White Paper: KnoDL Technology for Deterministic Data Matching and Unification

Version 1.0

Date: August 28, 2025

Abstract

In the modern digital economy, data serves as a strategic asset, yet its value is directly dependent on its quality. Organizations across various sectors—from government administration and logistics to telecommunications and retail—face a systemic problem of data fragmentation. The presence of duplicates, format inconsistencies, errors, and disparate reference systems leads to direct financial losses, operational risks, and erroneous strategic decisions.

KnoDL (Knowledge Definition Language) technology represents an engineering approach to solving this problem. Unlike probabilistic methods based on machine learning (ML/AI), KnoDL uses formal, deterministic algorithms for matching, cleansing, and unifying heterogeneous data. This ensures complete transparency, 100% reproducibility, and legal provability of results, which is a critically important requirement for regulated industries and high-risk systems.

This document presents a technical and business overview of KnoDL technology. It examines architectural principles, key application scenarios, and measurable implementation effects confirmed by pilot project results.

---

1. Introduction: The Crisis of "Dirty" Data

The problem of poor-quality data, often called "dirty" data, is systemic in nature and manifests at all organizational levels.

1.1. Sources of the Problem:

- **Multiple source systems:** Data about the same entities (customers, products, addresses) is stored in different systems (CRM, ERP, billing, WMS), each with its own structure and standards.

- **Disparate reference books and classifiers:** The absence of unified normative reference information (NRI) leads to the same entity being coded and named differently.
- **Human factor:** Manual input errors, use of abbreviations, typos, and non-compliance with data entry regulations.
- **Legacy data:** Outdated records and data migration from old systems without proper cleansing.

1.2. Measurable Business Consequences:

- **Direct financial losses:** According to the presented data, up to 8 out of 10 large companies suffer losses due to "dirty" data. In retail, this leads to inventory management errors; in telecom, to revenue loss and regulatory fines.
- **Reduced analytics accuracy:** Losses in analytical model accuracy can reach 25%, leading to incorrect forecasts and erroneous strategic decisions.
- **Operational inefficiency:** Manual data reconciliation and cleansing is a labor-intensive, expensive, and slow process. It diverts up to 80% of qualified specialists' time from performing core tasks.
- **Compliance and security risks:** Errors in customer verification procedures (KYC) can lead to fines. In control systems, for example, unmanned vehicles, data errors create direct security threats.

---

2. KnoDL Technological Foundations

KnoDL offers a fundamentally different approach to working with data, based on formal methods rather than statistical learning.

2.1. Architecture and Operating Principle:

KnoDL technology is based on computing similarity metrics between two textual or structured (JSON) objects.

1. **Vectorization (creating "convolutions"):** In the first stage, a compact digital "fingerprint" or "convolution" is created for each record (string, document). This process is deterministic: identical records always receive identical convolutions. The convolution size is significantly smaller than the original record.

2. **Comparison by "convolutions":** In the second stage, similarity comparison is performed not between the original data, but between their convolutions. This allows for orders of magnitude increase in processing speed for large data arrays.
3. **Similarity metric calculation:** The algorithm computes a continuous metric ranging from 0 (complete difference) to 1 (identity). It considers substring occurrences, word order, and structure, making it resistant to typos, permutations, and word form changes.

2.2. Key Difference from ML/AI Approaches:

Characteristic	KnoDL (Formal Approach)	ML/AI (Probabilistic Approach)
Transparency	**Complete.** The logic of each match can be explained and verified.	**Low ("black box").** Results are difficult to interpret, especially in case of errors.
Training	**Not required.** Works "out of the box," rules can be configured.	**Required.** Large, labeled, and quality datasets are needed for training.
Resource Requirements	**Low.** Works on standard CPU, requires ⩽4 GB RAM.	**High.** Often requires GPU and 8–16 GB VRAM.
Reproducibility	**100% determinism.** Guarantees identical results with identical input data.	**Probabilistic.** Results may vary, subject to data "noise."
Implementation Speed	**High.** Does not require lengthy training and data labeling phases.	**Low.** The process of data collection, training, and model tuning takes considerable time.

2.3. API and CLI Integration:

KnoDL provides integration interfaces: * **Command Line Interface (CLI):** A set of commands for data import (`kdl lines import`), matching (`kdl fuzzy match`), and database cleanup (`kdl db clean`). * **HTTP API:** A wrapper over CLI, allowing the same commands to be executed via POST requests, simplifying integration with existing IT systems.

---

3. Practical Application Scenarios and Case Studies

KnoDL technology has been tested on real industrial data across various industries.

3.1. GovTech — Government Registry Management

- **Problem:** Different ministries maintain their own registries (population, licenses, benefits) with duplicates, errors, and mismatches. This creates barriers to public service delivery and increases corruption risks.
- **KnoDL Solution:** Automatic consolidation of heterogeneous registries, duplicate elimination, and creation of a single source of truth.
- **Effect:**
    - Cost reduction of 30–40%.
    - Process acceleration (application/approval) from weeks to hours.
    - Reduced corruption risks and increased citizen trust.
    - Automatic matching accuracy reaches 98%.

3.2. Telco — Customer Base Management

- **Problem:** Millions of customers, duplicates in CRM and billing, errors in KYC procedures, revenue loss due to inaccurate data.
- **KnoDL Solution:** Customer data cleansing and unification, intelligent search across entire customer history (multi-ID), automatic deduplication.
- **Effect:**
    - ARPU growth of 10–15% through churn reduction and cross-selling.
    - Reduced regulatory fines.
    - Creation of a "clean" customer base to support new services (5G/IoT).

3.3. Logistics — Normative Reference Information Harmonization

- **Problem:** In cross-border transportation, participants (road, rail, maritime transport) use different reference books, encodings, and document formats, slowing processes and leading to errors.
- **KnoDL Solution:** Automatically matches commodity reference books, verifies cargo compliance with different jurisdictions' regulations, and creates a "single window" for normative information. Integrates with international standards such as GS1.

- **Effect:**
  - Document preparation time reduction of 30–70%.
  - Elimination of logical and structural conflicts in data exchange.
  - Increased speed and automation of decision-making.

3.4. Retail — Inventory Management

- **Problem:** Product data inconsistencies (SKU, GTIN, packaging) between WMS, SAP, and Excel systems lead to inventory errors and suboptimal stock management.
- **KnoDL Solution:** Performs deterministic matching for SKU unification, finds duplicates, and eliminates inconsistencies.
- **Effect (based on pilot project results):**
  - Reduced inventory discrepancies.
  - Lower manual data processing costs.
  - Complete traceability of corrections and data sources.

---

4. Conclusion

KnoDL is not a universal replacement for artificial intelligence technologies, but a specialized engineering tool that excels in tasks requiring **accuracy, transparency, and provability**. Its application is most effective when working with structured, semi-structured, and normative data.

KnoDL implementation enables organizations to: 1. **Transform "dirty" data into a reliable asset** by creating a unified, consistent source of truth. 2. **Significantly reduce operational costs** by automating routine reconciliation and data cleansing processes. 3. **Reduce risks** associated with poor-quality data in regulated and mission-critical processes. 4. **Establish a solid technological foundation** for further digital transformation, predictive analytics, and automation.

Thus, KnoDL is a key element for building a modern, data-driven organization.