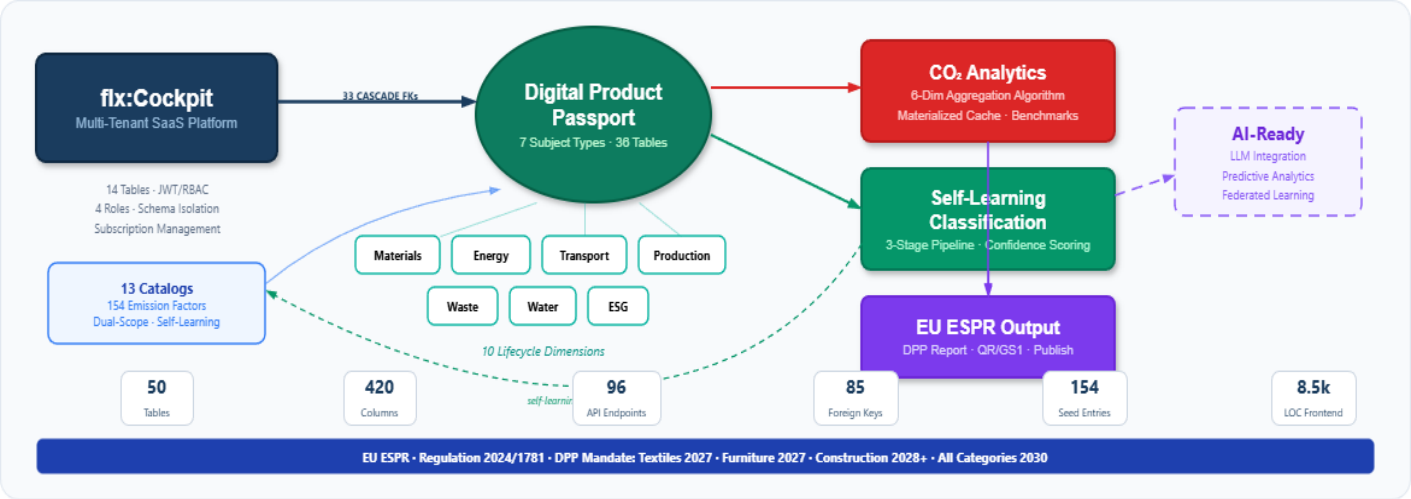
Architecture of a Multi-Tenant Digital Product Passport Platform

A Data-Centric Approach to EU ESPR Compliance with Integrated Analytics and AI-Ready Data Pipelines



Luisa Wehle

B.Sc. Student, Business Informatics, Hochschule der Medien, Stuttgart, Germany Co-Founder, flx-business GbR, Böblingen, Germany

Hans-Dieter Wehle

Dipl. Business Informatics (Wirtschaftsinformatik) Co-Founder & Managing Director, flx-business GbR, Böblingen, Germany

March 2026

Version 1.0

Keywords: *Digital Product Passport, EU ESPR, Multi-Tenant SaaS, Life Cycle Assessment, CO_2 Analytics, Self-Learning Classification, ESG Scoring, Data Architecture, Regulatory Compliance, AI-Ready Data Pipelines*

Abstract

The European Union's Ecodesign for Sustainable Products Regulation (ESPR, Regulation 2024/1781) mandates Digital Product Passports (DPPs) for an expanding set of product categories starting 2027.[1] The platform described in this paper originated from concrete customer requirements in the German construction and manufacturing industry, where companies faced the challenge of systematically capturing lifecycle data for regulatory compliance without adequate tooling. Rather than a theoretical framework, the architecture presented here was iteratively developed and validated through real-world deployment and client demonstrations. This paper presents the architecture of FLX:DPPplus, a production-grade, multi-tenant SaaS platform that implements a comprehensive data model for DPP generation across seven product types: construction, industrial products, batch goods, components, furniture, and textiles. We describe a 50-table relational schema spanning two PostgreSQL schemas with cross-schema tenant isolation, a multi-dimensional Life Cycle Assessment (LCA) engine covering 10 environmental dimensions with formal CO_2 computation algorithms, and a self-learning catalog classification system that bridges rule-based material matching with AI-ready data pipelines. The platform processes PDF bill-of-materials imports through a three-stage matching pipeline (keyword → catalog → fallback) with explicit confidence scoring, implements materialized CO_2 aggregates across materials, energy, transport, production, waste, and water dimensions, and supports EU ESPR-compliant publishing with GS1 Digital Link integration.[2] We formalize the CO_2 computation as a six-dimensional aggregation algorithm and demonstrate how the data architecture enables predictive analytics, anomaly detection, and AI-augmented material classification. With ~420 columns, ~85 foreign key relationships, and ~96 REST API endpoints, the system represents a significant contribution to the emerging field of regulatory data engineering for sustainability.

