CityData Harmonizer

AI-Assisted Smart City Data Integration - Energy Observation Use Case

MiMathon 2026 | GEX Team | Porto Use Case 3

Team Members:
Olaf-Gerd Gemein | Business Architect, OASC Germany | oggemein@googlemail.com
Mohamed Ben Kedim | Tech Lead, Software Engineer | mohamedbenkedim@gmail.com
Marwen Chaabouni | Project Manager | marwen123.c@gmail.com

Cities are drowning in data

Use Case 3: Energy Data Harmonization

Multiple formats

CSV, XLSX, JSON. Different delimiters, encodings, and shapes. Every provider does it differently.

Conflicting schemas

energy_kwh vs consumption_kWh vs energyConsumed. Same concept. Twelve names. Zero compatibility.

No shared model

Without a canonical target, comparing or aggregating across providers is impossible at scale.

Beyond Problem: Can we make the harmonization process replicable? How can we develop a reusable, domain agnostic system?

A new data model : EnergyConsumptionObserved

Core — measurement all required 8

id | type | dateObservedFrom | dateObservedTo | temporalResolution | energyType | consumption | unitCode

Context — who, what, how much 17

flowDirection | supplySource | sector | buildingType | floorArea | consumptionIntensity | energyService | consumptionPoint | cost | costCurrency | tariffPeriod | dataQuality | isNormalised | primaryEnergyFactor | co2EmissionFactor | co2Emissions | phase

Weather — energy drivers 10

temperature | heatingDegreeDays | coolingDegreeDays | degreeDayBaseTemp | relativeHumidity | windSpeed | globalSolarRadiation | precipitation | cloudCover | refWeatherObserved

Relationships — linked entities 6

refConsumptionPoint | refBuilding | refDevice | refOrganization | refOperatingArea | refWeatherObserved

Smart Data Models investigation result

EnergyConsumptionObserved

Capability	ACMeasure-ment	Consumption-Cost	Energy-Consumer	SmartMetering-Obs.	Ours
Multi-fuel (gas, heat, H_2...)	× elec only	◐ free text	× grid only	◐ elec focus	✓ 12-value enum
Interval timestamps (start/end)	◐ single string	× year+month	×	◐	✓ from/to split
Building/meter/org relationships	◐ refDevice	◐ refPoint	×	◐	✓ 6 relationships
Flow direction (import/export)	◐ separate fields	×	×	×	✓ enum
Supply source (grid/onSite/storage)	×	×	×	×	✓ enum
Sector classification	×	×	×	×	✓ Eurostat-aligned
CO_2 emission factor + emissions	×	×	×	×	✓ native
Weather context (inline)	×	×	×	×	✓ 9 fields

Meet the Rangers

They were trained in the OASC Academy - (MIM Champions)

Adapter

DET

Profiler

AI

Planner

AI

Resolver

AI

Knowledge Engine

DET

Executor

DET

Validator

DET

SchemaSelector

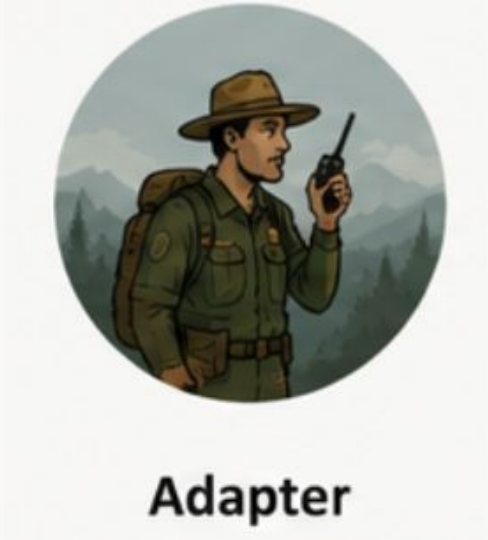
AI

PromotionManager

DET

AI-Powered

Deterministic (no LLM)



