

Agentic Mapping of National Transplant Registry Data to an Extended OMOP Common Data Model using LinkML



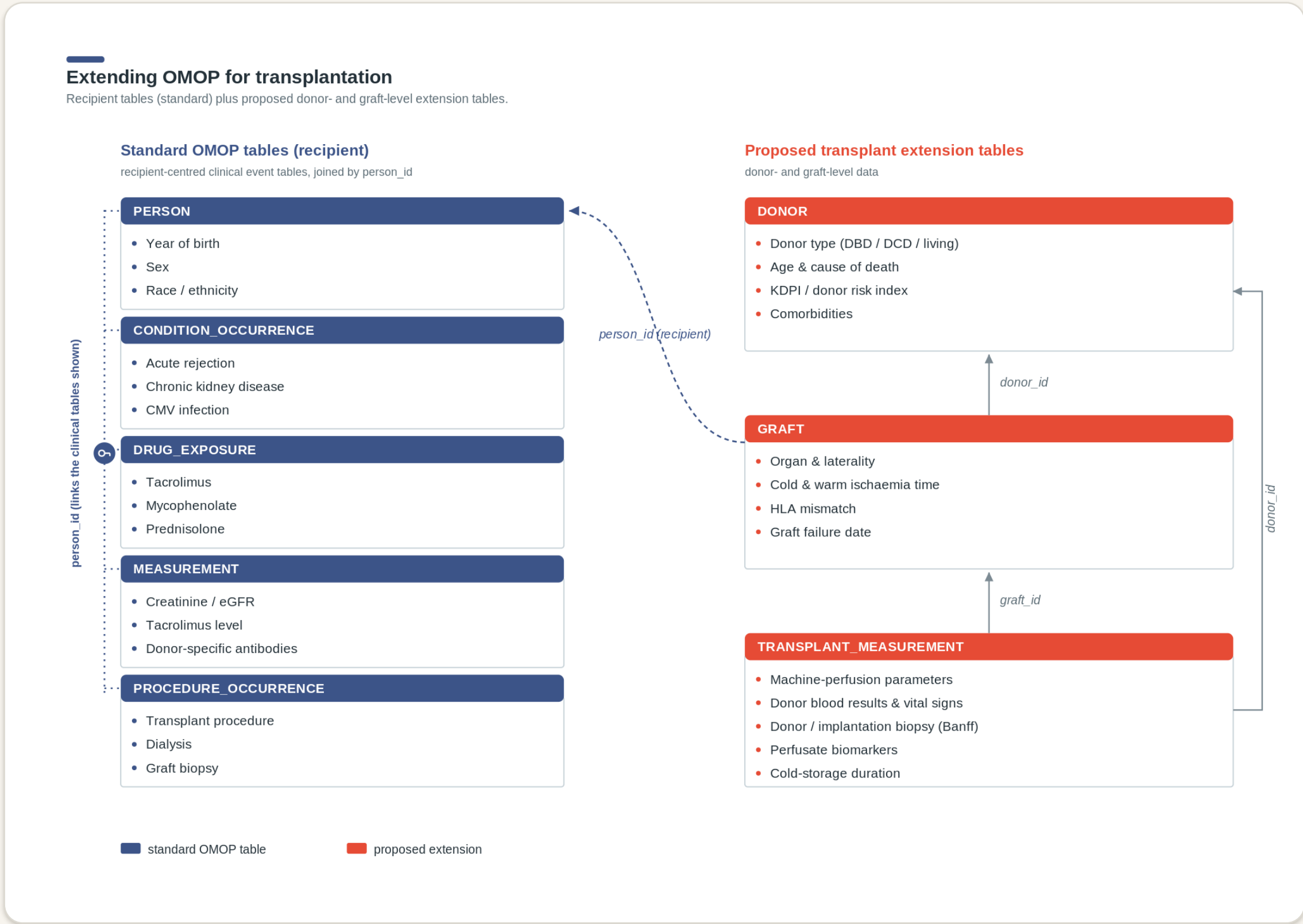
Aditya Borakati^{1,3}, Jack Wu¹, Chris Callaghan², Miriam Cortes-Cerisuelos³, James Teo³, William Bernals³, David Wallace^{3,4,5}

¹ King's College London · ² Guy's & St Thomas' NHS FT · ³ King's College Hospital NHS FT · ⁴ London School of Hygiene & Tropical Medicine · ⁵ Royal Free London NHS FT