Kurzfassung

Die europäische Ökodesign-Verordnung für nachhaltige Produkte (ESPR, Verordnung 2024/1781) schreibt Digitale Produktpässe (DPP) für eine wachsende Zahl von Produktkategorien ab 2027 vor. Die in diesem Paper beschriebene Plattform entstand aus konkreten Kundenanforderungen der deutschen Bau- und Fertigungsindustrie, in der Unternehmen vor der Herausforderung standen, Lebenszyklusdaten systematisch für die regulatorische Compliance zu erfassen, ohne dafür geeignete Werkzeuge zu haben. Anstelle eines theoretischen Frameworks wurde die hier vorgestellte Architektur iterativ entwickelt und durch realen Produktiveinsatz sowie Kundenpräsentationen validiert. Dieses Paper beschreibt die Architektur von FLX:DPPplus, einer mandantenfähigen SaaS-Plattform, die ein umfassendes Datenmodell zur DPP-Erstellung über sieben Produkttypen implementiert: Bauprojekte, Industrieprodukte, Chargen, Komponenten, Möbel und Textilien. Wir beschreiben ein relationales Schema mit 50 Tabellen über zwei PostgreSQL-Schemata mit schema-übergreifender Mandantenisolation, eine mehrdimensionale Ökobilanz-Engine (LCA) über 10 Umweltdimensionen mit formalisierten CO_2-Berechnungsalgorithmen sowie ein selbstlernendes Katalog-Klassifikationssystem, das regelbasiertes Material-Matching mit KI-fähigen Datenpipelines verbindet. Die Plattform verarbeitet PDF-Stücklisten-Importe durch eine dreistufige Matching-Pipeline

[1] European Commission, Regulation (EU) 2024/1781 on Ecodesign for Sustainable Products (ESPR), Official Journal of the European Union, 2024.

[2] GS1, Digital Link Standard, https://www.gs1.org/standards/gs1-digital-link, accessed Feb 2026.

(Schlüsselwort → Katalog → Fallback) mit expliziter Konfidenz-Bewertung, implementiert materialisierte CO_2-Aggregate über die Dimensionen Material, Energie, Transport, Produktion, Abfall und Wasser und unterstützt EU-ESPR-konforme Veröffentlichung mit GS1-Digital-Link-Integration. Wir formalisieren die CO_2-Berechnung als sechsdimensionalen Aggregationsalgorithmus und zeigen, wie die Datenarchitektur prädiktive Analytik, Anomalieerkennung und KI-gestützte Materialklassifikation ermöglicht.

1. Introduction

1.1 Regulatory Context and Motivation

The European Green Deal and its implementing legislation represent a paradigm shift in how products are documented throughout their lifecycle.[3] The ESPR (Regulation 2024/1781) requires manufacturers to create Digital Product Passports containing comprehensive environmental, social, and governance (ESG) data for each product placed on the EU market. Beginning with batteries and textiles in 2027, the regulation will progressively expand to cover construction products, furniture, electronics, and virtually all product categories by 2030.

The work presented in this paper is grounded in practice rather than theory. FLX:DPPplus was developed in direct response to customer inquiries from construction companies and industrial manufacturers who recognized the approaching ESPR deadlines but lacked the data infrastructure to comply. These organizations faced a common set of challenges: fragmented material data spread across PDF bills of materials, no systematic CO_2 accounting, and no connection between lifecycle data and regulatory reporting. The architecture described in the following sections represents the engineering response to these real-world requirements each design decision, from the self-learning catalog system to the materialized CO_2 aggregates, traces back to a specific operational need encountered during customer engagements. The system has been validated through production deployment and live demonstrations with prospective clients, confirming both technical feasibility and practical usability.

From a data science perspective, this regulation creates an unprecedented challenge: organizations must capture, compute, and publish structured environmental data across heterogeneous product types, integrate supply chain information from multiple tiers, and maintain audit trails that satisfy regulatory authorities. The data volumes are substantial a single construction project may generate hundreds of material entries, dozens of transport records, and multi-dimensional energy consumption profiles, all requiring accurate CO_2 quantification.

1.2 Research Questions and Contributions

This paper addresses three research questions at the intersection of data architecture and sustainability analytics:

- **RQ1:** How can a relational data model support polymorphic product types while maintaining computational efficiency for CO_2 aggregation across 10 lifecycle dimensions?
- **RQ2:** What data pipeline architecture enables self-learning material classification that improves with usage while maintaining full explainability, a critical requirement in regulated environments?

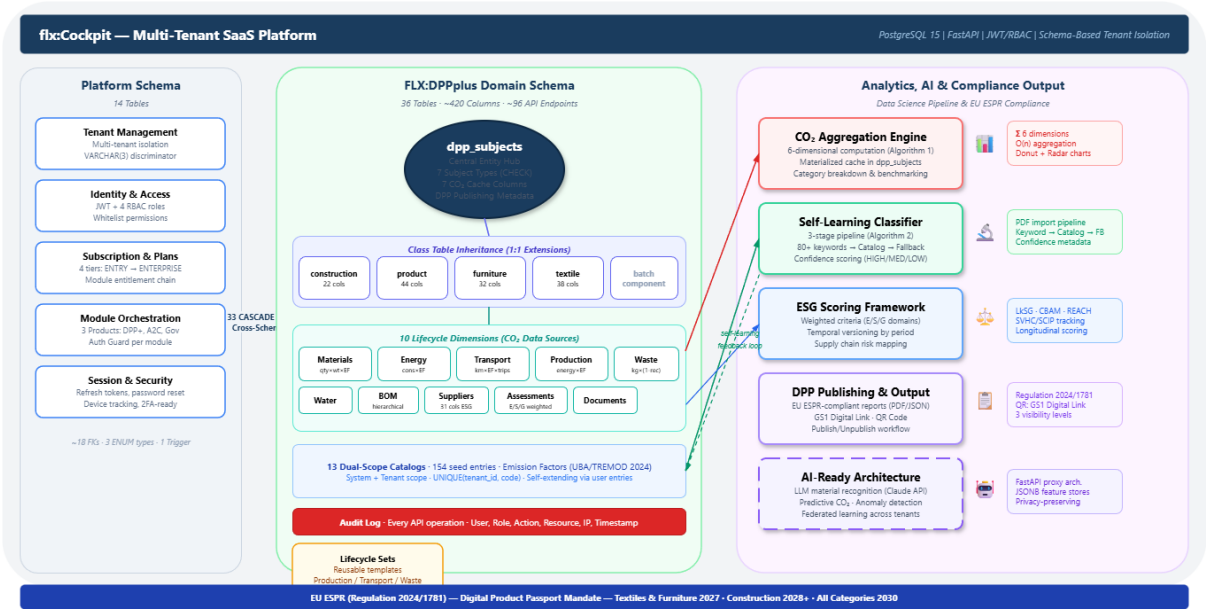
[3] Salhofer, S. et al., "Digital Product Passports as enablers for circular economy," Resources, Conservation and Recycling, vol. 192, 2023.