Adapter

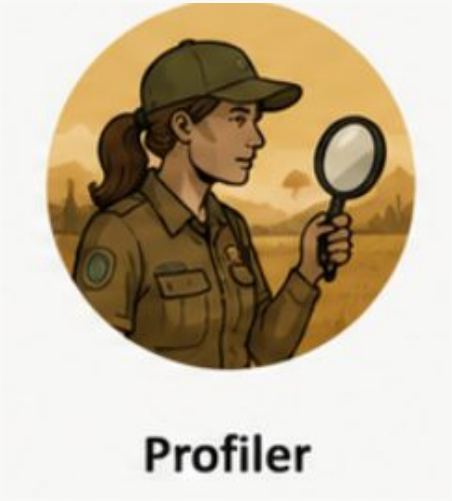
> *"I don't ask questions. I just read files."*

Input: CSV, XLSX, and JSON

Output: Turns any format into a clean list of records.

ACT 1

DETERMINISTIC



ACT 2

AI-POWERED

Profiler

> *"I observe. I never judge. Well, almost never."*

Input: Source file + Manifest files (Metadata)

Output: Source Intelligence File

Planner

> *"Give me a schema and I will give you a plan."*

ACT 3

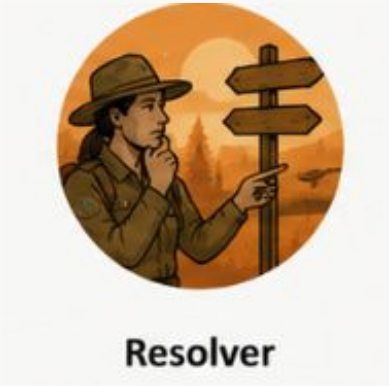
AI-POWERED

Maps source fields to canonical targets. Routes by confidence score.
Reason about required fields in case they are absent in the source data file.

Input: Source intelligence File + canonical schema + knowledge core.

Output: Harmonization Plan

Transformation functions:
`rename_field`, **`convert_unit`**, **`Aggregate`** ,**`map_value`**, **`parse_date`**, **`normalize_text`**, **`restructure`**, **`parse_geojson`**, **`combine_lat_lon`**



Resolver

> *"I handle the hard cases the Planner didn't want."*

Deep analysis of uncertain mappings. Final call before the gate.

ACT 4

AI-POWERED

Input: Uncertain mappings + full field context from SIF + canonical schema.
Output: Resolver Decision` per field: `INFERRED`, `UNMAPPED`, or `REVIEW`

Knowledge Engine

> *"Nothing passes without my approval. Nothing."*

Runs 4 rule-based checks. APPROVE, REJECT, or ESCALATE. No LLM involved.

ACT 5

DETERMINISTIC

Verifies every proposed mapping before data is written

schema_fit , vocabulary_membership , unit_compatibility , provenance_completeness

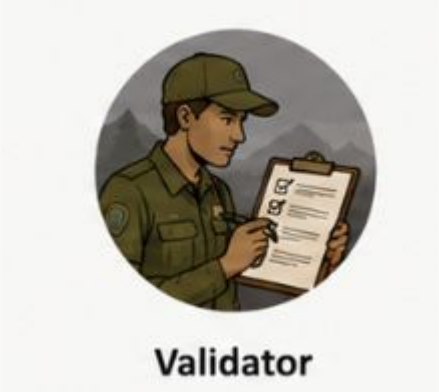
Executor

> *"No LLM. No drama. Just execution."*

Applies the verified plan row by row. 9 pure transform functions.

ACT 6

DETERMINISTIC



Validator

> *"I am the last line of defense. Don't disappoint me."*

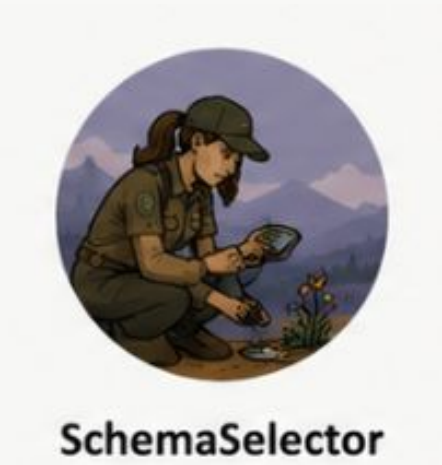
Checks every output record against the canonical JSON Schema.