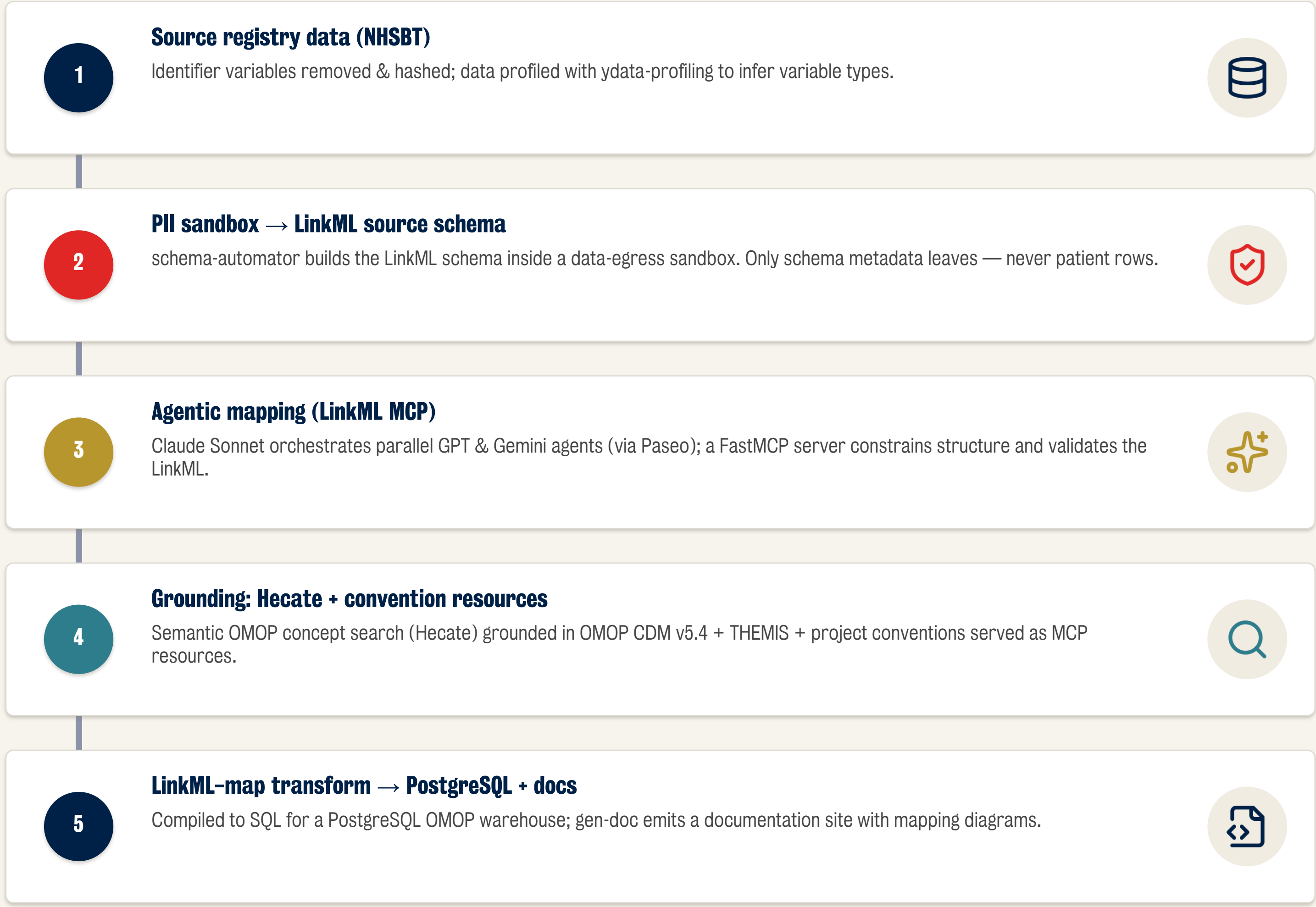
Background

- Solid-organ transplantation is a complex, life-saving procedure; donor organs are scarce and outcomes depend on interacting donor, graft, recipient and peri-operative factors.
- National registries (in the UK, NHS Blood & Transplant) record only high-level structured data in bespoke models — limiting linkage to granular EHR data and to external datasets such as HES.
- Transforming this data to the **OMOP Common Data Model** would create an internationally recognised standard for transplant research, audit and benchmarking.
- Registry volume prohibits manual mapping — so we developed an **AI-assisted (agentic) mapping** approach.

