yet there remained a gap between theoretical knowledge and practical application. To address this, we developed training materials illustrating UML class diagram modeling with simple everyday object associations, avoiding mechanical engineering content to mitigate bias. Next, we created a Word template for capturing essential metadata and structuring data models. It was divided into two main sections—a UML class diagram and tables populated with example data—supported by the following metadata: dataset name, contact person, institute name, work package, version number, date, and a brief description. This template design was iteratively refined through pilot trials at two institutes with disparate data management practices. One institute operates legacy machines requiring manual intervention at every step, from process planning to data analysis using MATLAB. The other institute operates a connected ecosystem where industrial machines relay data

to a time-series database via a message broker, harnessing visualization tools like Grafana. Feedback from these trials was critical in evolving the training materials and integrating a UML class diagram example directly into the template. Identifying a suitable modeling tool presented significant challenges. Web-based tools, while easily accessible, were limited in functionality, restricting the extent to which they could be utilized for our modeling tasks. Native applications, although potentially more robust, were out of reach due to administrative restrictions within the engineering institutes. Thus, PowerPoint and Visio were recommended for modeling as biggest common denominator.

The feedback gathered from the pilot trials not only informed the iterative improvement of these artifacts but also provided valuable insights into the practical challenges and preferences in data modeling practices across different engineering disciplines. Consequently, the outputs of the preparation cycle, specifically the refined Word template and the updated training materials, became critical inputs for the subsequent cycle.

Survey and Analysis Cycle: The study's design and development phase received strong management endorsement and was promoted at key project events, leading to significant participation over two months. A total of 81 data models were submitted, verified for completeness, and any gaps addressed through follow-up queries. These models were versioned and stored securely in a Git repository, adhering to DSRM principles for traceability and rigorous evaluation. A thorough screening to identify and correct errors preceded a detailed coding and classification process. This ensured a methodical assessment of each submission, with discrepancies resolved collectively, enhancing the study's categorization approach.

Rigorous coding and classification were conducted by a mixed team of senior researchers and PhD students, ensuring a comprehensive and methodical evaluation. Each submission was assigned a first coder. This decision was then reviewed by a second coder. Finally, deviations and conflicts were discussed in the whole group and decided in virtual meetings, ensuring refinement and improvement of the categorization along the iterative nature of DSRM. Thus, the survey led to a valuable repository of empirical data to inform future design science research within the manufacturing domain.

To uphold the confidentiality agreement with participants, which was pivotal in securing 81 submissions, the detailed datasets underpinning our study will not be published. This assurance of confidentiality was essential for participant engagement and the integrity of our research findings.

The classification of data models, enriched by empirical evidence and collaborative refinement, served as a critical input for the next cycle, guiding the design and validation of a model that addresses the identified needs and gaps within data management practices of machine data.

Model Creation and Validation Cycle: In the final phase, the data models were systematically consolidated according to their respective categories, leading

Tab. 1. Overall data model categorization and subcategories.

Category	#	Description (Examples)
		MACHINE
machine data (measurement)	44	Time-series machine and event data.
machine master data	40	Machine type designations, and location.
machine configuration	39	(Default) parameters.
robot	10	Robot configuration.
3D printer	3	3D-printer-specific master data.
maintenance	2	Maintenance schedules and configuration.
		PROCESS
process steps/operation/measurement	45	Assembly instructions and sequence.
process aggregation (case/event/log)	15	Preprocessed event data.
experiment	12	Experiment setup.
images	10	References to binary image files.
process evaluation	6	Evaluation of production processes.
		MATERIAL
material properties	41	Material characteristics.
Bill of Materials (BoM)	7	Parts and part-of relations.
material amount/inventory/stock	1	Inventory and stock of material.
		SIMULATION & OPTIMIZATION
CAD/3D models	14	References to 3D model files.
simulation	8	Descriptions of simulation experiments.
computed results	7	Results of simulation runs.
planning	6	Simulation plans.
mathematical model/optimization	5	References and descriptions of mathematical models.
		FACTORY
factory/machine arrangement	15	Shopfloor layouts.
factory master data	8	Factory descriptions.
finances	4	Financial information on shopfloor equipment.
		PRODUCTS
product (parts)	25	Planning and/or evaluation of product parts.
		SUPPLY CHAIN
jobs/sales order	13	Details of orders.
delivery	6	Delivery master data like shipping address.
supplier	5	Supplier data like origin.
purchase order (material)	4	Details of purchases.

to the identification of potential candidates for models in the middle. These candidates were refined with domain experts during a dedicated workshop.