ACT 7

DETERMINISTIC

Input: List of harmonized raws + canonical data model.

Output: ValidationResult` per record: `passed: True/False` + list of error messages.



SchemaSelector

> *"I pick the right model before the mission starts."*

What this ranger does:

Discovers the best OASC Smart Data Model via MCP. Runs pre-pipeline.

PRE-PIPELINE

AI-POWERED

PromotionManager

Promotion Manager

> *"I decide who gets promoted. Yes, really."*

Manages KB entry lifecycle: candidate to reviewed to approved to promoted.

ACT 5+

DETERMINISTIC

Key Concepts

HarmonizationPlan

The mapping table produced by the Planner: every source field labeled INFERRED, UNCERTAIN, or UNMAPPED with a confidence score and a transform.

Knowledge Base (KB)

The self-learning memory of the system. Confirmed mappings are stored and reused. LLM cost converges to zero for stable, recurring sources.

Quality Gate (KE)

Four deterministic checks: schema fit, vocabulary membership, unit compatibility, provenance completeness. Nothing passes without APPROVE.

Confidence Routing

Above 0.85: auto-approved as INFERRED. Between 0.60 and 0.84: sent to Resolver as UNCERTAIN. Below 0.60: excluded as UNMAPPED.

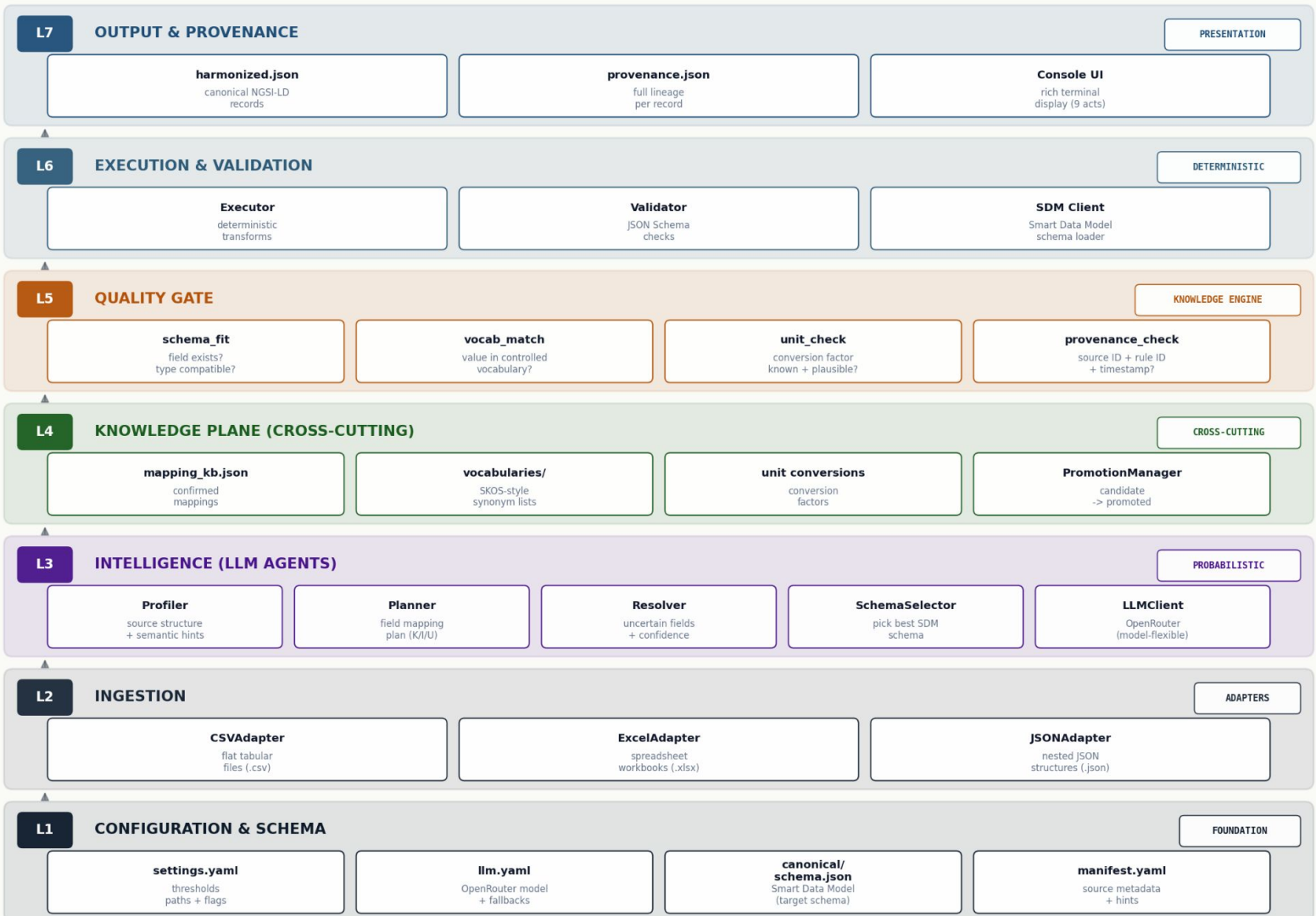
ProvenanceRecord

The full audit trail per output record: source row, all mappings applied, all transforms, KE verdict, validation result, agent versions.

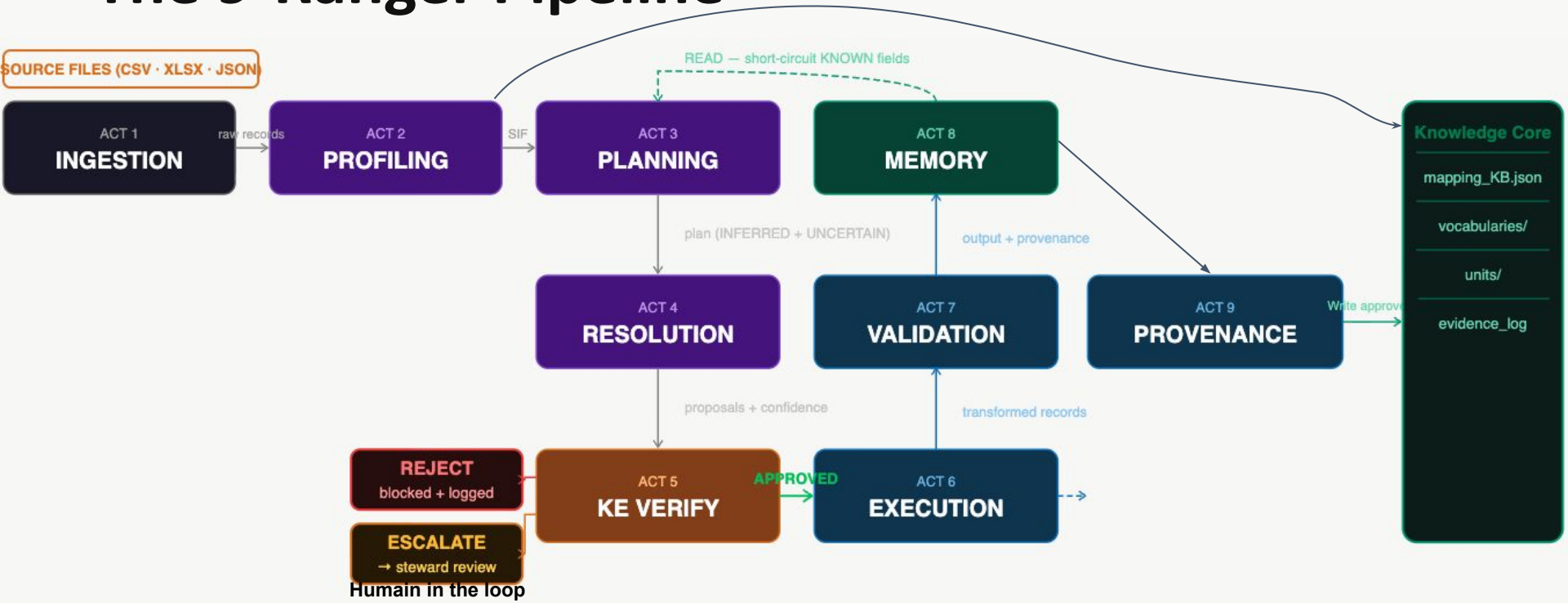
KB Short-circuit

If the Planner finds a promoted KB entry for a field, it skips the LLM call entirely. Zero cost, same quality, faster every run.
Positioning OASC as a for-front for data interoperability worldwide