- **RQ3:** How can the data architecture be designed as AI-ready from inception, enabling future machine learning integration without schema migration?

Our contributions include: (1) a formalized 50-table data model implementing Class Table Inheritance for polymorphic DPPs; (2) a six-dimensional CO_2 aggregation algorithm with materialized caching; (3) a three-stage self-learning classification pipeline with confidence scoring; and (4) an analysis of AI-ready architectural patterns for regulatory data systems.

1.3 System Overview

Figure 1 presents the complete system architecture, illustrating the three-layer design: the platform governance layer (left), the DPP domain model with its lifecycle dimensions and catalog system (center), and the analytics, compliance output, and AI-readiness layer (right). The following sections examine each layer in detail.



FLX:DPPplus is deployed as a module within flx:Cockpit, a multi-tenant SaaS platform built on FastAPI (Python 3.11), PostgreSQL 15, and a CDN-based JavaScript frontend (~8,460 lines). The platform currently manages ~96 API endpoints with JWT-based authentication, role-based access control (RBAC) with four tenant roles, and comprehensive audit logging of all operations.

Table 1: System Metrics at a Glance

Metric	Value	Scope
Database Tables	50	14 platform + 36 DPP
Total Columns	~420	Across both schemas
Foreign Keys	~85	Cross-schema referential integrity
Database Indexes	~65	Including partial unique indexes
API Endpoints	~96	RESTful with audit logging

Catalog Entries (Seed)	154	13 catalogs, system-level
Subject Types	7	Polymorphic with CHECK constraint
Lifecycle Dimensions	10	Materials through Certificates
Frontend Code	~8,460 LOC	8 JavaScript modules

2. Related Work

Research on Digital Product Passports has gained significant momentum since the EU's regulatory announcements.[4] Walden et al. (2023) provide a systematic review of DPP approaches, identifying data interoperability and lifecycle tracking as the primary technical challenges. Berger et al. (2023) propose a generic data architecture for DPPs in circular economy contexts[5], but their model is limited to single-product scenarios without multi-tenant isolation or cross-domain applicability.

In the domain of Life Cycle Assessment (LCA) software, established tools such as GaBi, SimaPro, and openLCA provide comprehensive environmental impact calculation.[6] However, these tools are designed as standalone desktop applications, lack multi-tenant SaaS capabilities, and do not address the specific data requirements of EU ESPR compliance. Furthermore, their material classification relies on expert-curated databases (e.g., ecoinvent) without self-learning capabilities.[7]

Our work differs in three key aspects: (1) we implement a multi-tenant data architecture with cross-schema isolation suitable for B2B SaaS deployment; (2) we provide a self-learning material classification pipeline that adapts to tenant-specific nomenclature without model retraining; and (3) we formalize the CO_2 computation model as a reproducible algorithm anchored to official emission factor databases (UBA/TREMOD 2024).[8]

3. Data Model Architecture

The following two diagrams provide a visual overview of the complete data model across both PostgreSQL schemas. Figure 2 shows the platform schema responsible for multi-tenant governance, authentication, and module entitlement. Figure 3 depicts the dpp_plus schema containing all domain-specific DPP tables, organized into six functional groups: catalogs, core entities, detail extensions, lifecycle data, output/compliance, and audit logging.

