machine learning & knowledge extraction

MDPI

Article

Demonstrating Data-to-Knowledge Pipelines for Connecting Production Sites in the World Wide Lab

Leon Gorissen [1,*], Jan-Niklas Schneider [1], Mohamed Behery [2], Philipp Brauner [3], Moritz Lennartz [4], David Kötter [5], Thomas Kaster [1], Oliver Petrovic [5], Christian Hinke [1], Thomas Gries [4], Gerhard Lakemeyer [2], Martina Ziefle [3], Christian Brecher [5] and Constantin Häfner [6,7]

1 Chair for Laser Technology, RWTH Aachen University, 52074 Aachen, Germany; jan-niklas.schneider@llt.rwth-aachen.de (J.-N.S.); thomas.kaster@llt.rwth-aachen.de (T.K.); christian.hinke@llt.rwth-aachen.de (C.H.)
2 Knowledge-Based Systems Group, RWTH Aachen University, 52074 Aachen, Germany; gerhard@cs.rwth-aachen.de (G.L.)
3 Chair for Communication Science, RWTH Aachen University, 52074 Aachen, Germany; brauner@comm.rwth-aachen.de (P.B.); ziefle@comm.rwth-aachen.de (M.Z.)
4 Institute for Textile Technology, RWTH Aachen University, 52074 Aachen, Germany; moritz.lennartz@ita.rwth-aachen.de (M.L.); thomas.gries@ita.rwth-aachen.de (T.G.)
5 Chair for Machine Tools, RWTH Aachen University, 52074 Aachen, Germany; d.koetter@wzl.rwth-aachen.de (D.K.); o.petrovic@wzl.rwth-aachen.de (O.P.); c.brecher@wzl.rwth-aachen.de (C.B.)
6 Faculty of Mechanical Engineering, RWTH Aachen University, 52074 Aachen, Germany
7 Fraunhofer-Gesellschaft, 80686 Munich, Germany
* Correspondence: leon.gorissen@llt.rwth-aachen.de

Abstract

The digital transformation of production requires methods for integrating, storing, and operationalizing data across organizational boundaries, yet most existing approaches remain siloed and unidirectional, lacking a systematic loop from raw data to actionable knowledge and back. We introduce Data-to-Knowledge (D2K) and Knowledge-to-Data (K2D) pipelines as a universal production concept built on networks of Digital Shadows. The Data-to-Knowledge (D2K) pipeline is realized as a cross-organizational proof of concept that captures and semantically annotates robotic trajectory data from three independent research institutes and uses those data to train an inverse-dynamics foundation model for robot control. Centralized aggregation via an existing FAIR-compliant research data repository was chosen deliberately over federated alternatives to maximize semantic interoperability and reuse of shared infrastructure; federated and privacy-preserving extensions are identified as a promising future direction. Fine-tuning the cross-organizationally trained foundation model reduces training time by approximately 85% relative to end-to-end training from scratch, while achieving comparable accuracy on a standardized inverse-dynamics benchmark. These gains are attributable to the combination of cross-site data aggregation and transfer learning; isolating the contribution of semantic annotation alone remains a topic for future ablation work. The implementation demonstrates that semantically enriched, cross-organizational D2K pipelines can accelerate model development and reduce redundant data collection within a constrained but practically relevant class of robotics tasks. We further discuss limitations, governance challenges, and how these pipelines can contribute to a broader World Wide Lab for collaborative production research.

Keywords: Industry 4.0; digital twin; digital shadow; foundation models; transfer learning; data pipelines; semantic web

Academic Editor: Mehmed Kantardzic

Received: 14 April 2026
Revised: 9 May 2026
Accepted: 14 May 2026
Published: 20 May 2026

1. Introduction

Modern manufacturing is undergoing a fundamental shift driven by the convergence of physical production systems with digital infrastructure. Industry 4.0 initiatives aim to integrate Artificial Intelligence (AI), Internet of Things (IoT), and data-driven decision-making into production processes, promising increased efficiency, flexibility, and sustainability [1–4]. Central to this transformation is the ability to collect, share, and exploit data across organizational boundaries: from individual machines and production cells, through supply chains, to global research collaborations. However, production data remains largely siloed. Existing data architectures—warehouses, lakes, lakehouses, meshes, and fabrics—each address specific aspects of the data challenge but rarely provide the bidirectional, semantically enriched flows required to turn raw sensor streams into reusable, cross-organizational knowledge assets [5–7]. The result is redundant experimentation, fragmented model development, and an inability to accumulate knowledge systematically across institutions.

D2K and Knowledge-to-Data (K2D) pipelines address this gap directly. A D2K pipeline transforms raw, context-dependent data—sensor readings, robot trajectories, process logs—into actionable knowledge through cleaning, semantic annotation, and machine learning. A K2D pipeline applies that knowledge to guide data collection strategies, closing the loop: knowledge informs what data to collect, how to process it, and how to reconfigure production activities accordingly. Together, these bidirectional pipelines provide the connective tissue for a World Wide Lab (WWL) —a globally interconnected ecosystem of data and knowledge-sharing infrastructures inspired by the World Wide Web (WWW) [8]. The World Wide Lab (WWL) vision emphasizes reducing redundant experimentation, improving model quality through diverse empirical data, and enabling cumulative, institutionally shareable model-building across organizational boundaries.

The key design choice motivating this work is the use of Digital Shadows (DSs)—task- and context-specific, modular data representations—as the foundational unit of both pipelines [9,10]. Unlike monolithic Digital Twins (DTs), DSs are purpose-driven projections that can be composed into networks, enabling hierarchical and cross-organizational data flows without requiring a single unified system. This modularity is what makes the D2K and K2D pipelines practically deployable in heterogeneous, multi-institutional settings.

We demonstrate these pipelines in a concrete proof of concept: trajectory data from three Franka Emika robots operating in distinct research settings—laser material processing, textile fiber draping, and gear assembly—are captured, semantically annotated, and aggregated in a shared Findable, Accessible, Interoperable, Reusable (FAIR)-compliant repository. This aggregated dataset trains an inverse-dynamics foundation model, which is then fine-tuned for instance-specific control. The D2K direction is fully realized and empirically evaluated. The K2D direction, in which knowledge (e.g., a fine-tuned model) guides data collection strategy, is architecturally anticipated but not yet empirically closed; realizing the full loop remains an open task.

The contributions of this paper are:

1. A conceptual framework for D2K and K2D pipelines built on networks of DSs, positioning them as the foundational elements of the WWL.
2. A fully realized cross-organizational D2K pipeline that aggregates semantically annotated trajectory data from three independent institutions using an existing FAIR-compliant research data infrastructure and trains a reusable inverse-dynamics foundation model.
3. Quantitative benchmark evidence showing that fine-tuning the foundation model reduces training time by approximately 85% while achieving accuracy within the torque

sensor noise floor, compared to an end-to-end baseline following the architecture of Schneider et al. [11].
4. A hybrid pipeline orchestration combining scheduled and event-driven data flows, validated in a live multi-institutional deployment.
5. An analysis of the resulting governance, scalability, and sustainability implications, including an identification of open challenges for K2D feedback and federated extensions.

This article is structured as follows. Section 2 reviews related work and identifies the research gap. Section 3 describes the conceptual foundation and the materials and methods of the D2K pipeline. Section 4 presents the benchmark results. Section 5 discusses findings, limitations, and implications. Section 6 concludes with an outlook on future research.

2. Background and Related Work

The industrial revolutions advanced manufacturing from steam-powered machinery and mass production to digital automation and data-driven decision making. Industry 4.0 aims at integrating legacy systems, disparate systems, and data silos into cohesive frameworks to utilize data effectively [2–4].

Data architectures evolved to address data challenges [5,6]. Data warehouses centralize structured data for analytics [12], yet are task-specific, as seen with Enterprise Resource Planning (ERP) systems, which lack broader integration with tools like Manufacturing Execution System (MES). Data lakes store raw, multi-format data but pose querying complexity [13]. Data lakehouses combine warehouse and lake capabilities [14]. More recently, data meshes enable decentralized, domain-specific data ownership [7], and data fabric integrates distributed data sources, enhancing governance [15]. While these architectures address storage and access, none provide native support for the bidirectional, semantically enriched flows between data and actionable knowledge that adaptive production systems require. Building on top of these data architectures are conceptual frameworks that partially address this gap.

2.1. From Siloed Data to Networked Production Ecosystems

The Internet of Production (IoP) and WWL extend inner-organizational data architectures to foster collaboration across supply chains and research labs [8,16]. The IoP connects various production entities from design to logistics, creating a network that integrates sensors, machines, and planning tools across supply chains. The WWL, inspired by the WWW, aims to provide a global platform for shared innovation, enabling cross-company, hierarchical, and horizontal data integration [8], preventing redundant experimentation, and improving model quality through empirical data from diverse manufacturing settings. Both are deliberately conceptual frameworks: they define the vision and the requirements for a globally collaborative production ecosystem without prescribing a specific technical implementation. This paper contributes a concrete realization of that vision—a working pipeline that closes the loop from raw cross-organizational data to reusable knowledge—demonstrating that the WWL concept is practically achievable with existing infrastructure.

Initiatives within the IoP, such as ProducTron (intra-company modularity [17]) and FactDAG (provenance-based data interoperability [18]), demonstrate specific aspects of networked production data but focus on data management rather than the full D2K/K2D pipeline. Blockchain-based systems such as the Trustworthy Information Store [19] address accountability but not cross-organizational knowledge transfer.

Academic shared-lab efforts—MIT's iLab [20], Spain's remote experimentation infrastructure [21], and Labicom [22]—demonstrate scalable laboratory sharing but predate the semantic web and FAIR data requirements that are central to our approach. Carnegie Mel-

lon's Manufacturing Futures Institute [23] and UCLA's Smart Manufacturing Institute [24] represent more recent convergences of digital twins and industrial data ecosystems, yet neither provides a bidirectional data-to-knowledge pipeline anchored in semantic Digital Shadows (DSs).