3.2 Results

Our analysis revealed a rich collection of 33 distinct model types, which we ultimately grouped into 11 categories. Some data models span multiple subcategories, highlighting the interconnected nature of manufacturing processes while underscoring potential integration points for bridging model design. For example, time series data frequently coincided with "experiment" or "process" categories, prompting multiple assignments.

Table 1 presents our categorization, listing both the primary categories and their subcategories alongside the count of data models in each. The majority of submissions fell under the MACHINE category, predominantly featuring time series measurements. This was followed closely by models describing PROCESS elements, like experiments or test runs. Due to space constraints, we omitted the four least-represented categories (quantity in brackets): HUMAN RESOURCES (21), METADATA (11), REQUIREMENTS (4), and SURVEY (4). These areas, while not the focus of this paper, represent valuable avenues for future exploration.

The 81 submissions collectively paint a heterogeneous picture, but nevertheless a striking similarity in challenges faced by different mechanical engineering processes across various disciplines. For instance, both aluminum die casting and plastic injection molding displayed a common issue: the internal control logic for pressure values operated at a higher frequency than what could be accessed via external interfaces. These shared challenges across disciplines are insightful for our endeavor to standardize and simplify data models, in particular towards the creation of automated data extractors and transformators. A common issue was handling external data such as 3D models or MATLAB files, which are often intricately integrated into the data models that merely outline their context.

Our analysis underscored not only the diversity of data models in manufacturing, but also common operational challenges such as the mentioned frequency discrepancies. These findings highlight critical caps that the "models-in-the-middle" aim to bridge. Specifically, the observed frequency differences between internal control logic and external data accessibility present a fundamental barrier to real-time AI analysis and decision-making. Before AI algorithms can be effectively applied, data must be synchronized and standardized, ensuring that AI tools can operate on real-time or near-real-time data seamlessly. Additionally, the integration of disparate data types into a cohesive model facilitates the development of automated data extractors and transformers, pivotal for AI's role in predictive maintenance, quality control, and process optimization. Thus, addressing these operational challenges is not merely a prerequisite but a foundational step towards realizing the full potential of AI integration in smart manufacturing.

3.3 Discussion and Implications for Data Model Design

While there is considerable diversity across the categories, a remarkable consistency exists within each category: certain modeling approaches and structures seem to be predominant in specific contexts within the manufacturing domain.

A frequently observed pattern was a triadic relationship encompassing (machine data) *measurements*, *processes*, and *products*. This relationship is a cornerstone in many submissions, albeit manifested differently across various stages of product development. In the inception phase (e.g., product development), this might include plans or sequences for robot movements and machine settings, while in the final stages, it shifts towards quality assessments and measurements.

Most models showcased intricate associations between different object types, yet these relationships were often not mirrored in foreign keys or similar in the example data. This observed discrepancy was made apparent by the fact that the majority of the data models were conceptualized retrospectively as part of the study. Initially, the data files (such as CSV files or database tables) were generated without an accompanying conceptual model. In the ex-post process of conceptual modeling, the modelers' inherent domain knowledge played a crucial role, enabling them to explicitly define relationships that were not initially apparent in the raw data. Inverting this approach—starting with a well-defined conceptual data model before data collection—holds significant potential for streamlining data handling.

Three data models documented cross-institute collaborations and two involved external industrial data, further highlighting the interdisciplinary potential by suitable bridging data models. This scarcity can be attributed to various factors, including NDAs and other confidentiality concerns.