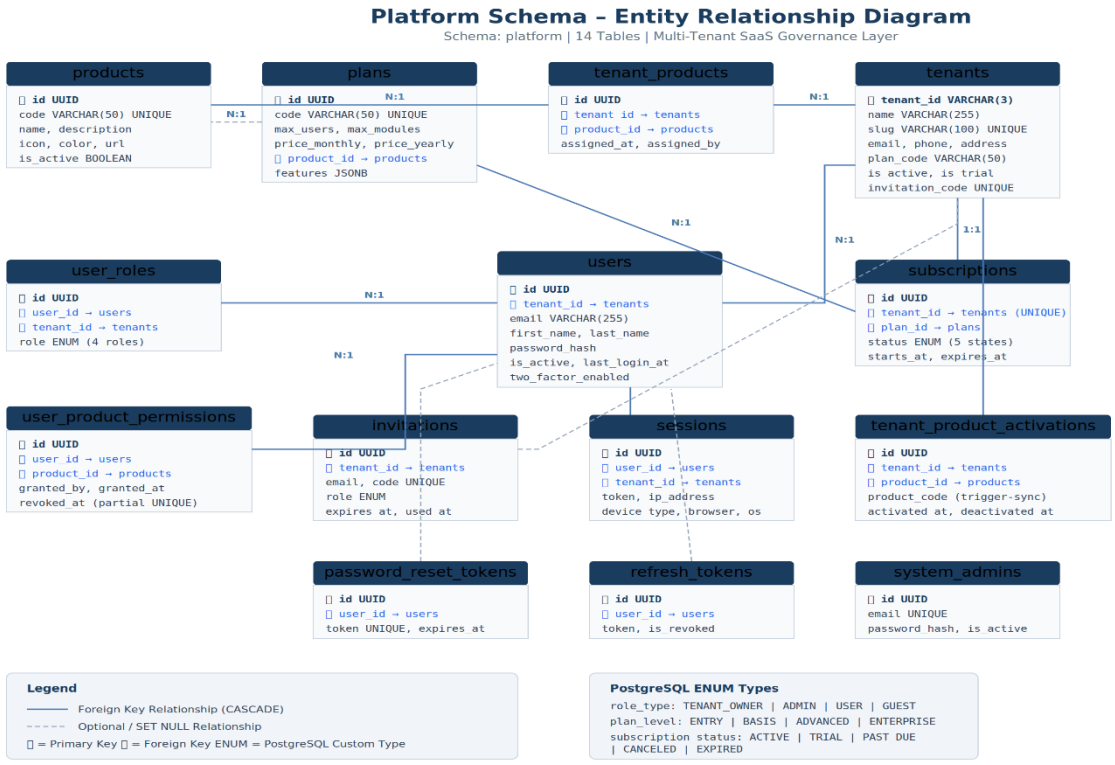
[4]Walden, J. et al., "Digital Product Passport: A systematic review," Journal of Cleaner Production, vol. 403, 2023.

[5]Berger, K. et al., "Data architecture for digital product passports in circular economy," Procedia CIRP, vol. 116, pp. 122–127, 2023.

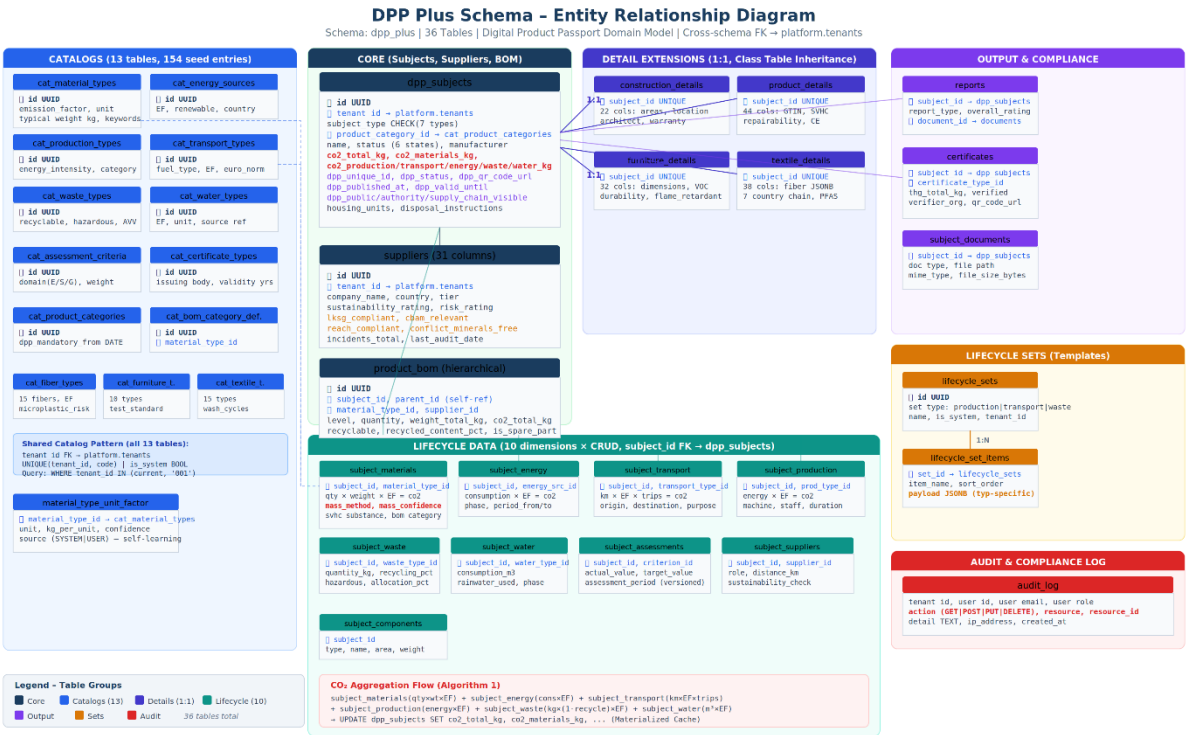
[6]ISO 14040:2006, Environmental management – Life cycle assessment – Principles and framework.

[7]Bezama, A., "Life cycle thinking approaches for sustainable material management," Journal of Industrial Ecology, vol. 26, pp. 1405–1420, 2022.

[8]Umweltbundesamt (UBA), TREMOD Transport Emission Model, Version 6.4, 2024.



The platform schema establishes tenant isolation through the central tenants table, which serves as the foreign key target for all 36 tables in the DPP domain schema. The subscription and permission model implements a three-tier entitlement chain: plans define capabilities, tenant_products assign modules, and user_product_permissions control individual access.



As illustrated in Figure 3, the dpp_subjects table acts as the central hub connecting all six table groups. The CO_2 aggregation flow (highlighted at the bottom) traces how lifecycle data from six dimensions is computed and cached as materialized aggregates directly in the subject record. The following subsections describe each architectural layer in detail.