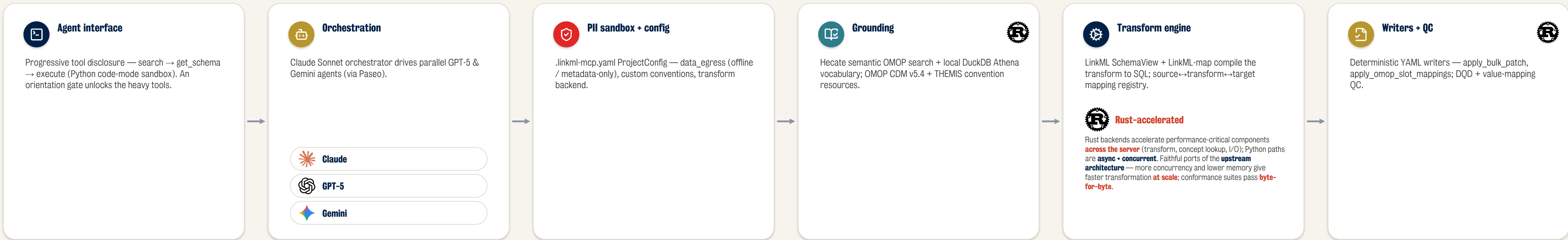
An Extended Target Schema



End-to-End Agentic ETL Pipeline

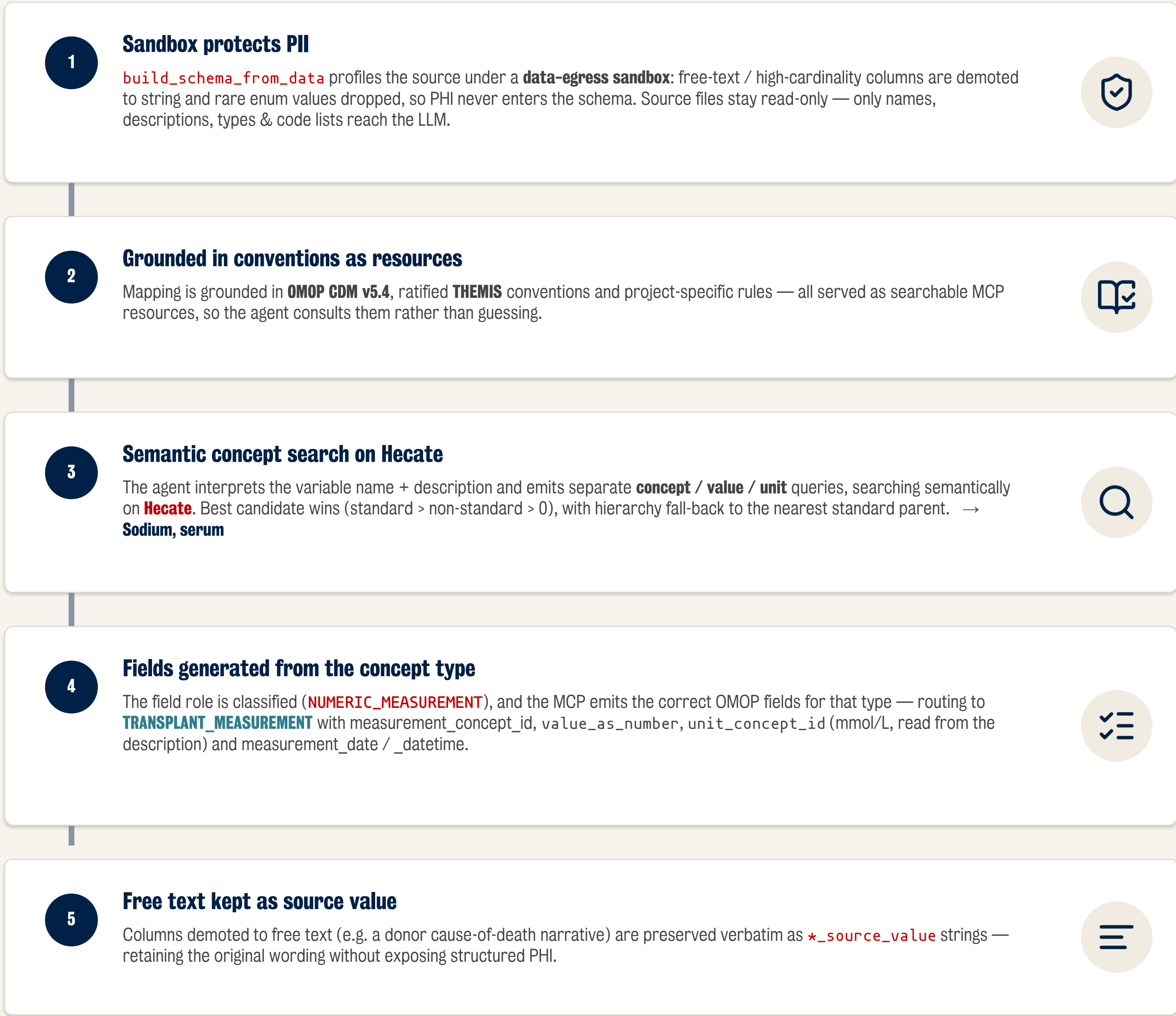


LinkML-MCP Server Architecture

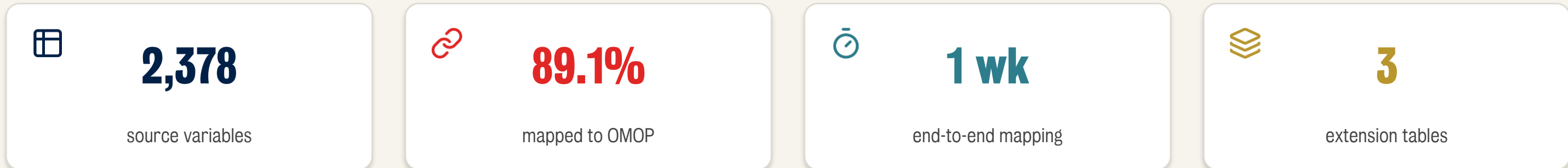


Inside the MCP: mapping one variable

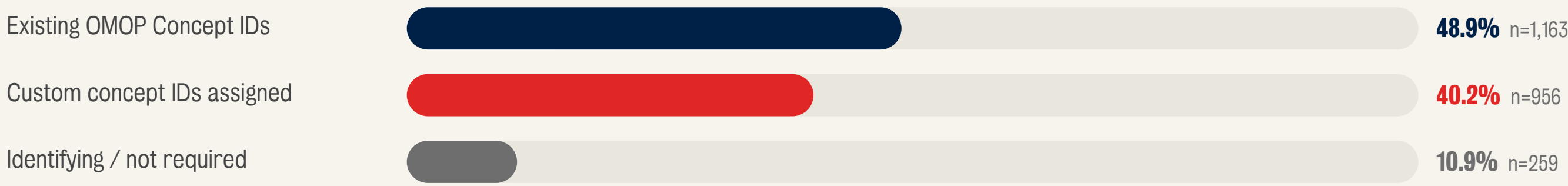
Worked example — source field **DON_NA** “Donor serum sodium at retrieval”



Results



Of 2,119 mapped variables:



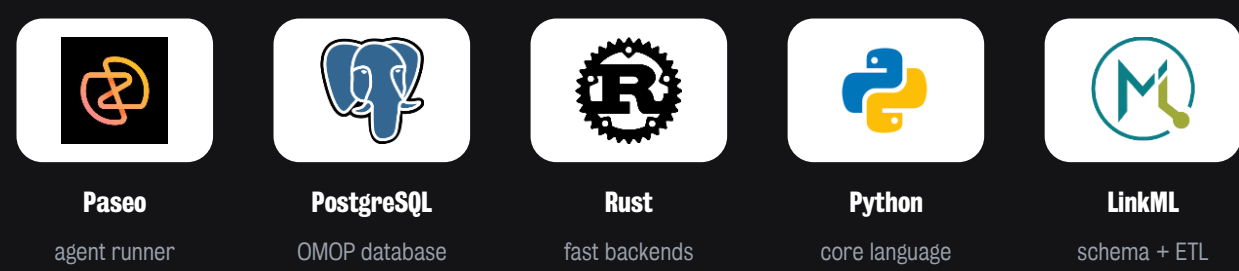
All LLM outputs were valid LinkML/YAML and generated a valid PostgreSQL OMOP database plus a live documentation site (QR, below).

Conclusions

- Agentic LLM workflows can **dramatically accelerate** mapping of large legacy registries to OMOP — completed here in one week.
- LinkML + OMOP yields an end-to-end ETL with **auto-generated SQL and documentation**.
- Transplantation **cannot** be represented by standard OMOP alone: donor, graft and recipient data must be delineated.
- 40.2%** of concepts lacked an OMOP-supported ontology match — motivating model and vocabulary extensions, feeding the OHDSI Transplant Working Group.

REFERENCES
[1] Moxon SAT et al. LinkML: An Open Data Modeling Framework. GigaScience 2025. [2] YData Profiling. [3] PrefectHQ/FastMCP. [4] Farah M. mikfarah/yq. [5] getpaseo/paseo. [6] streamlit/streamlit. [7] OHDSI/DataQualityDashboard.

Structural quality validated with the OHDSI Data Quality Dashboard; pipeline tested on Synthea synthetic and real NHSBT registry data.



Documentation site
pioneer-docs