Industrial initiatives including Gaia-X [25], Catena-X [26], and Manufacturing-X [27] promote federated, sovereignty-preserving data sharing, primarily within specific industries. NVIDIA Omniverse [28] focuses on simulation interoperability. These are complementary to the WWL but address different layers of the problem.

2.2. From Standalone Systems to Unified Data Pipelines

Modern production systems demand continuous and near real-time data integration across heterogeneous sensors and logs [29,30]. Approaches range from scheduled and poll-based pipelines to event-based architectures [31], with the latter enabling low-latency, scalable decoupling. Automated validation techniques such as Auto-Validate-by-History [32,33] address schema drift, while semantic annotation pipelines [34–37] add ontology-based context.

Bodenbrenner et al.'s FAIR Sensor Ecosystem [37] focuses on contextualizing high-frequency sensor data with a unified metamodel and temporal versioning, improving FAIRness—yet it does not close the loop into model training or knowledge feedback. Bi et al. [36] convert OPC UA models into RDF/OWL via Querying of Ontology Mapping-based OPC UA (QOMOU) and introduce semantic similarity scoring to handle heterogeneous industrial devices; an approach of this kind could, in the future, extend our pipeline to enable automatic discovery and onboarding of OPC UA-compliant machines, reducing the current manual RDF lifting effort. Critically, none of the existing semantic pipeline approaches provide comprehensive bidirectional flows where analytics outputs systematically inform data collection strategies. Our D2K/K2D framework is specifically designed to close this gap.

2.3. From Digital Twins to Digital Shadows

A cornerstone of the WWL, D2K, and K2D is the use of DSs, defined as "task- and context-dependent, purpose-driven, aggregated, multi-perspective, and persistent datasets" [9] and are analogous to views in relational databases [10,38]. Unlike DTs, which are often tied to a single comprehensive representation, DSs allow multiple purpose-specific projections of the same entity, enabling modularity and scalability [10]: for instance, one DS for high-precision simulation and another for real-time control [39]. This modularity is the key design choice enabling the proposed pipelines to function across organizational boundaries without a monolithic shared system.

DSs support hierarchical scalability at multiple levels, from individual machine instances to generalized machine types [40], informed by separation of concerns [41] and microservices architecture [42]. FactDAG [18] demonstrates how DSs leveraging Directed Acyclic Graphs (DAGs) can maintain data provenance aligned with FAIR principles. Recent work by Heithoff et al. [43] systematically applies the DS Reference Model [44] across automation pyramid levels in injection molding, reinforcing modularity and scalability. Validating DSs as core enablers for dynamic, multi-level pipelines that "share digital shadows across organization boundaries" [43] remains an open task that this work partially addresses.

2.4. Federated Learning as a Related Paradigm

Cross-organizational learning under data-sharing constraints is also addressed by federated learning. Tong et al. [45] propose a Federated Heterogeneity-aware Adaptive framework (FedHA), combining adaptive asynchronous aggregation with hierarchical knowledge distillation to improve prediction accuracy under heterogeneous client condi-

tions. Chen et al.'s Federated graph learning via Constructing and Sharing Feature spaces (FedCSF) [46] builds a globally consistent feature space for cross-domain IoT graph learning, avoiding feature-space contamination without sharing raw data. Chahoud et al. [47] use deep Reinforcement Learning (deep RL) to spin up containerized clients on demand, cutting training rounds by 20–50%. Wang et al. [48] designed an encryption scheme enabling public model integrity verification while maintaining approximately 95% accuracy.

In contrast to these federated approaches, we deliberately aggregate trajectories centrally to produce a reusable foundation model and fine-tune only a few layers per instance—achieving large runtime savings without the communication overhead of federated aggregation. This design choice is appropriate for our academic consortium setting, where all partners have agreed to share raw data. Federated and privacy-preserving extensions—such as those of [45,48]—are attractive for future D2K deployments where partners cannot or will not share raw data, and they are identified as a priority future direction in Section 6.

2.5. Research Gap

Synthesizing the above, existing work addresses individual components of the data-to-knowledge challenge: semantic annotation [36,37], data interoperability [18], shared infrastructure concepts [8,16], and cross-organizational learning [45,46]. What is missing is a fully realized, end-to-end pipeline that (a) captures and semantically annotates heterogeneous production data from multiple independent organizations, (b) uses that data to train a reusable cross-organizational foundation model, (c) enables instance-specific fine-tuning with quantitative runtime and accuracy evidence, and (d) explicitly anticipates and architecturally provisions for K2D feedback. This paper provides exactly that demonstration within a constrained but practically meaningful class of robotics tasks on the Franka Emika platform.

3. Materials and Methods

3.1. Conceptual Framework: Data, Knowledge, Agents, and Pipelines

The proposed framework rests on three formally defined concepts and two transformative pipeline types, formalized as follows.

> Data is raw, context-dependent information collected from sources like sensors, machines, and humans. It exists in structured, semi-structured, or unstructured formats, including measurements, logs, and images. Data undergoes transformations such as cleaning and semantic annotation to become actionable knowledge.
>
> Knowledge is actionable, semantically enriched information derived from data or inherent in autonomous agents. It supports decision-making, system optimization, and continuous improvement. Knowledge is operationalized here as a trained model (or model parameters) together with its provenance metadata: the data it was trained on, the training procedure, and performance bounds. This definition aligns with foundational knowledge management literature [49,50].
>
> Autonomous agent is an entity—human or artificial—capable of perceiving its environment, processing data, making decisions, and adapting over time [51]. Agents mediate between data and knowledge: they execute D2K transformations (e.g., a training pipeline) and apply K2D feedback (e.g., reconfiguring sensor sampling based on a trained model's uncertainty). This dual role is represented in Figure 1.
>
> Data-to-Knowledge pipeline is a directed sequence of transformations from raw sensor data to actionable knowledge:
>
> 1. Data collection from various sources.

2. Data processing to transform through cleaning and annotation.
3. Knowledge generation using analytics and learning methods.
4. Action application for process optimization and decision-making.

Each step may be automated, human-in-the-loop, or hybrid. The full pipeline forms a DAG that may branch, merge, and be triggered by schedules or events.
Knowledge-to-Data pipeline is the inverse transformation, applying existing knowledge to guide data collection:

1. Existing knowledge used to inform data collection (e.g., a foundation model).
2. Influence on data collection strategies (e.g., targeted trajectory sampling).
3. Data processing shaped by knowledge (e.g., filtering based on model uncertainty).
4. Modified production activities (e.g., reconfigured robot motion).

K2D outputs may themselves trigger new D2K cycles, creating a closed adaptive loop.

Formally, a D2K pipeline P_{D2K} is a DAG of transformations $f_i : DS_i \to DS_{i+1}$ over a network of DSs, terminating in actionable knowledge K consumed and decided on by an autonomous agent. A K2D pipeline P_{K2D} maps K back to updated DSs that reconfigure data collection or processing. The composition $P_{\text{K2D}} \circ P_{\text{D2K}}$ defines the self-reinforcing adaptive cycle that distinguishes this framework from unidirectional data architectures: knowledge generated at the terminal node feeds back to reconfigure source nodes, enabling continuous adaptation without discarding accumulated cross-organizational knowledge.

The design rationale for grounding both pipelines in DSs rather than full DTs is threefold: First, DSs are task-specific and purpose-driven, avoiding the maintenance overhead of comprehensive digital twins that must mirror the full physical state [9]. Second, multiple DSs (or, put simply, views) can coexist for the same physical entity (e.g., one for control, one for provenance), enabling modular composition without schema conflicts. Third, DSs are analogous to database views [10,38]: they expose relevant subsets of the underlying data while hiding irrelevant complexity, making cross-organizational federation more tractable. The alternative—a single unified Digital Twin (DT) shared across organizations—would require agreement on a comprehensive data model, raising governance and intellectual property barriers that are currently prohibitive at the scale of the WWL vision. Figure 1 illustrates the interplay of the proposed pipelines and the role of autonomous agents.

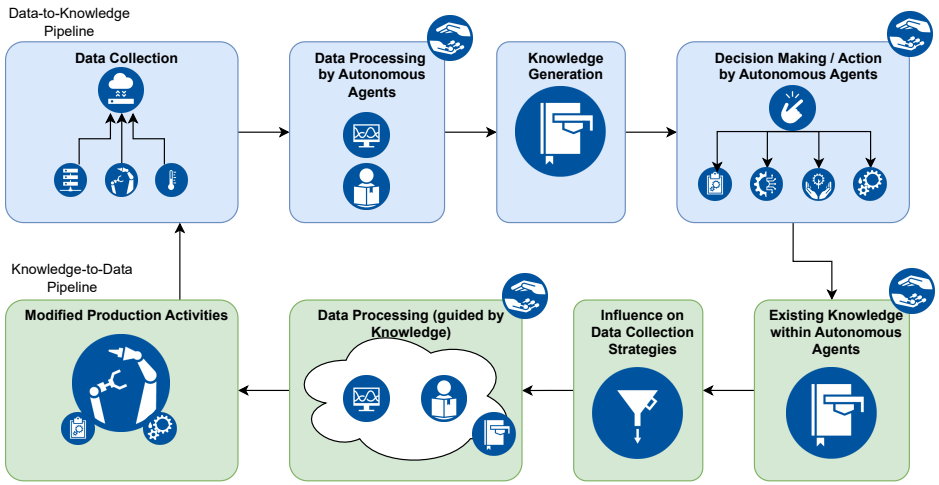


Figure 1. Illustration of the steps within the proposed pipelines. Each pipeline is a directed acyclic graph. The D2K pipeline can initiate a K2D pipeline and vice versa. The hands icon represents the agent's role (human and artificial) in each transformation step.