3.1 Dual-Schema Design with Cross-Schema Isolation

The database architecture follows a strict ***schema-per-concern*** pattern. The **platform** schema (14 tables) manages tenant lifecycle, authentication, subscription, and module entitlement. The **dpp_plus** schema (36 tables) contains all domain-specific DPP data. Every table in dpp_plus references **platform.tenants(tenant_id)** via foreign key with **ON DELETE CASCADE**, ensuring complete data removal upon tenant deprovisioning. This creates 33 cross-schema CASCADE relationships – a deliberate design decision that trades referential complexity for operational safety in a regulated environment.

The tenant identifier uses **VARCHAR(3)** (e.g., "001", "002") rather than UUID. While UUIDs offer universality, the short-string approach provides three advantages in a regulatory context: (1) human readability in audit logs and compliance reports; (2) efficient composite indexing (every DPP table uses *(tenant_id, ...)* indexes); and (3) natural partition boundaries for future horizontal scaling.

Table 2: Schema Distribution and Cross-References

Schema	Tables	FKs	Indexes	Purpose
Platform	14	~18	~22	IAM, Subscriptions, Modules
dpp_plus	36	~67	~43	DPP Domain + LCA + Catalogs

3.2 Polymorphic Subject Model: Class Table Inheritance

The central entity **dpp_subjects** implements a ***Class Table Inheritance*** pattern (Fowler, 2002) to support seven product types within a single table while maintaining type-specific extensions:

Table 3: Subject Type Hierarchy with Detail Extensions

Subject Type	Extension Table	Columns	EU ESPR Target	Cardinality
CONSTRUCTION	construction_details	22	2028+	1:1
PRODUCT	product_details	44	2027+	1:1
FURNITURE	furniture_details	32	2027	1:1
TEXTILE	textile_details	38	2027	1:1
BATCH / COMPONENT	– (base table only)	–	varies	–

The 1:1 relationship is enforced through **UNIQUE(subject_id)** constraints on each detail table, preventing accidental duplication. This design allows type-specific fields (e.g., *fiber_composition JSONB* for textiles, *voc_emission_class* for furniture) without NULL-heavy wide tables, while the base dpp_subjects table carries all 7 cached CO_2 values and the DPP publication metadata shared across all types.

3.3 Dual-Scope Catalog System

The 13 catalog tables implement a ***dual-scope pattern***: system-level seed entries (*is_system=true, tenant_id='001'*) provide a curated baseline of 154 entries with validated emission factors, while tenant-specific entries (*is_system=false*) extend the catalog without affecting other tenants. The composite unique index **(tenant_id, code)** ensures namespace isolation. Catalog queries use *tenant_id IN (current_tenant, '001')* to merge both scopes – a pattern critical for the self-learning classification described in Section 5.

Table 4: Catalog Classification and Analytics Relevance

Catalog	Seeds	Group	Analytics Application
cat_material_types	23	Domain	CO_2 emission factors, keyword matching, weight estimation
cat_energy_sources	12	Domain	Renewable vs. fossil classification, country-specific EF
cat_transport_types	14	Domain	Vehicle×Fuel EF matrix (UBA/TREMOD), Euro norm
cat_production_types	15	Domain	Energy intensity classification
cat_waste_types	14	Domain	Recycling rates, AVV codes, hazardous classification
cat_fiber_types	15	Domain	Natural/synthetic, microplastic risk, water usage
cat_assessment_criteria	16	Master	ESG scoring: weighted criteria, E/S/G domain mapping

3.4 Analytical Data Architecture: From Transactional Store to Structured Data Lake

While the schema described above serves as a transactional system for DPP data entry, its layered structure simultaneously forms a **structured Data Lake** suitable for advanced analytics and machine learning workflows. From a Data Science perspective, the architecture naturally decomposes into five analytical zones that mirror established Data Lake design patterns (Inmon, 2016; Gorelik, 2019):

The **Raw Zone** captures unprocessed input data: PDF bill-of-materials imports arrive as unstructured text extracted via pdfplumber, producing raw material designations before any classification. Entries with confidence = NONE and match_source = UNMATCHED represent the rawest form of data, preserving the original semantics for future reprocessing as classification capabilities improve. The mass_explain text field retains human-readable provenance for each weight determination, forming a natural language corpus for future NLP applications.

The **Curated Zone** encompasses the 13 dual-scope catalog tables with 154 validated seed entries. These catalogs serve as reference data with validated emission factors sourced from UBA/TREMOD 2024, providing the ground truth for all CO_2 calculations. The matching_keywords field in cat_material_types and the material_type_unit_factor table with its source (SYSTEM/USER) and confidence attribution represent curated, quality-controlled reference datasets – the backbone of any supervised learning pipeline.

The **Structured Zone** comprises the 10 lifecycle tables (subject_materials, subject_energy, subject_transport, subject_production, subject_waste, subject_water, subject_assessments, product_bom, subject_suppliers, subject_logistics), the four detail extension tables, and the suppliers master data. Each record carries a foreign key to both the subject and the relevant catalog entry, creating a fully normalized, join-ready analytical dataset. The hierarchical product_bom with parent_id self-reference and level column enables recursive queries for multi-level material composition analysis – a prerequisite for accurate Scope 3 emission attribution.

