Declarative Data Standardization to OMOP-CDM using LinkML

Alberto Labarga[1], Tom Stone[2]
[1]MirrorHealth SL, Pamplona, Spain, [2]Roche Products Limited, Welwyn Garden City, UK

The secondary use of clinical data for research is severely constrained by data heterogeneity across healthcare systems. Although the OMOP Common Data Model (OMOP-CDM)[1] provides a robust and widely adopted standard for observational research, the Extract–Transform–Load (ETL) process required to populate OMOP remains a major barrier. Traditional ETL pipelines rely on bespoke, imperative code (e.g., SQL or Python), making them costly to develop, difficult to maintain, and hard to reproduce. This poster presents a paradigm shift in OMOP standardization, reframing ETL as a declarative, model-driven, and AI-augmented process using the Linked Data Modeling Language (LinkML)[2].

LinkML

Declarative data model

A LinkML schema defines the structure and semantics of a data model. It's the top-level specification that declares:

- What kinds of things exist (**classes**)
- What attributes or properties they have (**slots**)
- What controlled values those properties can take (**enums**)
- How these elements relate to each other

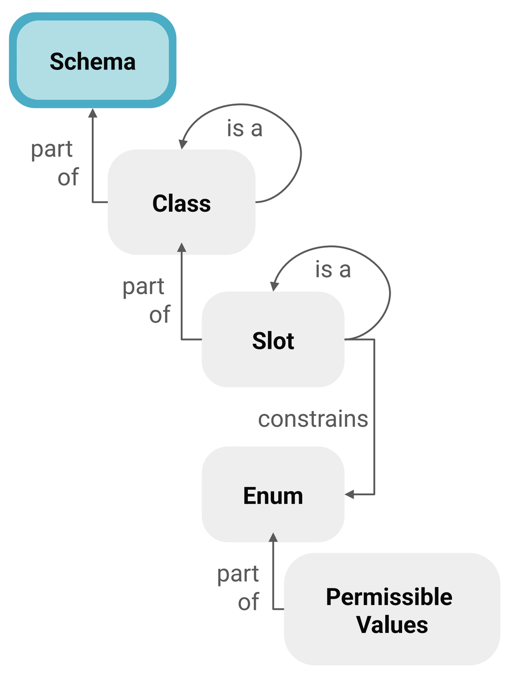

Figure 1. LinkML schema outline


```yaml
id: Synthea
prefixes:
  linkml: https://w3id.org/linkml/
enums:
  gender_enum:
    permissible_values:
      M:
        description: M
      F:
        description: F
classes:
  patient:
    slots:
      ID:
        examples:
        - value: cff50893-315f-1450-dce9-ceaeac1668c6
        identifier: true
        range: string
      BIRTHDATE:
        examples:
        - value: '1946-04-27'
        range: datetime
      DEATHDATE:
        examples:
        - value: '1989-01-28'
        range: datetime
      GENDER:
        examples:
        - value: M
        range: gender_enum
```


Figure 2. Source data LinkML profile generated with *schemauto*


```yaml
id: omop_cdm_v54
prefixes:
  linkml: https://w3id.org/linkml/
  omop: https://athena.ohdsi.org/search-terms/terms/
enums:
  omop_gender_values:
    reachable_from:
      source_ontology: omop:Gender
classes:
  person:
    description: The Person Domain contains records that uniquely identify each patient.
    attributes:
      person_id:
        description: A unique identifier for each person.
        range: integer
        identifier: true
        required: true
      gender_concept_id:
        description: A foreign key that refers to an identifier in the CONCEPT table.
        range: integer
        bindings:
          - range: omop_gender_values
        domain: concept
        identifier: false
        required: true
      year_of_birth:
        description: The year of birth of the person.
        range: integer
        identifier: false
        required: true
```


Figure 3. Target data LinkML profile for OMOP-CDM v5.4.1

LinkML-map

Declarative transformation

We implemented a declarative clinical data to OMOP-CDM standardization framework built on the LinkML ecosystem. First, source database schemas were automatically generated using LinkML Schema Automator[3], producing formal class–slot–enum definitions directly from relational databases. These schemas were semantically enriched using BioPortal-based annotators to suggest candidate mappings to OMOP standard vocabularies. Second, a canonical LinkML representation of the OMOP-CDM was used as the target schema. Third, a large language model (LLM) was employed to generate a LinkML-Map[4] transformation specification aligning source and OMOP schemas. To ensure reliable structured output, we used BAML's schema-aligned parsing[5], constraining the LLM to produce syntactically valid LinkML-Map artifacts. Finally, the LinkML-Map Transformer executed the declarative mappings to generate OMOP-CDM–conformant data, with validation performed against the OMOP LinkML schema.


```yaml
class_derivations:
  person:
    populated_from: patient
    slot_derivations:
      person_id:
        expr: "surrogate_key"
      person_source_value:
        populated_from: ID
      gender_concept_id:
        populated_from: GENDER
      gender_source_value:
        expr: GENDER
      year_of_birth:
        expr: BIRTHDATE.year
      month_of_birth:
        expr: BIRTHDATE.month
      day_of_birth:
        expr: BIRTHDATE.day
      birth_datetime:
        populated_from: BIRTHDATE
```

```yaml
enum_derivations:
  GENDER_enum:
    populated_from: GENDER_enum
    permissible_value_derivations:
      8507:
        populated_from: "M"
      8532:
        populated_from: "F"
```


Figure 5. LinkML-map transformation definition

Generate profile data using **schemauto**	Generate mapping proposal using **BAML**	Transform data using **linkml-map**	Validate data using **linkml validate**