3.2. System Architecture

Figure 2 outlines the realized D2K pipeline within the WWL. Three distinct production sites contribute trajectory data to a shared FAIR-compliant research data repository (Coscine [52]), from which a centralized training instance queries semantically annotated data to produce and update a foundation model. Instance-specific fine-tuned models are then derived and deployed back to the individual use cases.

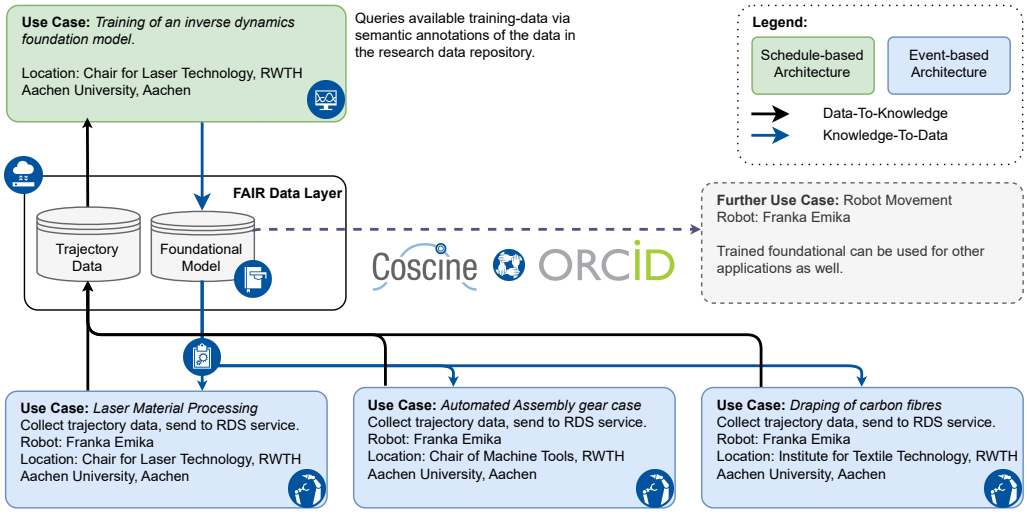


Figure 2. Network of D2K pipelines: trajectory data from three independent organizations are stored and semantically annotated in a shared research data repository (lower use cases). The training instance (upper use case) queries available training data via semantic annotations and provides a foundation model. Instance-specific models are derived by the original use cases or third parties.

The design choice of centralized aggregation over federated or privacy-preserving alternatives was made for the following reasons. All three contributing institutions are part of an academic consortium with explicit data-sharing agreements, making centralized aggregation feasible and appropriate. Centralization maximizes the semantic interoperability benefit of the shared DS infrastructure: queries can span the full aggregated dataset without communication overhead, and the FAIR metadata layer provides provenance without requiring local data to leave its source. Federated alternatives would impose communication overhead and require more complex coordination protocols. However, we explicitly acknowledge that this choice limits applicability to settings where raw data sharing is permitted; federated and privacy-preserving extensions (Section 6) are a priority for broader WWL deployment.

Instead of creating an additional data silo, we build on Coscine [52], an existing centralized research data infrastructure that manages storage, findability, and accessibility, modeling all data as FAIR Digital Objects (FDOs) based on Resource Description Format (RDF). This choice demonstrates that semantically enriched repositories already in place can unite siloed data without new infrastructure investment, and it provides a transferable blueprint adaptable to other FAIR-compliant data management solutions.

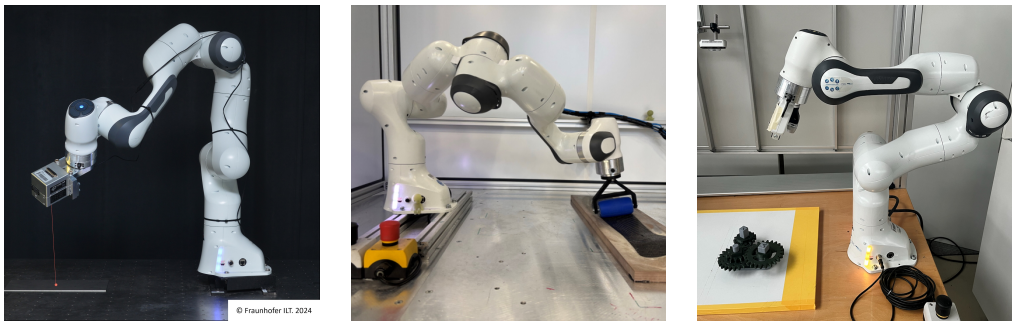
The pipeline combines schedule-based and event-based operations. Data is pushed from robots to the repository immediately following trajectory execution (event-based). Model training is triggered nightly at 2 a.m. (schedule-based), ensuring fresh aggregated training data without continuous compute overhead. This hybrid orchestration reflects a deliberate design choice: event-driven ingestion maximizes data freshness and traceability, while scheduled training decouples the computationally intensive learning step from real-time production operations.

At the machine level, DSs represent specific robots through command and attained trajectory data, fine-tuned dynamics models, and virtual scenes. Machine-level DSs aggregate data across robot types, enabling knowledge transfer and scalability, akin to sharing vision–language–action models across tasks in large-scale robotics [53]. Google DeepMind's Genie 3 demonstrates how generative world models can produce constraint-aware simulations [54]; combined with DSs, such models could complement physical simulators with greater scenario diversity.

3.3. Domain-Specific Use Cases

Three distinct production use cases contribute trajectory data to the shared pipeline, spanning different domains, task structures, and workspace constraints (Figure 3).

- Laser Material Processing (Lehrstuhl fuer Lasertechnik (Chair for Laser Technology) (LLT)/Fraunhofer Institute for Laser Technology (ILT)): A Franka Emika robot equipped with a fast beam steering device performs laser engraving on steel, where trajectory inaccuracies directly affect geometric tolerances in laser–matter interaction [11,55]. The robot operates with relatively few workspace restrictions, resulting in a broad, near-uniform joint-space distribution (Figure 4).
- Textile Fiber Draping (Institut für Textiltechnik (Institute for Textile Technology) (ITA)): A Franka Emika robot automates fiber composite preform draping for flexible manufacturing in small and medium-sized enterprises (SMEs), where careful force control is required to avoid damaging delicate textiles [56]. The constrained draping geometry yields a narrower joint-space distribution.
- Gear Assembly (Werkzeugmaschinenlabor (Chair for Machine Tools) (WZL)): A Franka Emika robot performs peg-in-hole gear assembly using inverse-dynamics-based torque control for precise positioning [57]. The highly structured assembly environment similarly constrains joint-space coverage.



(**a**) (**b**) (**c**)

Figure 3. Franka Emika robots integrated into the D2K pipeline. (**a**) LLT and ILT: laser material processing. (**b**) ITA: automated draping of fiber composites. (**c**) WZL: collaborative gear assembly.

3.4. Inverse-Dynamics Data-to-Knowledge Pipeline

The concrete D2K pipeline targets the robot inverse dynamics problem: given the desired joint configuration $\mathbf{q}$, velocity $\dot{\mathbf{q}}$, and acceleration $\ddot{\mathbf{q}}$, determine the required joint torques $\boldsymbol{\tau}$, i.e., $\boldsymbol{\tau} = f(\mathbf{q}, \dot{\mathbf{q}}, \ddot{\mathbf{q}})$ [58]. Rather than computing f analytically via inertia, Coriolis, and gravity matrices, we learn data-driven mapping from empirical trajectory data, capturing unmodelled dynamics (friction, cable effects) without explicit physical modeling.

We use the stacked Long Short-Term Memory (LSTM) architecture of Schneider et al. [11], which demonstrated competitive accuracy against conventional and data-driven state-of-the-art methods on this task. Here, we extend that architecture to the cross-organizational setting: θ_0 is trained on $\mathcal{D} = \mathcal{D}_{\text{LLT}} \cup \mathcal{D}_{\text{ITA}} \cup \mathcal{D}_{\text{WZL}}$, yielding a foundation model that generalizes

across robot instances. Instance-specific parameters θ_i are then obtained by fine-tuning up to five layers of θ_0 on a site-specific subset.