The results provide valuable insights into the common patterns, variance, and limitations observed in the submitted data models. The recurring triadic relationship across models indicates a fundamental structure in manufacturing data modeling, while discrepancies between models and example data highlight a crucial area for improvement. The limited collaboration and external data integration also point to systemic challenges in data sharing and inter-institutional cooperation. These insights not only inform the current understanding but also shape our approach to future research and development in this area, especially in creating more integrated, real-world applicable "models in the middle".

4 Stage 2: Design and Preliminary Evaluation of an Intermediary Machine Data Model

We contribute towards a theory of data modeling by identifying a number of formal criteria that an intermediary model for navigating the data model divide should satisfy. While some of these criteria stem from decades of conceptual data model and model implementation research, others are inspired by specific experiences gained from an early success story in object-centric process mining.

Within this context, we then present a specific bridging model addressing the problem of linking event logs to their measurement data provenance, called

MAED (Measurement And Event Data). In addition to testing this proposal with respect to the mentioned criteria, we also offer an initial practical validation through an expert panel from different engineering disciplines, and the experimental development of transformers from actual measurement data to the model. Moreover, we study one exploitation potential of MAED on the AI analytics side, i.e., its potential usage to integrate concepts of the measurement stage into Digital Shadow creation by object-centric process mining via OCEL 2.0.

4.1 Formal and Technical Design Criteria for Bridging Data Models: Insights from OCEL 2.0

Before we embark on the data model design, it seems worthwhile to fix some formal properties such models should have, as well as on the requirements concerning the used database technologies. The requirements formulated here can be seen as a database-centric IS engineering view on experiences gained originally in a process mining context, culminating in the Object-Centric Event Log (OCEL 2.0) bridging model [7]. To follow the subsequent discussion, please also refer to Figure 3.

The need to include both *static and dynamic aspects* in conceptual modeling and data management goes back to early efforts to combine ideas from Entity-Relationship and relational databases, with Petri net models and transaction processing in the late 1970's. Yet, data-oriented and the process-oriented IS engineering subcommunities remain clearly recognizable in conferences such as CAiSE even today. However, a bridging data model must clearly address both perspectives to enable sufficiently rich and selective analytics. In the process mining community, the quest for a "model in the middle" started with standardized file formats such as IEEE XES [2] which serve as an intermediary format between data extraction from ERP systems, and process analytics software. Only from problematic experiences with early attempts at object-centric process mining, the new OCEL 2.0 has emerged from research to address object-centric process mining use cases [3] which carefully differentiates the *object* concept to a degree that significantly extends the versatility with respect to many different object-focused as well as process-focused types of analysis, based on a growing catalog of reusable analytics tools [7].

Such broad applicability, however, requires two additional formal aspects. First, it is extremely important not just to elaborate the important aspects of objects and events, but also to offer a *rich set of relationships* among them, not just structurally but also positioned with shared context aspects such as time or—in geo-intensive applications—space; filters (qualifiers) enable a more narrow focus of analysis in such relationships. At the implementation level, foreign keys are essential to materialize these relationships – one more reason that their use including underlying *unique identifiers* must be included in more engineering management practice.

Second, the evolving landscape of data and analytical methods necessitates adaptive perspectives on data management, particularly for decision makers

seeking diverse viewpoints for strategic analysis. An optimal "model in the middle" must facilitate not only schema evolution but also support the coexistence of multiple schema organizations. This concept, rooted in the innovations of deductive database research from the early 1980s, involves integrating data and its schema within a unified framework. This amalgamation approach, now pivotal in various semantic data management areas, enables dynamic schema modifications and multiple, parallel data representations, enhancing flexibility and responsiveness to changing analytical and operational requirements.

However, most of these attempts required significant algorithmic research to address the *performance challenges* associated with amalgamation. For example, research in [14] employs RDF knowledge graphs for comprehensive modeling (schema) and execution (instance) of Digital Shadow structure and process as in [25]. While it demonstrated many of the needed aspects, massive performance problems have prevented its use in practice. Figure 3 shows how OCEL 2.0 addresses the amalgamation in a relational setting, having tables for both schema and instance data. This approach cannot just profit from long experience with similar methods in SQL servers, but also permits, e.g., special-purpose main memory databases for interactive analytics even with massive event data.