The **Aggregated Zone** is implemented through the seven materialized CO_2 columns in dpp_subjects (co2_total_kg plus six dimensional breakdowns), the weighted ESG scores from subject_assessments, and the supplier compliance aggregates (sustainability_rating, incidents_total). These pre-computed features represent the serving layer for dashboards and reports, but equally function as labeled training data for predictive models: each subject's CO_2 vector constitutes a multi-dimensional outcome variable linked to its complete lifecycle input features.

The **Event Zone** is formed by the audit_log table, which records every API operation with user_id, action (CRUD types), resource, detail JSONB, ip_address, and timestamp. This event stream enables temporal analytics: user behavior modeling, data entry pattern recognition, anomaly detection (e.g., bulk deletions before audit deadlines), and process mining across the DPP lifecycle. The granularity of one event per API call creates a high-resolution behavioral dataset rarely available in compliance systems.

A distinctive feature of this Data Lake architecture is the **schema-on-read capability** provided by JSONB columns throughout the schema: fiber_composition in textile_details, svhc_details in subject_materials, features in product_details, and certifications store semi-structured data that can be queried with PostgreSQL's native JSON operators. This hybrid approach – relational structure for known dimensions, JSONB for evolving attributes – mirrors the schema-on-read philosophy of traditional Data Lakes while retaining the ACID guarantees and referential integrity of a relational database. For a Data Scientist, this means feature extraction can operate on both structured columns (standard SQL) and semi-structured fields (JSON path queries) within a single analytical query, eliminating the ETL overhead typically associated with moving data from operational stores to analytical environments.

The multi-tenant isolation via tenant_id adds a further analytical dimension: **federated analytics** across tenants become possible while preserving data sovereignty. Cross-tenant queries (with appropriate anonymization) can establish industry benchmarks, identify best practices in CO_2 reduction, and train shared models through federated learning – all without exposing individual tenant data. The confidence metadata (HIGH/MED/LOW/NONE) on every classified record provides built-in data quality scoring, enabling a Data Scientist to filter training sets by confidence level or to implement active learning strategies that prioritize low-confidence items for human review and model retraining.

4. CO_2 Computation Model

4.1 Six-Dimensional Aggregation Algorithm

The CO_2 computation follows ISO 14040/14044 principles[9] adapted for real-time web-based calculation. We formalize the aggregation as Algorithm 1, executed via the **POST /subjects/{id}/recalc-co2** endpoint:

Algorithm 1: Multi-Dimensional CO_2 Aggregation (recalc-co2)

Input: subject_id, tenant_id
Output: co2_vector = (co2_mat, co2_prod, co2_trans, co2_energy, co2_waste, co2_water)

1: co2_mat ← ∑ (quantity_i × weight_per_unit_i × EF_material_i)
where EF_material_i ∈ cat_material_types ∀ i ∈ subject_materials

2: co2_prod ← ∑ (energy_consumption_j × EF_energy_source_j)
where EF_energy_source_j ∈ cat_energy_sources ∀ j ∈ subject_production

3: co2_trans ← ∑ (distance_km_k × EF_transport_k × trips_count_k)
where EF_transport_k = f(vehicle_type, fuel_type) via UBA/TREMOD matrix

4: co2_energy ← ∑ (consumption_l × EF_energy_source_l)
where phase_l ∈ {BAUPHASE, FIRMA_GESAMT, NUTZUNG}

5: co2_waste ← ∑ (quantity_kg_m × (1 – recycling_pct_m) × EF_waste_m)

6: co2_water ← ∑ (consumption_m3_n × EF_water_type_n)

7: co2_total ← co2_mat + co2_prod + co2_trans + co2_energy + co2_waste + co2_water

8: UPDATE dpp_subjects SET co2_total_kg = co2_total,
co2_materials_kg = co2_mat, co2_production_kg = co2_prod, ...
WHERE id = subject_id -- Materialized Aggregate Cache

Time Complexity: O(n) where n = total lifecycle entries across 6 dimensions

4.2 Materialized Aggregate Pattern

The seven CO_2 columns directly embedded in **dpp_subjects** (co2_total_kg, co2_materials_kg, co2_production_kg, co2_transport_kg, co2_energy_kg, co2_waste_kg, co2_water_kg) implement a ***Materialized Aggregate*** pattern. This denormalization is deliberate: dashboard queries retrieving CO_2 data for dozens or hundreds of subjects avoid expensive 6-way JOINs across lifecycle tables. The cache is invalidated and recomputed via explicit *recalc-co2* API calls, providing eventual consistency with user-controlled refresh – an appropriate trade-off for a compliance system where data entry and reporting are temporally separated.

4.3 Transport Emission Factor Matrix

Transport CO_2 calculation uses a **vehicle-type × fuel-type** emission factor matrix derived from UBA/TREMOD 2024 data.[10] The matrix is implemented as a dual-source system: the **cat_transport_types** database table stores official EF values with *fuel_type*, *load_capacity_kg*, and *euro_norm* metadata, while a hardcoded JavaScript matrix in the frontend provides instant UI feedback for common vehicle-fuel combinations. This hybrid approach balances latency (no API call for common lookups) with extensibility (new combinations added via catalog).

The transport CO_2 formula: $\boldsymbol{CO_2 = distance\_km \times EF(vehicle, fuel) \times trips\_count}$ uses a km-based model rather than the traditional tonne-kilometre (tkm) approach, reflecting real-world logistics where load weight is often unknown or variable. This design decision is documented in the schema comments

and produces comparable results for last-mile delivery scenarios typical in construction and furniture logistics.