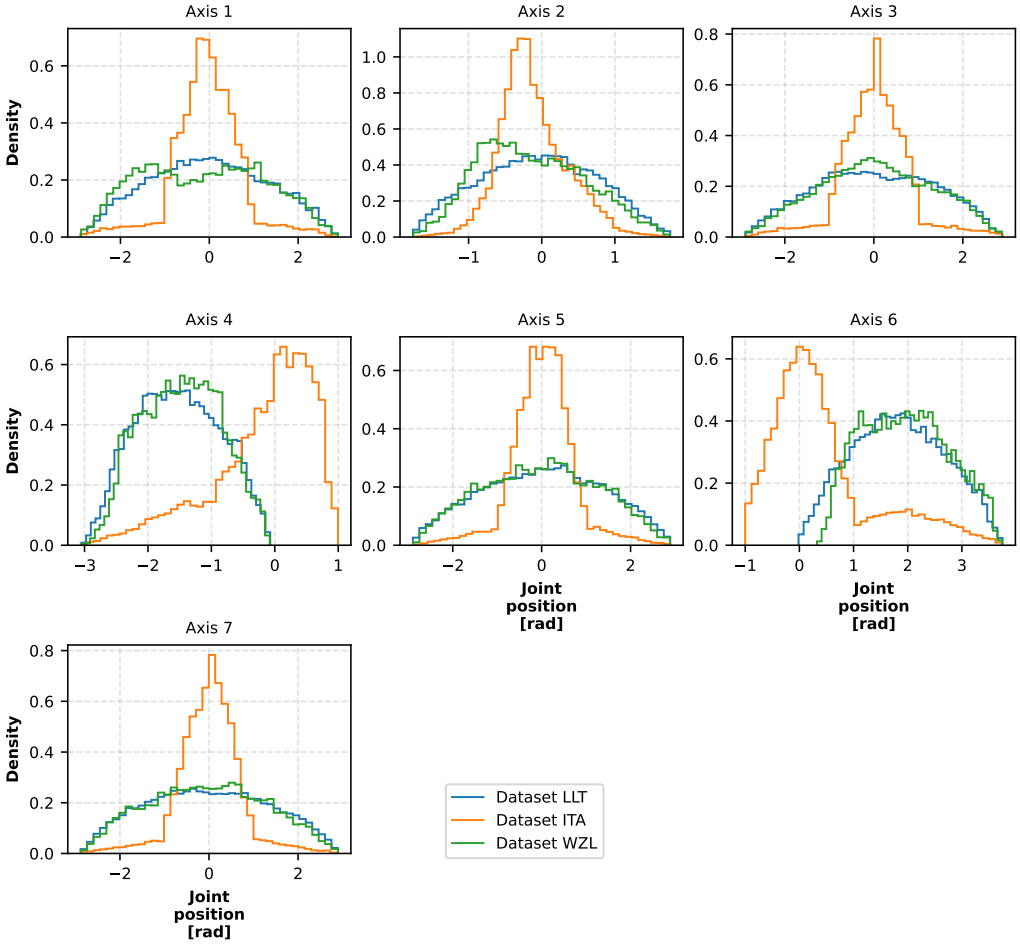


Figure 4. Histograms of joint positions (in radians) for each of the seven robot axes and the three sources: LLT (Chair for Laser Technology), ITA (Institute for Textile Technology), and WZL (Chair for Machine Tools). Bars are colorless for readability. Axis 2 highlights the broader LLT distribution (fewer workspace constraints), whereas ITA and WZL distributions are noticeably narrower due to task-specific motion limits.

Foundation model fine-tuning is chosen over independent end-to-end training for each new instance for two reasons. First, end-to-end training requires a full hyperparameter search and long training time per instance, whereas fine-tuning θ_i from θ_0 exploits the shared dynamics structure across Franka Emika instances, reducing the search space and starting from a well-initialized parameter set. Second, reusing a shared foundation model makes per-instance model development cumulative: knowledge accumulated across sites is not discarded when deploying to a new robot. The limitation of this approach is that the foundation model inherits any biases present in $\mathcal{D}$; the coverage of the joint-space distributions across sites is therefore an important data quality criterion (Figure 4).

The realized pipeline operates as follows:

1. Event-driven data ingestion: Following each trajectory execution, robot data are pushed to the Coscine repository. Data are modeled as FDOs with RDF-based metadata including velocity and acceleration scaling factors, robot instance identifiers, and git commit hashes to ensure traceability.
2. Nightly training sweep (schedule-based): At 2 a.m., the current DSs are pulled from the repository. A sweep agent initiates $n = 10$ training runs, each with a new hyperpa-

rameter configuration H sampled by the sweep server (Weights and Biases). Dataset statistics are analyzed and uploaded back.

3. Model selection: If a new model achieves $\mathcal{L}_{\text{CV}}$ below the current champion, the repository is updated with the new θ_0 and H. The selected model is evaluated on the held-out test set and results are stored.
4. On-demand fine-tuning: Instance models θ_i are fine-tuned on demand by adapting up to five layers of θ_0.

3.5. Benchmark Setup

This benchmark (https://doi.org/10.18154/RWTH-2025-00519, [59]: Source code for the D2K pipeline implementation and benchmark experiments) evaluates four training configurations for inverse-dynamics modeling. The baseline is End-to-End training, replicating the single-site approach of Schneider et al. [11]: a model is trained from random initialization on site-specific data, with a full hyperparameter search. This baseline was previously shown to achieve accuracy on par with conventional model-based methods and other data-driven approaches [11]. Three D2K variants are compared against this baseline: Foundation (fine-tuning θ_0 with a new hyperparameter search), Instance Known (fine-tuning θ_0 with the hyperparameters of the foundation model, no additional search), and Instance Unknown (fine-tuning θ_0 with a new hyperparameter search on site-specific data).

3.5.1. Dataset and Splits

The benchmark (https://doi.org/10.18154/RWTH-2025-00466, [60]: Trajectory dataset collected from the three contributing robot instances (LLT, ITA, WZL)) uses a centralized dataset $\mathcal{D}$ aggregated from three institutes (Table 1). Training trajectories consist of random joint-space motions generated via MoveIt target joint samples, with velocity and acceleration scaling factors independently sampled in $[0.1, 0.5]$. The training corpus of 2533 trajectories and 554,679 measurements per axis was divided 80/20 into training and validation samples after concatenating the sorted trajectory files; validation loss $\mathcal{L}_{\text{CV}}$ is computed on the held-out 20% and used for model selection throughout the sweep.

The held-out test set was collected on the LLT robot instance only, consisting of 28 ISO 9283 evaluation trajectories with fixed velocity and acceleration scaling factors of 0.25 and 0.1, respectively. The benchmark therefore evaluates whether models trained or fine-tuned from cross-organizational data improve performance on a held-out LLT target instance; it does not constitute a per-site evaluation across all three institutes.

Table 1. Dataset summary: measurements per axis and trajectory counts by site and split. The held-out test set was collected on LLT only.

	Measurements	Train	Val	Subtotal	Test
LLT	230,627	1027	257	1284	28
ITA	87,950	252	64	316	–
WZL	236,102	746	187	933	–
Total	554,679	2025	508	2533	28

3.5.2. Hyperparameter Search

Hyperparameter optimization used Weights and Biases Bayesian sweeps minimizing validation loss, with Hyperband early stopping (minimum three iterations) and $n = 10$ candidate models per nightly cycle. The search space is summarized in Table 2.

Table 2. Hyperparameter search space.

Hyperparameter	Range/Values	Distribution
Optimizer	Adam, SGD	Categorical
Learning rate	10^{-7}–0.9	Log-uniform
Clipnorm	1–10,000	Log-uniform
Window size	0–100	Integer-uniform
Batch size	2048, 4096	Categorical
LSTM units	1–1000	Integer-uniform
Dropout	10^{-5}–1	Log-uniform
LSTM layers	1–100	Integer-uniform
Epochs	100 (fixed)	—

3.5.3. Success Criterion

Model selection is based on $\mathcal{L}_{CV}$; a new model replaces the champion only if it achieves strictly lower cross-validation loss. Final evaluation uses mean absolute error (MAE) on the ISO 9283 test set. Values below 0.15 Nm are considered within the torque sensor inaccuracy range of the Franka Emika robot and therefore represent the practical lower bound of achievable performance.

4. Results

4.1. Dataset Characteristics

Figure 4 shows the joint-position distributions across the seven robot axes for each contributing site. LLT exhibits the broadest, most uniform distribution—consistent with its relatively unconstrained laser processing workspace. ITA and WZL show narrower, task-specific distributions reflecting their constrained operational ranges. This heterogeneity motivates cross-site aggregation: the combined dataset provides broader joint-space coverage than any single site alone.

4.2. Training Time

Figure 5 shows per-run training time across all four configurations. The End-to-End baseline requires approximately 60 h total, with a median per-run time of approximately 12 min and 3 s. Foundation fine-tuning reduces total training time to approximately 8 h and 47 min (median 1 min 45 s per run). Instance Known (reusing foundation hyperparameters) further reduces time to approximately 5 h and 33 min (median 1 min 6 s), while Instance Unknown (new hyperparameter search on site data) completes in approximately 4 h and 51 min (median 58 s). All three fine-tuning approaches outperform the end-to-end baseline in every timing metric.

Hyperparameter reuse (Instance Known) provides the largest marginal gain by eliminating exhaustive search, which can take up to 5 h per run without a good starting point.

4.3. Validation Accuracy

Figure 6 shows validation MAE across hyperparameter runs for all four configurations. All configurations achieve MAE values well below the theoretical maximum of 108.71 Nm, and the best runs across all configurations approach the 0.15 Nm sensor noise floor.

The End-to-End baseline exhibits higher initial MAE and slower convergence, with some runs showing outlier behavior during early hyperparameter search (indicative of poor initializations). With sufficient tuning, end-to-end training achieves a comparable lower-bound MAE to the fine-tuned approaches.

Foundation fine-tuning achieves the lowest MAE in early runs, reaching $\pm 1\%$ of its lower-bound MAE on the first run without additional hyperparameter refinement. This

rapid convergence is the primary practical advantage of the D2K approach: foundation models enable fast, reliable deployment in new production contexts. With extended hyperparameter tuning (beyond 200 runs), the end-to-end baseline slightly surpasses the Foundation MAE lower bound, though the difference is within the sensor noise floor and thus not practically significant.

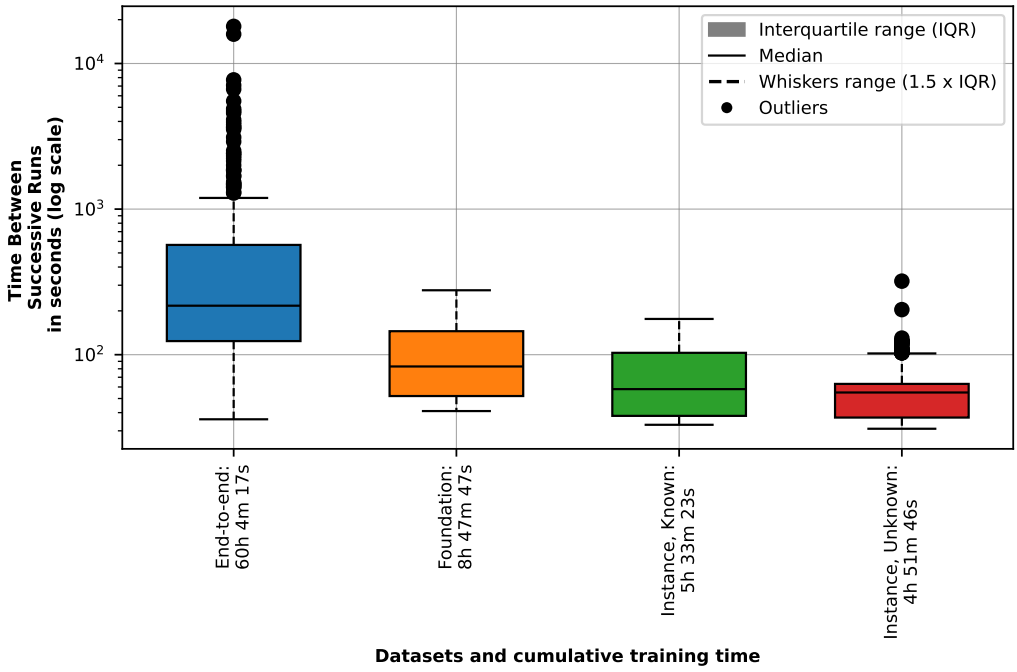


Figure 5. Per-model training time (log scale) for the four configurations: End-to-End (trained from scratch), Foundation (fine-tuned from θ_0 with new hyperparameter search), Instance Known (fine-tuned with the foundation hyperparameters), and Instance Unknown (fine-tuned with a new hyperparameter search). Fine-tuning approaches are faster than the end-to-end baseline in every metric.