From a practical viewpoint, a bridging data model is only useful if its content can be easily filled using simple, generic, and robust transformation mechanisms from legacy, use-case-specific data models. Such transformers are not only pivotal for integrating diverse data sources but also for ensuring the scalability and adaptability of data models in dynamic industrial settings.

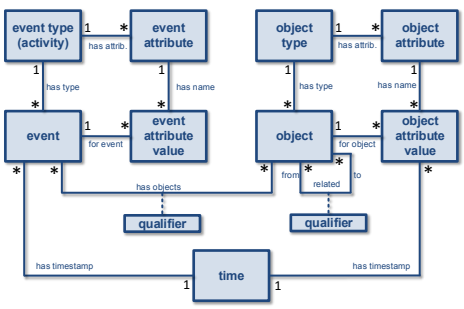

Fig. 3. OCEL 2.0 Metamodel [7].

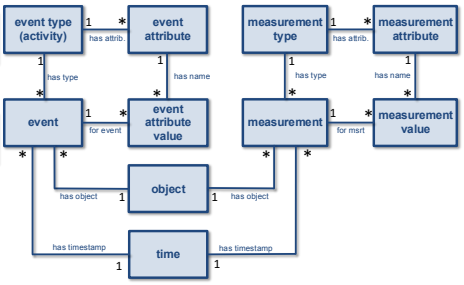

Fig. 4. Measurement and Event Data (MAED) model for Machine Data.

4.2 MAED: A Bridging Data Model for Harmonizing Sensor-Generated Industrial Measurement Data

In manufacturing, the challenge of harmonizing vast streams of sensor-generated data with the analytical needs of CPPS is substantial. We propose the MAED data model for capturing and standardizing time-series data and event records

from manufacturing processes (see Figure 4). For a detailed introduction to the MAED data model, see [30]. Recognizing the pioneering efforts of OCEL 2.0 in establishing a robust framework for object-centric process mining, our approach to MAED was informed and inspired by the foundational principles and structural components of OCEL 2.0. This was a deliberate choice, grounded in the rationale that event-driven data points, central to both OCEL 2.0 and MAED, present a complex domain where prior advancements can significantly accelerate innovation and applicability in related fields. The seamless integration of measurement and event data is pivotal for enriching analytics, enhancing decision-making, and ultimately fostering the development of more responsive and efficient CPPS.

At its core, the MAED format requires minimal, yet critical data attributes for each entry: the precise time of data capture, the nature of the recorded information, and the identification of its physical origin within the manufacturing system. Data points within the MAED schema are categorized as "events" (Figure 4, left) or "measurements" (Figure 4, right). Events are singular occurrences that mark transitions or alterations in state, carrying significance even when devoid of detailed data. A simple event like "machine overheated" suffices to signal a system's condition. In turn, measurements are systematically captured and expected readings that depict a machine's operational state through their values, which can reveal normal function or indicate anomalies like sensor faults.

"Time" is central, providing the temporal context and enabling the chronological reconstruction of events and states. The "object" identifier is equally critical, enriching the data with spatial context and relevance.

By consolidating events and measurements into a uniform structure with clear specifications, the MAED metamodel facilitates the assembly of individual data points into comprehensive sequences for advanced analysis. This provides a framework for creating data sets that are more readily comparable and analyzable across different machines or processes.

4.3 Preliminary Evaluation

In accordance with the validation phase of the DSRM, the proposed data model underwent a preliminary user evaluation during a workshop. It convened around 30 engineering researchers from diverse domains, leading to the collective affirmation of the fundamental principles of the proposed model.