5. Self-Learning Classification Pipeline

5.1 Three-Stage Material Matching

The PDF bill-of-materials (BOM) import pipeline implements a **three-stage cascading classifier** that improves classification accuracy with each tenant's usage. When a PDF is uploaded, the system extracts material designations via pdfplumber and attempts classification through the following pipeline:

Algorithm 2: Three-Stage BOM Classification Pipeline

```
Input:  designation (raw text from PDF), tenant_id
Output: (material_type_id, confidence, match_source)

Stage 1 – Keyword Matching (80+ hardcoded rules):
  FOR each keyword_rule IN KEYWORD_MAP:
    IF lowercase(designation) CONTAINS keyword_rule.keyword:
      RETURN (lookup_material_type(keyword_rule.category),
              HIGH if keyword_rule.priority = 1 else MED,
              'KEYWORD')
  → Priority: longest match first (e.g., 'kvh fichte' > 'holz')

Stage 2 – Catalog Name Matching (tenant + system):
  FOR each catalog_entry IN cat_material_types
    WHERE tenant_id IN (current_tenant, '001'):
      IF lowercase(designation) CONTAINS lowercase(entry.name)
         OR lowercase(designation) CONTAINS lowercase(entry.code):
        RETURN (entry.id, MED, 'CATALOG')
  → Self-learning: new tenant catalog entries auto-expand matching

Stage 3 – BOM Category Fallback:
  fallback ← cat_bom_category_defaults[bom_category]
  IF fallback EXISTS:
    RETURN (fallback.material_type_id, LOW, 'FALLBACK')

DEFAULT: RETURN (NULL, NONE, 'UNMATCHED')
```

5.2 Confidence Scoring and Data Quality Framework

Each material classification carries explicit provenance metadata stored in the **subject_materials** table: *mass_method* records the weight determination approach (EXPLICIT, CATALOG_PROFILE, CATALOG_TYPICAL, UNIT_IS_MASS, MANUAL_OVERRIDE, or MISSING), while *mass_confidence* provides a four-level quality indicator (HIGH, MED, LOW, NONE). This framework enables downstream analytics to weight CO_2 calculations by confidence level – a feature that is unusual in commercial LCA tools and represents a contribution to transparent environmental accounting.

5.3 Unit Factor Learning System

The **material_type_unit_factor** table implements a ***self-learning unit conversion system***. When users specify that 1 piece of "KVH 60×120mm" weighs 3.5 kg, this fact is stored as a tenant-specific conversion factor with source attribution (SYSTEM vs. USER) and confidence scoring. Subsequent imports of the same material-unit combination automatically apply this learned factor. The **UNIQUE(tenant_id, material_type_id, unit)** constraint ensures one canonical factor per combination, while the multi-tenant design prevents cross-contamination of learned data between organizations.

6. AI-Ready Data Architecture and Analytics Potential

6.1 Architectural Patterns for ML Integration

While the current system operates on rule-based classification, the data architecture has been designed with explicit **AI-readiness** in mind. Several architectural decisions enable future machine learning integration without schema migration:

Table 5: AI-Ready Architectural Patterns in the Data Model

Pattern	Current Implementation	ML Extension Path
Confidence Metadata	mass_confidence (HIGH/MED/LOW/NONE) stored per material	Training labels for classification model; active learning on LOW-confidence items
Provenance Tracking	mass_method, mass_explain text field for audit	Feature engineering from explanation text; model interpretability anchoring
Keyword Repository	matching_keywords TEXT in cat_material_types	Training corpus for NLP-based material NER; embeddings for similarity search
Audit Trail	Full operation log with user, action, resource, timestamp	User behavior modeling; anomaly detection; recommendation engine
JSONB Fields	fiber_composition, svhc_details, certifications	Flexible schema for ML model outputs; feature stores; prediction caching
Tenant Isolation	Every table filtered by tenant_id	Federated learning across tenants; privacy-preserving analytics

6.2 Predictive Analytics Opportunities

The accumulated lifecycle data enables several predictive analytics applications that go beyond simple reporting:

- **CO_2 Prediction for Early-Stage Projects.** Given the structured relationship between subject_type, product_category, and CO_2 outcomes across completed projects, a regression model could predict total CO_2 from partial lifecycle data (e.g., materials-only). The materialized CO_2 cache in dpp_subjects provides immediate training data with labeled outcomes.
- **Material Substitution Recommender.** The cat_material_types catalog with emission_factor, is_sustainable, recyclable, and recycled_content_pct fields contains sufficient features to build a similarity-based recommender that suggests lower-carbon material alternatives. The self-learning catalog ensures the feature space grows with tenant usage.
- **Supply Chain Risk Scoring.** The suppliers table captures 31 compliance dimensions (LkSG, CBAM, REACH, conflict minerals, sustainability rating, incident tracking). Combined with the subject_suppliers distance data and material_type linkage, this enables multi-factor supply chain risk models with geographic and compliance-weighted scoring.
- **Anomaly Detection in Lifecycle Data.** The audit_log table with action, resource, and timestamp enables temporal pattern analysis. Statistical process control methods could flag unusual data entry patterns (e.g., CO_2 values outside expected ranges for a material category), supporting data quality assurance in a regulatory context.

- **ESG Score Forecasting.** The subject_assessments table with assessment_period versioning enables time-series analysis of ESG scores. With the weighted criteria from cat_assessment_criteria (including weight NUMERIC and eu_regulation_ref), trend analysis and regulatory gap prediction become feasible.