Architectural Planes



The 9-Ranger Pipeline



Demo

Key Architectural Properties

Control / Data Separation

LLM decisions isolated from deterministic execution

Deterministic Reproducibility

Same HarmonizationPlan + same data = identical output every time

Self-Learning Convergence

KB grows with each run — LLM cost converges to zero for stable sources

Full Provenance

Every output record traceable to source row, mappings, transforms, and all verdicts

MCP Interoperability

Auto-discovers any FIWARE Smart Data Model via SmartDataModels MCP server

Quality Gate

4-check deterministic firewall between AI-generated plans and production data

Format-Agnostic

CSV, XLSX, JSON via pluggable adapters — add new formats without pipeline changes

Schema-Agnostic

Switch canonical model by changing schema.json or using --auto-schema flag

What the rangers learned from the MIMs

MIM	MIM Logic	How it is satisfied in the system
MIM1	Every entity must carry a unique, persistent identifier	The executor auto-generates urn:ngsi-ld:EnergyConsumptionObserved:<uuid> for every harmonised record
MIM1	Cross-system entity linking must be possible	refMeter and refBuilding fields carry NGSI-LD URNs linking each observation to its physical meter and building across any system
MIM1	Semantic typing of entities	Every record carries "type": "EnergyConsumptionObserved" — a declared, unambiguous semantic type aligned to NGSI-LD
MIM1	Ontology-level mapping across data sources	The AI agent pipeline (Profiler -> Planner -> Resolver) maps heterogeneous field names to canonical semantic equivalents with confidence scores and persisted evidence
MIM2	Data models must be explicit, documented, and semantically unambiguous	EnergyConsumptionObserved.schema.json defines all attributes with types, units, enumerations, and semantic references to schema.org and FIWARE SDMs
MIM2	Models must build on standardised community models where possible	The canonical schema composes FIWARE Smart Data Models via allOf/$ref — GSMA-Commons, Location-Commons, EnergyConsumption core
MIM2	Cross-model transformation to a common model is required	The full 9-act harmonisation pipeline transforms any provider schema (different field names, units, vocabularies) into one canonical model automatically

What the rangers learned from the MIMs

MIM	MIM Logic	How it is satisfied in the system
MIM3	Data assets must carry adequate metadata, human and machine-readable	Each source manifest.yaml is a structured machine-readable metadata descriptor: owner, domain, format, columns, units, data quality notes
MIM3	Provenance and trust lineage must be traceable	provenance.json captures full lineage per record: source row -> mapping IDs -> every transform applied with parameters -> Knowledge Engine verdict -> validation result
MIM7	Geospatial data must be encoded in open standards	The combine_lat_lon and parse_geojson transforms produce GeoJSON Point geometry — OGC/MIM7 compliant encoding
MIM7	Contextual datasets must comply with MIM1 and MIM2	The location field is embedded inside the same NGSI-LD record that carries the MIM1 URN and MIM2 canonical schema — natively integrated

Next Steps

1. **NGSI-LD Standard Output format integration**
2. **MCP Client Integration to pick the best fitting SDM**
3. **Human in the loop : Expert Interaction with Agents**
4. **Setting up a community : OASC members**
5. **Funding and budgeting until 2026 and further**
6. **“Inspire Regulation” Contribution**
   The new metadata for AI systems
7. **Alignment with OASC mission on improving smart data model**

CityData Harmonizer

Any source. Any domain. One pipeline.

GEX Team | MiMathon 2026 | Porto Use Case 3