Further, we collaborated with five data owners within the IoP and an external partner, on transformers of their datasets into our specified format to test its practical applicability and effectiveness. The original data formats included collections of CSV files, JSON files with complex nestings, untyped text files from a MinIO database export, and a complete PostgreSQL database dump. As a consequence, no two datasets could be processed or visualized using the same methodology or tools initially. However, once the datasets were transformed into the MAED format, they were seamlessly integrated and became compatible with preliminary tooling, underscoring the robustness and versatility of the approach

in standardizing and automating data processing for effective analysis and visualization. As initial proofs-of-concept, we created a Python library for handling the data, and two visualization widgets. In this widget, measurement and event types can be specified to be rendered below each other.

The successful transformation of datasets into the MAED format across five distinct examples not only substantiates the feasibility of our approach but also highlights intricacies of data structures at both logical and physical levels. Further easing the transformation process requires foundational prerequisites, like the inclusion of explicit foreign key relationships, thereby streamlining data integration and enhancing effective automation. Looking at the previously specified formal design criteria for such models, the successful transformer experiments and positive workshop feedback offer strong evidence of a good match of MAED to current practices and its potential.

However, the formal criteria are only partially satisfied. While we have rich relationships and schema-instance amalgamation for measurement and event data, and a time concept as in OCEL 2.0, the same has not yet been achieved for the integration of the object concept. Thus, one main usage idea of MAED—embedding extremely fine-grained and massive measurement data from continuous processes into the object-centric process mining world of OCEL 2.0—remains a non-trivial challenge for more sophisticated analyses and thus opens the avenue for significant further technical research.

Regarding the envisioned enhanced AI integration, the "model in the middle" approach enables a seamless and standardized application of advanced AI services across various domains. This standardization unlocks the potential for employing advanced AI methodologies, such as few-shot learning with large language models (LLMs) for domain-specific language (DSL) model generation, where previously, the absence of uniform data models limited the applicability of such technologies [5]. Beyond this, standardized data formats pave the way for AI-driven anomaly detection, predictive maintenance, and optimization algorithms that can now be more readily integrated and operationalized across different manufacturing environments.

5 Conclusions, Limitations, and Future Work

This paper addressed the critical challenge of bridging the divide in Industry 4.0 between a multitude of data models and diverse data-driven analytical technologies. It proposed the use of standardized intermediary models, a strategy that reduces complexity and enhances reuse across various organizational contexts.

In summary, our contributions are manifold. By utilizing empirical methods, we have opened a novel avenue to structure data diversity into categories, provided a practical example of a "model in the middle" in a mechanical engineering context, and yielded positive initial experiences with the new data format. This advancement marks a significant step towards enabling artificial intelligence methods to work more effectively with comprehensive, real-time manufacturing data, leading to smarter, more adaptive, and efficient production systems.

The empirical study of over 80 data model practices in an applied research context confirmed that there are several clusters of sufficiently similar practices within and beyond individual engineering disciplines that could scope the requirements and potential advantages for such models. Further validation directly in industry or from analysis of published case studies should promote deeper understanding and identification of other "high potentials".

Transitioning from a diverse array of data models to a small number of standardized models presents a series of organizational implementation challenges. Organizations may encounter resistance due to existing investments in custom data models. The transition may necessitate significant effort, time, and resources, potentially acting as a deterrent for some stakeholders. Moreover, the absence of established metrics for evaluating the efficiency and effectiveness of the proposed "model in the middle" approach poses another limitation. Without a benchmark, it becomes challenging to quantitatively assess the impact and benefits of our approach, beyond the formal and practice-oriented criteria proposed in this paper.

The integration of the MAED model with additional proposed data models offers substantial benefits, particularly in enhancing AI integration within production engineering. Effectively linking the data dimensions—machine, process, and product—facilitates the creation of comprehensive event logs, which are instrumental for analysis through generic process mining tools. This integration not only requires a more nuanced representation of entities like product types and hierarchies but also marks a critical step towards realizing a holistic and integrated data analysis approach. Such an approach significantly contributes to the advancement of AI applications in production engineering, as it leverages the comprehensive insights provided by the "models in the middle", ensuring that AI algorithms can access a richer, more structured pool of manufacturing data for enhanced decision-making and optimization.