6.3 Planned AI Integration Architecture

The roadmap includes a **FastAPI proxy architecture** for AI integration: a secure intermediary service mediates between the DPP frontend and the Anthropic Claude API for three use cases: (1) *intelligent material recognition* from PDF imports, augmenting the rule-based pipeline with LLM-based entity extraction; (2) *EU ESPR compliance checking*, where the AI evaluates completeness of DPP data against regulation requirements; and (3) *contextual chat assistance*, providing domain-specific guidance for data entry. The proxy architecture ensures API keys remain server-side, rate limiting is centralized, and all AI interactions are logged in the existing audit framework.

7. ESG Assessment and Supply Chain Compliance Model

7.1 Multi-Domain ESG Scoring Framework

The ESG assessment system implements a **weighted, multi-domain scoring framework** through the *cat_assessment_criteria* and *subject_assessments* tables. Each criterion belongs to one of three ESG domains (Environmental, Social, Governance) with optional sub-domains, carries a numeric weight (*weight NUMERIC(5,3)*), and references the applicable EU regulation. The *assessment_period VARCHAR(20)* field enables temporal versioning (e.g., '2025-Q1', '2025-H1'), supporting longitudinal ESG tracking – a requirement for continuous improvement reporting under ESPR.

7.2 Supply Chain Compliance Data Model

The **suppliers** table (31 columns) encodes a comprehensive compliance framework covering the German Lieferkettensorgfaltspflichtengesetz (LkSG), EU Carbon Border Adjustment Mechanism (CBAM), REACH regulation, and conflict minerals requirements. Boolean flags (*lksg_compliant, cbam_relevant, reach_compliant, conflict_minerals_free*) provide binary compliance indicators, while *sustainability_rating VARCHAR(10)* and *risk_rating VARCHAR(20)* offer graduated assessment. The SVHC tracking in *subject_materials* (*svhc_substance, svhc_concentration_pct, scip_notified*) implements ECHA SCIP database requirements[11], creating an integrated compliance chain from raw material to published DPP.

8. Evaluation and Discussion

8.1 Extensibility Validation

The architecture's extensibility was validated through the Stage 8 expansion, which added FURNITURE and TEXTILE subject types. The expansion required: 5 new SQLAlchemy models, 3 new catalog tables (40 seed entries), 2 new detail tables with 70 combined columns, 2 new API endpoints, and 925 lines of new frontend code across 2 new JavaScript modules. Critically, the expansion required **zero changes to existing lifecycle tables or the CO_2 aggregation algorithm** – the

[11] European Chemicals Agency (ECHA), SCIP Database for Substances of Concern In articles as such or in complex objects (Products), 2021.

polymorphic Subject model and generic lifecycle router absorbed the new types without modification. This demonstrates that the Class Table Inheritance pattern successfully decouples type-specific complexity from shared lifecycle logic.

8.2 Query Performance Considerations

The materialized CO_2 aggregate pattern reduces dashboard query complexity from O(6 × JOIN) to O(1) per subject. For a tenant with 100 active subjects, the subject list query retrieves pre-computed CO_2 totals directly, avoiding the cost of aggregating across potentially thousands of lifecycle entries. The trade-off – eventual consistency requiring manual recalculation – is acceptable because DPP data entry and reporting follow distinct temporal patterns: users typically complete data entry over days or weeks, then trigger recalculation before generating reports.

8.3 Limitations

Several limitations merit discussion. First, the *keyword-based classification* is inherently limited to the 80+ hardcoded rules and tenant-specific catalog entries; novel materials without matching keywords require manual classification. The planned AI integration (Section 6.3) addresses this gap. Second, the *VARCHAR(3) tenant_id* limits the platform to 999 tenants – sufficient for the current B2B market segment but requiring migration for high-volume scenarios. Third, the current system lacks *real-time CO_2 updates* via database triggers; the explicit recalculation model was chosen for predictability and debuggability in a compliance context.

9. Conclusion and Future Work

This paper presented the data architecture of FLX:DPPplus, a multi-tenant Digital Product Passport platform implementing EU ESPR compliance across seven product types. Our key contributions are: (1) a formalized 50-table data model with Class Table Inheritance, dual-scope catalogs, and cross-schema tenant isolation; (2) a six-dimensional CO_2 aggregation algorithm with materialized caching; (3) a three-stage self-learning classification pipeline with explicit confidence scoring; and (4) an analysis of AI-ready architectural patterns that enable future machine learning integration without schema changes.

The platform has been validated through production deployment and client demonstrations, with the Stage 8 furniture/textile expansion confirming the architecture's extensibility. The self-learning catalog system demonstrates that rule-based approaches with transparent confidence scoring can provide effective classification in regulated domains where model explainability is paramount.

Future work will focus on three areas: (1) **AI-augmented material classification** via LLM integration through the planned FastAPI proxy, measuring improvement in classification accuracy and confidence distribution; (2) **federated CO_2 benchmarking** across tenants using differential privacy to enable industry comparisons without exposing proprietary data; and (3) **predictive DPP completeness scoring** that uses accumulated data patterns to guide users toward regulation-compliant data entry. These extensions leverage the AI-ready data architecture described in this paper, requiring no structural changes to the existing 50-table schema.

Acknowledgements

The FLX:DPPplus platform was developed by flx-business GbR, Böblingen, Germany. The authors thank the EU ESPR regulatory framework for providing the motivation and the Umweltbundesamt (UBA) for the publicly available TREMOD emission factor data that underpins the transport CO_2 calculations.