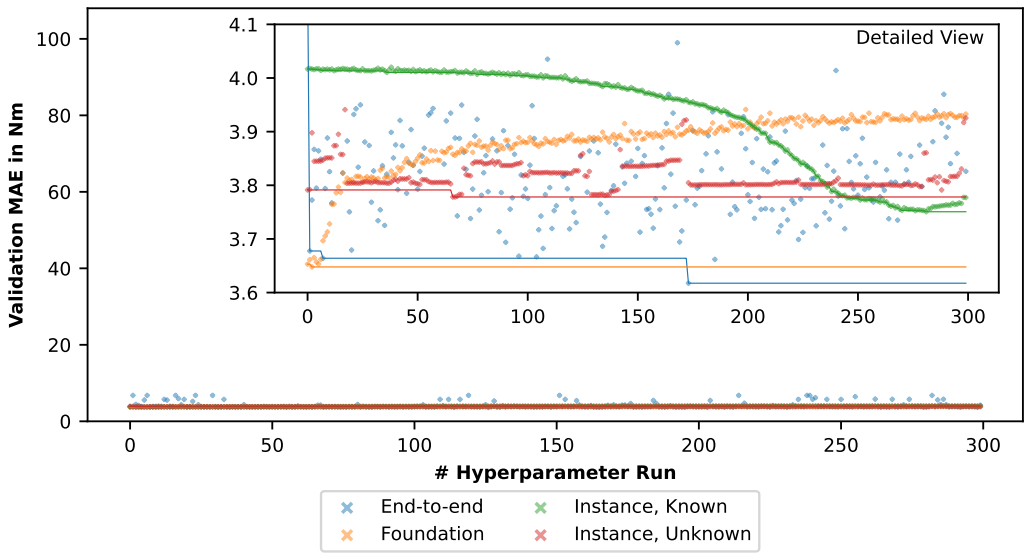


Figure 6. Validation MAE (in Nm) across hyperparameter runs (# = run index) for all four configurations. The outer plot is scaled to the maximum possible MAE; the inset details the region of interest near the sensor noise floor (0.15 Nm). Fine-tuning a foundation model (D2K approach) achieves the lowest MAE in early runs. With extensive end-to-end tuning, the baseline slightly surpasses the fine-tuning lower bound, though the difference is within torque sensor inaccuracy.

Instance Known and Instance Unknown exhibit similar convergence behavior to Foundation, with Instance Unknown showing marginally higher early variability due to the new hyperparameter search.

5. Discussion

5.1. Summary of Findings and Limitations

The realized cross-organizational D2K pipeline demonstrates that semantically enriched, FAIR-compliant data infrastructure can serve as the backbone of a multi-institutional machine learning workflow. By building on Coscine rather than a new silo, and by anchoring data flows in DSs, the system achieves data federation without requiring a monolithic shared schema.

The main empirical finding is that foundation model fine-tuning substantially reduces training time (approximately 85% reduction in total compute) while maintaining accuracy within the torque sensor noise floor. This acceleration is attributable to the combination of cross-site data aggregation and parameter transfer; we cannot currently isolate the contribution of semantic enrichment alone, since all aggregated data in this study carries the same semantic annotations. Disentangling the contribution of semantic interoperability from that of simple data volume increase is an important open question and a priority for future ablation work.

The implementation also reveals several limitations. First, all use cases employ the same robot platform (Franka Emika), which means the demonstrated transferability is within a constrained but practically meaningful class of tasks. Extending to heterogeneous robot platforms or other production domains (e.g., CNC machining, additive manufacturing) would require additional standardization of trajectory representations and may yield different transfer learning dynamics. Second, while the D2K pipeline is fully realized, the K2D direction—in which knowledge (e.g., model uncertainty) guides data collection strategy—remains an open task. Realizing the full K2D loop will require safe write-back mechanisms to production equipment, real-time model inference at data collection time, and semantic drift detection as ontologies evolve. Third, the current implementation does not address privacy or verifiability beyond the inherent access controls of the Coscine repository; all three sites share raw data. Long-term sustainability hinges on establishing cross-institutional governance structures that promote stakeholder alignment and trust.

Regarding the semantic annotation infrastructure, the current RDF-based mapping from raw robot data to FDOs involves manual ontology alignment when adding a new machine type or sensor configuration. Specifically, a domain expert must map the new machine's data schema to the existing DS metamodel and define the relevant RDF predicates. This process currently takes on the order of days for a new Franka variant and would likely require weeks for a fundamentally different robot platform. Automated semantic alignment tools (e.g., OPC UA to RDF mapping via QOMOU [36]) are an active research direction and would substantially reduce this barrier.

On federated SPARQL query performance, the current deployment issues queries against a single centralized Coscine endpoint, so federated query performance was not a bottleneck at the scale of this proof of concept (three sites, roughly 550 k measurements). For future distributed WWL deployments spanning dozens of institutions and billions of measurements, federated SPARQL performance will become a critical engineering concern; partitioning strategies, caching, and approximate query processing are identified as necessary areas of investment. With respect to end-to-end latency, the present implementation should be understood as an asynchronous model-improvement pipeline: data ingestion is event-triggered after trajectory execution, whereas model updates are deliberately deferred

to the nightly training cycle, making the system suitable for next-day model updates but not for near-real-time adaptive control.

5.2. Comparison with State-of-the-Art

The field of semantic data pipelines and agentic AI-based knowledge extraction is still emergent. Existing contributions focus on specific aspects: semantic annotation [36,37], federated learning [45–48], and shared infrastructure concepts [8,16]. Our work operationalizes bidirectional D2K and K2D pipelines over a network of DSs, closing the loop from raw data capture to actionable control knowledge—and architecturally anticipating the feedback path back to data acquisition policies—which is not addressed in the existing literature.

In robot dynamics, Schneider et al. [11] established the end-to-end LSTM approach as competitive with conventional model-based methods and other data-driven approaches on a single-site benchmark. We build directly on that architecture and extend it to the cross-organizational setting. Here, we show that Foundation (cross-site pre-training + fine-tuning) achieves comparable accuracy to End-to-End (single-site training from scratch) while offering substantially faster deployment. Ongoing work investigates transformer architectures and Physics-Informed Neural Network (PINN)-based training paradigms to further improve robustness.

Widely adopted layered frameworks—most prominently the Purdue Enterprise Reference Architecture (ISA-95), RAMI 4.0 [61], and the International Data Spaces architecture [62]—organize production data from field devices up to enterprise systems and share the form of vertical layering with our approach. However, they differ from the D2K/K2D framework in three substantive ways. First, all three are primarily descriptive and unidirectional: they define where data resides and how it flows upward, but do not prescribe a feedback loop in which knowledge actively reconfigures data collection—the K2D direction is absent. Second, they treat data at each layer as a monolithic, generic asset; our framework substitutes DSs as typed, purpose-driven nodes that can be independently evolved without altering the underlying store [9,10]. Third, all three target intra-enterprise or bilateral deployments, whereas our approach is explicitly scoped to cross-organizational data federations with FAIR governance and provenance across institutional boundaries. D2K and K2D pipelines are therefore best understood as a behavioral specification that can be layered on top of any of these structural frameworks, defining how data and knowledge must cycle between collection and application in a networked production ecosystem.

Our contribution is novel and superior to existing work in four specific aspects: (1) a fully realized, cross-organizational D2K pipeline anchored in DSs, yielding a reusable inverse-dynamics foundation model with demonstrably faster, equally accurate fine-tuning; (2) explicit architectural provisions for forthcoming K2D feedback, going beyond passive storage or one-shot aggregation; (3) a hybrid orchestration of scheduled and event-driven processes validated in a live multi-institutional deployment; (4) quantitative evidence on training-time reduction and MAE bounds relative to a well-established single-site baseline.

5.3. Implications

5.3.1. Technical Implications

The primary technical implication of this work is that cross-organizational foundation model fine-tuning is a practically viable alternative to per-site end-to-end training for inverse dynamics. The approximately 85% reduction in training time translates directly to faster deployment of new robot instances and lower compute cost per deployment. The pipeline architecture—event-driven ingestion, nightly scheduled training, and a FAIR repository as the single source of truth—is conceptually domain-agnostic. However, the empirical validation in this paper is limited to inverse-dynamics learning for one robot family.