Our "models in the middle" approach strategically positions itself between domain-specific standards, such as OPC-UA Companion Specifications, and general AI frameworks and libraries. This unique placement facilitates a critical linkage, enabling integration of specialized industrial protocols with advanced AI analytical frameworks. Future work will provide interoperability tests with existing IoT platforms and AI analytics tools to validate and refine this connection, aiming to close a significant gap in the current ecosystem.

Furthermore, while our approach offers a promising framework for enhancing AI integration in smart manufacturing, the aspects of scalability and real-time data processing have not been extensively explored in this paper. Future research will need to assess the scalability of our models, identifying computational and architectural optimizations to handle large-scale, real-time data streams effectively. This evaluation is crucial for ensuring that our approach can support the dynamic and expansive nature of smart manufacturing environments.

Acknowledgements Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC-

2023 Internet of Production - 390621612. We thank the Alexander von Humboldt (AvH) Stiftung for supporting our research. We express our gratitude to all participants in our study.

References

1. Engineering data exchange format for use in industrial automation systems engineering - Automation markup language. Standard IEC 62714-1 (Jun 2014)
2. IEEE Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams (2016), iSBN: 9781504424219
3. van der Aalst, W.: Concurrency and Objects Matter! Disentangling the Fabric of Real Operational Processes to Create Digital Twins. In: Intl. Colloquium on Theoretical Aspect of Computing. pp. 3–17. Springer LNCS 12819 (2021)
4. Bader, S., Pullmann, J., Mader, C., Tramp, S., Quix, C., Müller, A.W., Akyürek, H., Böckmann, M., Imbusch, B.T., Lipp, J., Geisler, S., Lange, C.: The International Data Spaces Information Model – An Ontology for Sovereign Exchange of Digital Content. In: Pan, J.Z., Tamma, V., d'Amato, C., Janowicz, K., Fu, B., Polleres, A., Seneviratne, O., Kagal, L. (eds.) The Semantic Web – ISWC 2020, vol. 12507, pp. 176–192. Springer International Publishing, Cham (2020)
5. Baumann, N., Diaz, J.S., Michael, J., Netz, L., Nqiri, H., Reimer, J., Rumpe, B.: Combining retrieval-augmented generation and few-shot learning for model synthesis of uncommon dsls. Gesellschaft für Informatik e.V. (2024)
6. Bazaz, S.M., Lohtander, M., Varis, J.: Availability of manufacturing data resources in digital twins. Procedia Manufacturing **51**, 1125–1131 (2020)
7. Berti, A., Koren, I., Adams, J.N., et al.: OCEL (Object-Centric Event Log) 2.0 Specification. Chair of Process and Data Science, RWTH Aachen University (2023)
8. Brauner, P., Dalibor, M., Jarke, M., et al.: A computer science perspective on digital transformation in production. ACM Transactions on Internet of Things **3**(2, article 15), 1–32 (2022)
9. Brecher, C., Padberg, M., Jarke, M., van der Aalst, W., Schuh, G.: Internet of Production: Interdisciplinary Visions and Concepts for the Production of Tomorrow. Springer Nature (2023)
10. Brockhoff, T., Heithoff, M., Koren, I., et al.: Process Prediction with Digital Twins. In: Models@run.time Workshop at MODELS'21 (2021)
11. Correia, J., Abel, M., Becker, K.: Data management in digital twins: a systematic literature review. Knowledge and Information Systems **65**, 3165–3196 (2023)
12. Gantz, J., Reinsel, D.: The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC Analyze the Future (2013)
13. Geisler, S., Vidal, M.E., Cappiello, C., et al.: Knowledge-driven data ecosystems towards data transparency. ACM Journal of Data and Information Quality JDIQ **14**(1, article 3), 1–13 (2022)
14. Gleim, L., Pennekamp, J., Liebenberg, M., et al.: FactDAG: Formalizing Data Interoperability in an Internet of Production. IEEE Internet of Things Journal **7**(4), 3243–3253 (2020)
15. Groeger, C.: There is no AI without data. Comm. ACM **64**(11), 98–108 (2021)
16. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design Science in Information Systems Research. MIS Quarterly: Management Information Systems **28**(1), 75–105 (2004)