Figure 4. LinkML-based standardization pipeline

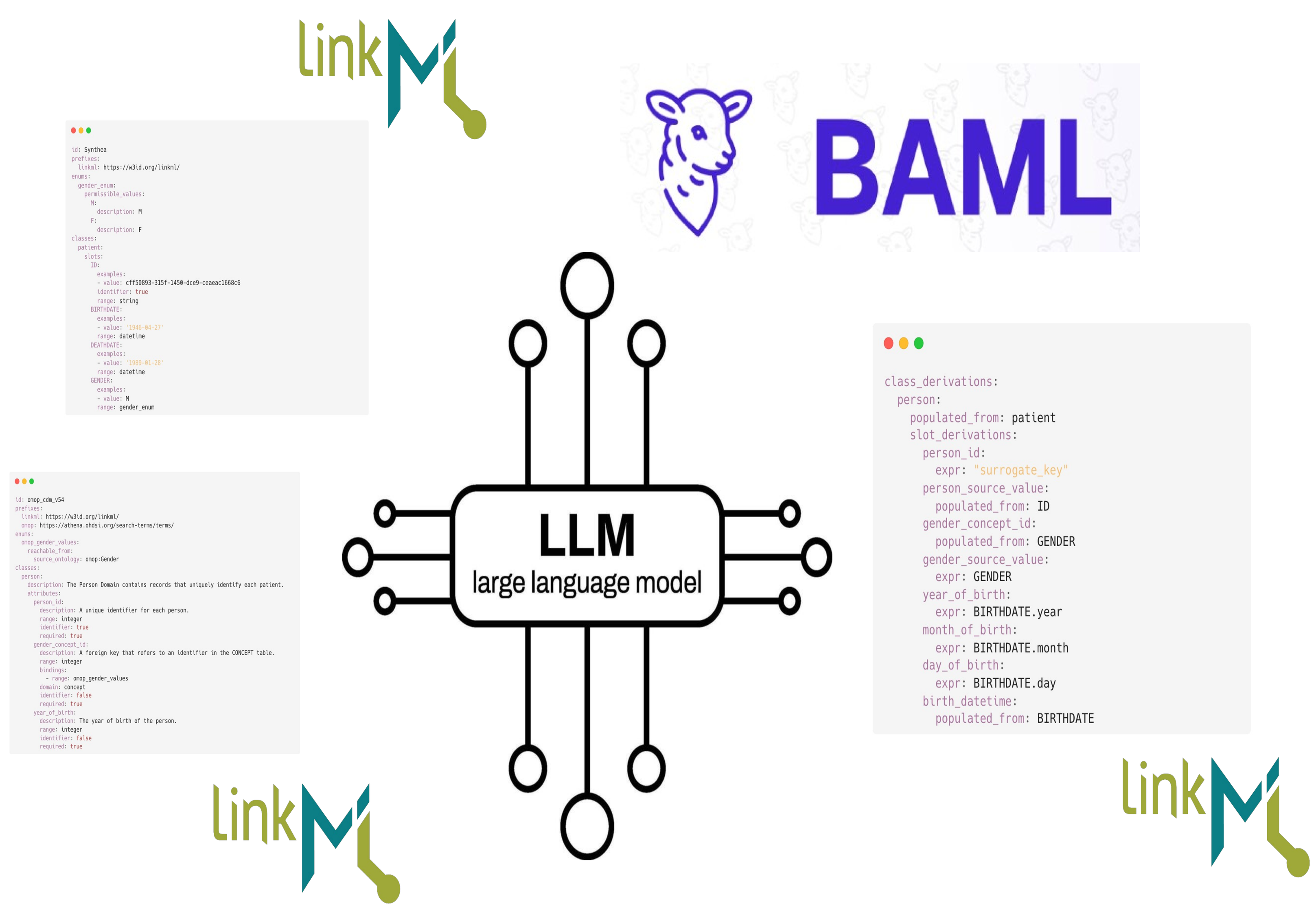

Figure 6. BAML's schema-aligned parsing is used to generate proper LinkML-map definition for the transformation between two LinkML data models using any LLM

Results

The proposed approach yielded qualitative improvements over traditional ETL pipelines. All transformation logic was externalized into human-readable, version-controllable YAML artifacts, improving transparency, auditability, and reproducibility. The use of structured schemas and schema-constrained LLM outputs substantially reduced malformed mappings and simplified expert review. The workflow promoted modularity by separating source modeling, vocabulary grounding, mapping generation, and execution, allowing independent evolution of each component. Rather than replacing human expertise, the framework functioned as a mapping co-pilot, automating repetitive tasks while enabling domain experts to review and refine mappings.

Conclusion

This work demonstrates that declarative, model-driven ETL using LinkML can significantly lower the technical barrier to OMOP-CDM adoption. By replacing brittle imperative scripts with explicit schemas and mapping specifications, the framework enhances maintainability, reproducibility, and trust. Structured AI integration improves efficiency while preserving human-in-the-loop governance, which remains essential given the semantic ambiguity of clinical data and the limitations of LLMs. The approach balances standardization and local customization, enabling institutions to adapt mappings to local data while ensuring strict OMOP compliance. Declarative data standardization provides a scalable foundation for federated observational research and aligns well with OHDSI's goals of reusable, transparent, and shareable analytics pipelines.

References

[1] OHDSI. The Book of OHDSI. https://ohdsi.github.io/TheBookOfOhdsi/

[2] Moxon, Sierra AT, et al. "LinkML: An Open Data Modeling Framework." GigaScience (2025)

[3] LinkML Schema Automator. https://linkml.io/schema-automator/

[4] LinkML-Map Documentation. https://linkml.io/linkml-map/

[5] Boundary ML (BAML). https://github.com/baml-org

mirrorhealth.eu