Three further scenarios illustrate the expected transferability of the approach and identify where additional work is required. (1) In additive manufacturing quality monitoring, heterogeneous sensor streams from laser powder bed fusion machines across sites could be aggregated to train a cross-site anomaly-detection foundation model [63]; the K2D loop would then adjust in situ sampling rates based on predicted process instability. (2) In textile production, draping robots from different SMEs could share fiber composite trajectory data to build shared motion priors analogously to the robotics case presented here; the key additional challenge is semantic alignment across incompatible machine vocabularies, which OPC UA–RDF mappings [36] could partially automate. (3) In predictive maintenance, vibration and power-draw DSs from heterogeneous Cyber–Physical Production System (CPPS) platforms could be aggregated to pre-train shared anomaly models fine-tuned per machine class; the transfer-learning efficiency gains observed here are expected to generalize since the underlying mechanism—shared pre-training reducing fine-tuning cost—is independent of the specific signal domain.

In each scenario, the primary bottleneck shifts from data collection to semantic alignment and cross-organizational governance, precisely the open challenges identified throughout this paper. Transfer to other production settings beyond these illustrative cases remains plausible but requires domain-specific semantic alignment, data standardization, and empirical validation. Future technical work should focus on (a) automated semantic alignment to reduce the manual effort of onboarding new machine types, (b) transformer and PINN architectures for improved model robustness, and (c) realizing the K2D feedback loop with safe write-back mechanisms.

5.3.2. Organizational Implications

The more distinctive practical value of the D2K/K2D approach is not simply that shared production data improves model performance—this is expected—but that it reduces redundant experimentation across organizational boundaries and makes model-building cumulative and institutionally shareable. Each new institution that contributes data benefits from the foundation model without needing to collect the data that other sites already provided. This cumulative structure is the organizational analog of the WWL vision: a global collaborative laboratory where knowledge compounds rather than being rediscovered independently. Well-designed Human–Machine Interfaces minimize automation bias and support operator trust in knowledge-driven recommendations [64], while upskilling employees ensures successful adoption. The pipeline also contributes to sustainability by reducing the redundant data collection and experimentation that would otherwise occur at each site independently [65].

5.3.3. Governance Implications

The governance implications of shared digital infrastructure are substantial and distinct from the technical ones. Trust, provenance, and data sovereignty are not solved by FAIR metadata alone; they require explicit governance frameworks covering data usage rights, liability when autonomous agents reconfigure production processes, and mechanisms for provable provenance across partners. Privacy-preserving techniques [66–69] and clear governance frameworks [70,71] are essential for protecting sensitive information while fostering collaboration. Data sovereignty frameworks such as those proposed by Gaia-X [25] and International Data Spaces provide relevant technical and legal building blocks [62,72].

Shoomal et al. [73] offer a complementary analytical lens: their framework for evaluating blockchain and IoT use cases in sustainable supply chains identifies when shared digital infrastructure creates operational value and when additional governance mechanisms be-

come necessary to make multi-actor data ecosystems viable. Applied to the WWL context, their framework not only suggests that the value of shared D2K pipelines scales with the number of contributing organizations and the degree of task similarity, but that governance overhead also increases non-linearly with organizational heterogeneity—underscoring the importance of federated, auditable governance frameworks as non-technical enablers for WWL expansion.

6. Conclusions

This paper introduced D2K and K2D pipelines as the foundational elements of the WWL, grounded in networks of DSs. The key conceptual contribution is the articulation of a bidirectional framework: D2K pipelines transform raw production data into reusable knowledge assets, while K2D pipelines apply that knowledge to guide future data collection—creating a closed adaptive loop that prevents knowledge from remaining trapped in a single organization or training run.

The empirical contribution is a fully realized cross-organizational D2K pipeline: trajectory data from three distinct production settings (laser processing, textile draping, gear assembly) were captured, semantically annotated as FDOs, and aggregated in a shared FAIR repository. An inverse-dynamics foundation model trained on this aggregated dataset achieved approximately 85% reduction in per-instance training time relative to the end-to-end single-site baseline of Schneider et al. [11], while maintaining accuracy within the torque sensor noise floor (below 0.15 Nm MAE). A hybrid pipeline orchestration—event-driven ingestion plus nightly scheduled training—was validated in a live multi-institutional deployment. The K2D direction is architecturally anticipated but not yet empirically closed; it remains the primary open task.

These results demonstrate that cross-organizational data pipelines built on DSs can make model development cumulative and institutionally shareable, reducing the redundant experimentation that currently occurs independently at each site. The architecture is domain-agnostic and transferable to any production setting where heterogeneous data must be turned into reusable knowledge assets.

Future research directions are grounded in the specific limitations encountered during this implementation.

1. Ablation studies: The current benchmark cannot isolate the individual contributions of cross-site data aggregation, parameter transfer from θ_0, and semantic annotation to the observed training-time reduction and accuracy. Controlled ablation experiments—for example, comparing cross-site aggregation without fine-tuning, single-site fine-tuning without a cross-organizational foundation model, and varying the number of fine-tuned layers systematically—would clarify which factors drive the gains and inform the design of future D2K deployments.
2. Realizing the K2D feedback loop: Safe write-back to legacy production equipment, real-time model inference at data collection time, and semantic drift detection as ontologies evolve are the concrete technical hurdles identified.
3. Extending to heterogeneous platforms: The current proof of concept is limited to a single robot family; extending to different robot platforms or other production domains requires additional standardization and may yield different transfer dynamics.
4. Federated and privacy-preserving extensions: For deployments where raw data cannot be shared, federated learning approaches [45,46] combined with cryptographic verification [48] are the natural next step; benchmarking centralized versus federated D2K would clarify the performance/privacy trade-off.

5. Automated semantic alignment: Reducing the manual effort of onboarding new machine types via tools such as OPC UA to RDF mapping [36] is a prerequisite for broader WWL adoption.
6. Governance frameworks: Federally auditable provenance and enforceable data usage agreements are critical non-technical enablers for multi-actor WWL expansion, as the organizational and liability challenges may prove to be as demanding as the technical ones.

By addressing these directions, the envisioned WWL can evolve from the robust academic consortium prototype demonstrated here toward a resilient, privacy-preserving, and self-optimizing global production data ecosystem.

Table 3 summarizes all mathematical symbols used in this manuscript.

Table 3. List of symbols used in this manuscript.

Symbol	Description
Inverse dynamics	
$\mathbf{q}$	Joint configuration vector
$\dot{\mathbf{q}}$	Joint velocity vector
$\ddot{\mathbf{q}}$	Joint acceleration vector
$\boldsymbol{\tau}$	Joint torque vector
$f(\cdot)$	Inverse dynamics mapping, $\boldsymbol{\tau} = f(\mathbf{q}, \dot{\mathbf{q}}, \ddot{\mathbf{q}})$
Data and datasets	
$\mathcal{D}_{\text{site}}$	Trajectory dataset from a specific institute
$\mathcal{D}$	Aggregated cross-organizational dataset, $\mathcal{D} = \mathcal{D}_{\text{LLT}} \cup \mathcal{D}_{\text{ITA}} \cup \mathcal{D}_{\text{WZL}}$
n	Number of training runs per sweep cycle ($n = 10$)
Models and training	
θ_0	Foundation model parameters trained on $\mathcal{D}$
θ_i	Instance-specific model parameters fine-tuned from θ_0
H	Hyperparameter configuration (architecture, learning rate, etc.)
$\mathcal{L}_{\text{CV}}$	Cross-validation loss used for model selection
Evaluation	
MAE	Mean absolute error on predicted joint torques (Nm)

Author Contributions: Conceptualization: all authors; Methodology: L.G., J.-N.S. and P.B.; Software: L.G. and J.-N.S.; Validation: L.G.; Formal analysis: L.G.; Investigation: L.G.; Resources: M.L., D.K., T.K., O.P., C.H. (Christian Hinke), T.G., G.L., M.Z., C.B. and C.H. (Constantin Häfner); Data curation: L.G. and J.-N.S.; Writing—original draft preparation: L.G., M.B., P.B., M.L. and D.K.; Writing—review and editing: L.G., M.B., P.B., J.-N.S., T.K., O.P., C.H. (Christian Hinke), T.G., G.L., M.Z., C.B. and C.H. (Constantin Häfner); Visualization: L.G. and P.B.; Supervision: T.G., G.L., M.Z., C.B. and C.H. (Constantin Häfner); Funding acquisition: T.G., G.L., M.Z., C.B. and C.H. (Constantin Häfner). All authors have read and agreed to the published version of the manuscript.

Funding: Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC-2023 Internet of Production—390621612. Open access funding provided by the Open Access Publishing Fund of RWTH Aachen University.

Data Availability Statement: The data supporting the findings of this study are openly available as part of the replication package at https://s.fhg.de/gorissen-2025a and via the persistent identifiers [59,60] and https://doi.org/10.18154/RWTH-2025-00450, [74]: Trained benchmark models resulting from the four evaluated training configurations.

Acknowledgments: This manuscript was reviewed for spelling, grammar, and clarity using AI-based tools, including ChatGPT (OpenAI, GPT-4o and later versions) and Claude Sonnet 4.6 (Anthropic).