17. Jarke, M.: Data sovereignty and the internet of production. In: International Conference on Advanced Information Systems Engineering – CAiSE 20. pp. 549–558. Springer (2020)
18. Jones, D., Snider, C., Nassehi, A., Yon, J., Hicks, B.: Characterizing the digital twin: A systematic literature review. CIRP Journal of Manufacturing Science and Technology **29**, 36–52 (2020)
19. Lefrançois, M., Garcia-Castro, R., Bouter, C., Poveda-Villalon, M., Daniele, L., Gnabasik, D.: SAREF: the smart applications REFerence ontology (2020)
20. Lenzerini, M.: Direct and Reverse Rewriting in Data Interoperability. In: International Conference on Advanced Information Systems Engineering. pp. 3–13. Springer (2019)
21. Liebenberg, M., Jarke, M.: Information systems engineering with Digital Shadows: Concept and use cases in the Internet of Production. Information Systems **114**, 102182 (2023)
22. Lin, S.W., Watson, K., Shao, G., Stojanovic, L., Zarkout, B.: Digital Twin Core Conceptual Models and Services. Industrial IoT Consortium Framework Publication (2023-08-01)
23. Loucopoulos, P., Kavakli, E., Chechina, N.: Requirements Engineering for Cyber Physical Production Systems. In: International Conference on Advanced Information Systems Engineering. pp. 276–291. Springer (2019)
24. Merino, J., Xie, X., Moretti, N., Chang, J., Parlikad, A.: Data integration for digital twins in the built environment based on federated data models. In: Smart Infrastructure and Construction. pp. 1–18. No. 2300002, Proceedings of the Institutions of Civil Engineers (2023)
25. Michael, J., Koren, I., Dimitriadis, I., et al.: A Digital Shadow Reference Model for Worldwide Production Labs. In: C. Brecher et al., eds: Internet of Production. pp. 1–29. Springer (2023)
26. OPC-Foundation: The industrial interoperability standard (2023), `https://opcfoundation.org/developer-tools/documents/?type=Specification`, accessed 2023-07-27
27. Pennekamp, J., Henze, M., Schmidt, S., et al.: Dataflow Challenges in an Internet of Production: A Security & Privacy Perspective. In: Proc. ACM Workshop on Cyber-Physical Systems Security & Privacy. pp. 27–38. ACM (2019)
28. Plebani, P., Salnitri, M., Vitali, M.: Fog computing and data as a service: a goal-based modeling approach to enable effective data movement. In: CAiSE 2018 Tallinn/Estonia. pp. 203–219. Springer LNCS 10816 (2018)
29. Stary, C.: Digital twin generation: re-conceptualizing agent systems for behavior-centered cyber-physical system development. Sensors **21**(1096), 1–24 (2021)
30. Tacke Genannt Unterberg, L., Koren, I., van der Aalst, W.M.: Maximizing reuse and interoperability in industry 4.0 with a minimal data exchange format for machine data. In: Modellierung 2024, pp. 103–118. Gesellschaft für Informatik e.V., Bonn (2024)
31. Vila, M., Sancho, M.R., Teniente, E.: Modeling Context-Aware Events and Responses in an IoT Environment. In: CAiSE 2023, Zaragoza/Spain. pp. 71–87. Springer LNCS 13901 (2023)
32. Volz, F., Sutschet, G., Stojanovic, L., Uslaender, T.: On the role of digital twins in data spaces. Sensors **23**(7601), 1–21 (2023)
33. Ziegler, J., Reimann, P., Keller, F., Mitschang, B.: A Metadata Model to Connect Isolated Data Silos and Activities of the CAE Domain. In: CAiSE 2021, Leuven/Belgium. pp. 213–228. Springer LNCS 12751 (2021)