Conflicts of Interest: The authors have no competing interests to declare that are relevant to the content of this article.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
D2K	Data-to-Knowledge
DAG	Directed Acyclic Graph
deep RL	Deep Reinforcement Learning
DS	Digital Shadow
DT	Digital Twin
ERP	Enterprise Resource Planning
FAIR	Findable, Accessible, Interoperable, Reusable
FDO	FAIR Digital Object
FedCSF	Federated graph learning via Constructing and Sharing Feature spaces
FedHA	Federated Heterogeneity-aware Adaptive framework
ILT	Fraunhofer Institute for Laser Technology
IoP	Internet of Production
IoT	Internet of Things
ITA	Institut für Textiltechnik (Institute for Textile Technology)
K2D	Knowledge-to-Data
LLT	Lehrstuhl für Lasertechnik (Chair for Laser Technology)
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MES	Manufacturing Execution System
PINN	Physics-Informed Neural Network
QOMOU	Querying of Ontology Mapping-based OPC UA
RDF	Resource Description Format
SME	Small and Medium-sized Enterprise
WWL	World Wide Lab
WWW	World Wide Web
WZL	Werkzeugmaschinenlabor (Chair for Machine Tools)

References

1. Bruner, J. *Industrial Internet—The Machines Are Talking*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2013.
2. Kagermann, H. Change Through Digitization—Value Creation in the Age of Industry 4.0. In *Management of Permanent Change*; Albach, H., Meffert, H., Pinkwart, A., Reichwald, R., Eds.; Springer Fachmedien Wiesbaden: Wiesbaden, Germany, 2015; pp. 23–45. [CrossRef]
3. Lu, Y. Industry 4.0: A survey on technologies, applications and open research issues. *J. Ind. Inf. Integr.* **2017**, *6*, 1–10. [CrossRef]
4. Zhong, R.Y.; Xu, X.; Klotz, E.; Newman, S.T. Intelligent Manufacturing in the Context of Industry 4.0: A Review. *Engineering* **2017**, *3*, 616–630. [CrossRef]
5. Nargesian, F.; Zhu, E.; Miller, R.J.; Pu, K.Q.; Arocena, P.C. Data lake management: Challenges and opportunities. *Proc. VLDB Endow.* **2019**, *12*, 1986–1989. [CrossRef]
6. Nambiar, A.; Mundra, D. An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management. *Big Data Cogn. Comput.* **2022**, *6*, 132. [CrossRef]
7. Goedegebuure, A.; Kumara, I.; Driessen, S.; Van Den Heuvel, W.J.; Monsieur, G.; Tamburri, D.A.; Nucci, D.D. Data Mesh: A Systematic Gray Literature Review. *ACM Comput. Surv.* **2024**, *57*, 1–36. [CrossRef]
8. Behery, M.; Glawe, F.; Koren, I.; Ziefle, M.; Lakemeyer, G.; Brauner, P. Vision Paper: Leveraging Industrial Big Data—Past, Present, and Future of the World Wide Lab. In Proceedings of the 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 15–18 December 2023; pp. 1308–1313. [CrossRef]
9. Brauner, P.; Dalibor, M.; Jarke, M.; Kunze, I.; Koren, I.; Lakemeyer, G.; Liebenberg, M.; Michael, J.; Pennekamp, J.; Quix, C.; et al. A Computer Science Perspective on Digital Transformation in Production. *ACM Trans. Internet Things* **2022**, *3*, 1–32. [CrossRef]

10. Liebenberg, M.; Jarke, M. Information Systems Engineering with Digital Shadows: Concept and Case Studies: An Exploratory Paper. In Proceedings of the Advanced Information Systems Engineering: 32nd International Conference, CAiSE 2020, Grenoble, France, 8–12 June 2020; pp. 70–84. [CrossRef]
11. Schneider, J.N.; Gorissen, L.; Kaster, T.; Walderich, P.; Hinke, C. LSTM-based Inverse Dynamics Learning for Franka Emika Robot. In Proceedings of the 2024 International Conference on Control, Automation and Diagnosis (ICCAD), Lyon, France, 1–3 July 2024; pp. 1–6. [CrossRef]
12. Inmon, W.H. *Building the Data Warehouse*, 3rd ed.; Wiley Computer Publishing: Hoboken, NJ, USA, 2002.
13. Dixon, J. Pentaho, Hadoop, and Data Lakes. James Dixon's Blog, 2010. Available online: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/ (accessed on 13 May 2026).
14. Harby, A.A.; Zulkernine, F. From Data Warehouse to Lakehouse: A Comparative Review. In Proceedings of the 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 17–20 December 2022; pp. 389–395. [CrossRef]
15. IBM. *What is a Data Fabric?* IBM Website: Armonk, NY, USA, 2024.
16. Schuh, G.; Prote, J.P.; Gützlaff, A.; Thomas, K.; Sauermann, F.; Rodemann, N. Internet of Production: Rethinking production management. In *Proceedings of the Production at the Leading Edge of Technology*; Wulfsberg, J.P., Hintze, W., Behrens, B.A., Eds.; Springer: Berlin/Heidelberg, Germany, 2019; pp. 533–542. [CrossRef]
17. Pallasch, C.; Hoffmann, N.; Storms, S.; Herfs, W. ProducTron: Towards Flexible Distributed and Networked Production. In Proceedings of the 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES), Las Palmas de Gran Canaria, Spain, 21–23 June 2018; pp. 000287–000292. [CrossRef]
18. Gleim, L.; Pennekamp, J.; Liebenberg, M.; Buchsbaum, M.; Niemietz, P.; Knape, S.; Epple, A.; Storms, S.; Trauth, D.; Bergs, T.; et al. FactDAG: Formalizing Data Interoperability in an Internet of Production. *IEEE Internet Things J.* **2020**, *7*, 3243–3253. [CrossRef]
19. Pennekamp, J.; Matzutt, R.; Kanhere, S.S.; Hiller, J.; Wehrle, K. The Road to Accountable and Dependable Manufacturing. *Automation* **2021**, *2*, 202–219. [CrossRef]
20. Auer, M.; Zutin, D.G. A grid of online laboratories based on the iLab shared architecture. In Proceedings of the ASEE Annual Conference and Exposition, San Antonio, TX, USA, 10–13 June 2012. [CrossRef]
21. Salzmann, C.; Gillet, D.; Esquembre, F.; Dormido, S. Web 2.0 open remote and virtual laboratories in engineering education. In *Cyber Behavior: Concepts, Methodologies, Tools, and Applications*; IGI Global Scientific Publishing: Palmdale, PA, USA, 2014. [CrossRef]
22. Titov, I.; Glotov, A.; Mikolnikov, J. Labicom.net—The online laboratories platform demonstration 2014. In Proceedings of the 2014 International Conference on Interactive Collaborative Learning (ICL), Dubai, United Arab Emirates, 3–5 December 2014. [CrossRef]
23. Carnegie Mellon University. *Manufacturing Futures Institute—Building the Factory of the Future*; Carnegie Mellon University: Pittsburgh, PA, USA, 2023.
24. The Smart Manufacturing Institute. *Smart Manufacturing Innovation Platform*; The Smart Manufacturing Institute: Los Angeles, CA, USA, 2023.
25. Gaia-X. Gaia-X: A Federated Data Infrastructure for Europe. Gaia-X Project Website, 2020. Available online: https://www.gaia-x.eu (accessed on 1 October 2024).
26. Catena-X. Catena-X: The Automotive Network. Catena-X, 2021. Available online: https://catena-x.net/ (accessed on 1 October 2024).
27. Plattform Industrie 4.0. Manufacturing-X: Data Ecosystem for Manufacturing. Plattform Industrie 4.0 Website, 2022. Available online: https://www.plattform-i40.de/IP/Navigation/DE/Manufacturing-X/Initiative/initiative-manufacturing-x.html (accessed on 1 October 2024).
28. NVIDIA. Omniverse—Plattform für Open USD. Available online: https://www.nvidia.com/de-de/omniverse/ (accessed on 8 August 2025).
29. Munappy, A.R.; Bosch, J.; Olsson, H.H. On the Trade-off Between Robustness and Complexity in Data Pipelines. In *Proceedings of the Quality of Information and Communications Technology*; Paiva, A.C.R., Cavalli, A.R., Ventura Martins, P., Pérez-Castillo, R., Eds.; Springer: Cham, Switzerland, 2021; pp. 401–415. [CrossRef]
30. Munappy, A.R.; Bosch, J.; Olsson, H.H. Maturity Assessment Model for Industrial Data Pipelines. In Proceedings of the 2023 30th Asia-Pacific Software Engineering Conference (APSEC), Los Alamitos, CA, USA, 4–7 December 2023; pp. 503–513. [CrossRef]
31. Yadranjiaghdam, B.; Pool, N.; Tabrizi, N. A Survey on Real-Time Big Data Analytics: Applications and Tools. In Proceedings of the 2016 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 15–17 December 2016; pp. 404–409. [CrossRef]
32. Tu, D.; He, Y.; Cui, W.; Ge, S.; Zhang, H.; Han, S.; Zhang, D.; Chaudhuri, S. Auto-Validate by-History: Auto-Program Data Quality Constraints to Validate Recurring Data Pipelines. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 6–10 August 2023; pp. 4991–5003. [CrossRef]

33. Song, J.; He, Y. Auto-Validate: Unsupervised Data Validation Using Data-Domain Patterns Inferred from Data Lakes. In Proceedings of the 2021 International Conference on Management of Data, New York, NY, USA, 20–25 June 2021; pp. 1678–1691. [CrossRef]
34. Mesbah, S.; Fragkeskos, K.; Lofi, C.; Bozzon, A.; Houben, G.J. Semantic Annotation of Data Processing Pipelines in Scientific Publications. In *Proceedings of the Semantic Web*; Blomqvist, E., Maynard, D., Gangemi, A., Hoekstra, R., Hitzler, P., Hartig, O., Eds.; Springer: Cham, Switzerland, 2017; pp. 321–336. [CrossRef]
35. Zheng, Z.; Zhou, B.; Zhou, D.; Soylu, A.; Kharlamov, E. ExeKG: Executable Knowledge Graph System for User-friendly Data Analytics. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management, New York, NY, USA, 17–21 October 2022; pp. 5064–5068. [CrossRef]
36. Bi, J.; Wu, R.; Yuan, H.; Wang, Z.; Zhang, J.; Zhou, M. Ontology-Based Semantic Reasoning for Multisource Heterogeneous Industrial Devices Using OPC UA. *IEEE Internet Things J.* **2025**, *12*, 25020–25032. [CrossRef]
37. Bodenbenner, M.; Pennekamp, J.; Montavon, B.; Wehrle, K.; Schmitt, R.H. FAIR Sensor Ecosystem: Long-Term (Re-)Usability of FAIR Sensor Data through Contextualization. In Proceedings of the 2023 IEEE 21st International Conference on Industrial Informatics (INDIN), Helsinki, Finland, 13–16 June 2023; pp. 1–8. [CrossRef]
38. Date, C.J. *Database Design and Relational Theory: Normal Forms and All That Jazz*; Apress: Berkley, CA, USA, 2019. [CrossRef]
39. Behery, M.; Brauner, P.; Kluge-Wilkes, A.; Baier, R.; Mertens, A.; Schmitt, R.H.; Ziefle, M.; Lakemeyer, G. Digital Shadows for Robotic Assembly in the World Wide Lab. *Procedia CIRP* **2023**, *120*, 165–170. [CrossRef]
40. Bauernhansl, T.; Hartleif, S.; Felix, T. The Digital Shadow of Production—A Concept for the Effective and Efficient Information Supply in Dynamic Industrial Environments. *Procedia CIRP* **2018**, *72*, 69–74. [CrossRef]
41. Kulkarni, V.; Reddy, S. Separation of concerns in model-driven development. *IEEE Softw.* **2003**, *20*, 64–69. [CrossRef]
42. Nadareishvili, I.; Mitra, R.; McLarty, M.; Amundsen, M. *Microservice Architecture: Aligning Principles, Practices, and Culture*, 1st ed.; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2016.
43. Heithoff, M.; Hopmann, C.; Köbel, T.; Michael, J.; Rumpe, B.; Sapel, P. Application of digital shadows on different levels in the automation pyramid. *Data Knowl. Eng.* **2025**, *158*, 102442. [CrossRef]
44. Michael, J.; Koren, I.; Dimitriadis, I.; Fulterer, J.; Gannouni, A.; Heithoff, M.; Hermann, A.; Hornberg, K.; Kröger, M.; Sapel, P.; et al. A Digital Shadow Reference Model for Worldwide Production Labs. In *Internet of Production: Fundamentals, Applications and Proceedings*; Brecher, C., Schuh, G., van der Aalst, W., Jarke, M., Piller, F.T., Padberg, M., Eds.; Springer International Publishing: Cham, Switzerland, 2023; pp. 1–28. [CrossRef]
45. Tong, C.; Zhang, L.; Ding, Y.; Yue, D. A Heterogeneity-Aware Adaptive Federated Learning Framework for Short-Term Forecasting in Electric IoT Systems. *IEEE Internet Things J.* **2025**, *12*, 15388–15403. [CrossRef]
46. Chen, J.; Zhuo, S.; He, J.; Qiu, W.; Zhang, Q.; Xiong, Z.; Zheng, Z.; Tang, Y.; Chen, M.; Wang, C.; et al. Federated Graph Learning via Constructing and Sharing Feature Spaces for Cross-Domain IoT. *IEEE Internet Things J.* **2025**, *12*, 26200–26214. [CrossRef]
47. Chahoud, M.; Sami, H.; Mourad, A.; Otrok, H.; Bentahar, J.; Guizani, M. On-Demand Model and Client Deployment in Federated Learning With Deep Reinforcement Learning. *IEEE Internet Things J.* **2025**, *12*, 26685–26698. [CrossRef]
48. Wang, X.; Chen, T.; Dai, H.N.; Long, P.; Yang, H.; Xiong, Z.; Susilo, W. A Privacy-Enhanced Method for Privacy-Preserving and Verifiable Federated Learning. *IEEE Internet Things J.* **2025**, *12*, 26768–26781. [CrossRef]
49. Alavi, M.; Leidner, D.E. Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. *MIS Q.* **2001**, *25*, 107–136. [CrossRef]
50. National Institute of Standards and Technology. *NIST Big Data Interoperability Framework (NBDIF) Version 3.0*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2019.
51. Russell, S.J.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 3rd ed.; Pearson Education: London, UK, 2016.
52. Coscine. Coscine. Coscine Project Website, 2016. Available online: https://about.coscine.de/en/ (accessed on 17 March 2024).
53. Bjorck, J.; Castañeda, F.; Cherniadev, N.; Da, X.; Ding, R.; Fan, L.J.; Fang, Y.; Fox, D.; Hu, F.; Huang, S.; et al. GR00T N1: An Open Foundation Model for Generalist Humanoid Robots. *arXiv* **2025**. [CrossRef]
54. Ball, P.J.; Bauer, J.; Belletti, F.; Brownfield, B.; Ephrat, A.; Fruchter, S.; Gupta, A.; Holsheimer, K.; Holynski, A.; Hron, J.; et al. Genie 3: A New Frontier for World Models, 2025. Available online: https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/ (accessed on 8 August 2025).
55. Gorißen, L.; Schneider, J.N.; Kaster, T.; Hinke, C.; Häfner, C. Towards the Application of Low-Cost Collaborative Robots in Laser Materials Processing. *J. Laser Micro/Nanoeng.* **2026**, *21*, 91–102. [CrossRef]
56. Dammers, H.; Lennartz, M.; Liebe, P.; Gries, T. AI-Driven Robotic-Tool Selection for Draping Composite Preforms Based on a Geometric Surface Segmentation Approach. In *Proceedings of the SAMPE 2024*; NA SAMPE: Diamond Bar, CA, USA, 2024. [CrossRef]
57. Arents, J.; Abolins, V.; Judvaitis, J.; Vismanis, O.; Oraby, A.; Ozols, K. Human—Robot Collaboration Trends and Safety Aspects: A Systematic Review. *J. Sens. Actuator Netw.* **2021**, *10*, 48. [CrossRef]

58. Siciliano, B.; Khatib, O. (Eds.) *Springer Handbook of Robotics*; Springer International Publishing: Berlin/Heidelberg, Germany, 2016. [CrossRef]
59. Gorißen, L.M.; Schneider, J.N.; Behery, M.; Brauner, P.; Lennartz, M.; Kötter, E.D.; Kaster, T.; Petrovic, O.; Hinke, C.R.; Gries, T.; et al. *Demonstrating Data-to-Knowledge Pipelines for Connecting Production Sites in the World Wide Lab: Source Code*; RWTH Aachen University: Aachen, Germany, 2025. [CrossRef]
60. Gorißen, L.M.; Schneider, J.N.; Behery, M.; Brauner, P.; Lennartz, M.; Kötter, E.D.; Kaster, T.; Petrovic, O.; Hinke, C.R.; Gries, T.; et al. *Demonstrating Data-to-Knowledge Pipelines for Connecting Production Sites in the World Wide Lab: Trajectory Data*; RWTH Aachen University: Aachen, Germany, 2025. [CrossRef]
61. Plattform Industrie 4.0. *Reference Architecture Model Industrie 4.0 (RAMI 4.0)*; Plattform Industrie 4.0: Berlin, Germany, 2015.
62. Otto, B.; Jarke, M. Designing a Multi-sided Data Platform: Findings from the International Data Spaces Case. *Electron. Mark.* **2019**, *29*, 561–580. [CrossRef]
63. Behery, M.; Brauner, P.; Zhou, H.A.; Uysal, M.S.; Samsonov, V.; Bellgardt, M.; Brillowski, F.; Brockhoff, T.; Farhang Ghahfarokhi, A.; Gleim, L.; et al. Actionable Artificial Intelligence for the Future of Production. In *Internet of Production*; Springer: Cham, Switzerland, 2023; pp. 1–46. [CrossRef]
64. Endsley, M.R. From Here to Autonomy: Lessons Learned from Human–Automation Research. *Hum. Factors* **2017**, *59*, 5–27. [CrossRef]
65. Bernhard, S.; Pütz, S.; Röhl, C.; Baier, R.; Brauner, P.; Christou, E.; Dammers, H.; Flaig, R.; Gorißen, L.M.; Heilinger, J.C.; et al. Sustainability in the Internet of Production: Interdisciplinary Opportunities and Challenges. In Proceedings of the 2023 IEEE International Symposium on Technology and Society (ISTAS), Cape Town, South Africa, 15–17 November 2023; pp. 1–8. [CrossRef]
66. Sweeney, L. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* **2002**, *10*, 557–570. [CrossRef]
67. Li, N.; Li, T.; Venkatasubramanian, S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 15–20 April 2007; pp. 106–115. [CrossRef]
68. Machanavajjhala, A.; Gehrke, J.; Kifer, D.; Venkitasubramaniam, M. L-diversity: Privacy beyond k-anonymity. In Proceedings of the 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, 3–7 April 2006; p. 24. [CrossRef]
69. Zheng, X.; Cai, Z. Privacy-Preserved Data Sharing Towards Multiple Parties in Industrial IoTs. *IEEE J. Sel. Areas Commun.* **2020**, *38*, 968–979. [CrossRef]
70. Abraham, R.; Schneider, J.; vom Brocke, J. Data governance: A conceptual framework, structured review, and research agenda. *Int. J. Inf. Manag.* **2019**, *49*, 424–438. [CrossRef]
71. Hummel, P.; Braun, M.; Tretter, M.; Dabrock, P. Data sovereignty: A review. *Big Data Soc.* **2021**, *8*, 2053951720982012. [CrossRef]
72. Cuñat, S.; Julian, M.; Belsa, A.; Valero, C.I.; Esteve, M.; Palau, C.E. Secure, Trusted, Privacy-Protected Data Exchange in an Edge-Cloud Continuum Environment. In *Internet of Things*; Springer: Cham, Switzerland, 2024; pp. 201–231. [CrossRef]
73. Shoomal, A.; Jahanbakht, M.; Componation, P.J. An analytical framework for evaluating blockchain and IoT use cases in sustainable supply chains. *Supply Chain. Anal.* **2026**, *13*, 100198. [CrossRef]
74. Gorißen, L.M.; Schneider, J.N.; Behery, M.; Brauner, P.; Lennartz, M.; Kötter, E.D.; Kaster, T.; Petrovic, O.; Hinke, C.R.; Gries, T.; et al. *Demonstrating Data-to-Knowledge Pipelines for Connecting Production Sites in the World Wide Lab: Benchmark Models*; RWTH Aachen University: Aachen, Germany, 2025. [CrossRef]