

# A Demonstration of Layered Schema Architecture as a Semantic Harmonization Tool

Burak Serdar<sup>1</sup>

<sup>1</sup>Cloud Privacy Labs, Colorado, USA

## Abstract

We will demonstrate Layered Schema Architecture (LSA) as a novel semantic harmonization tool to make health-related records research and AI-ready. LSA is designed to capture data, contextual metadata, and semantics to preserve the correct meaning of data through transformations in an ecosystem where multiple data standards and conventions exist. The specific examples in this demonstration will deal with the harmonization of clinical data and social needs screening survey data collected from disparate systems using multiple data standards and ontologies.

## Keywords

Layered schemas, semantic interoperability, semantic harmonization, data warehousing, FHIR, OMOP,

## 1. Introduction

The longitudinal health data from large diverse populations with varying social, economic, geographic, and environmental conditions is a highly valuable resource for medical and public health researchers through the creation of various data commons where disparate data are structured and harmonized to expand research options. Many challenges hinder the complete and efficient capture and exchange of health data, including: 1) a lack of semantic interoperability across systems; 2) the varying adoption of data standards within and between systems; 3) a lack of standardized metadata; and 4) the poor integration of electronic health records (EHR) data with data from other relevant sources such as social services, environmental measurements, patient-entered, and data collected from wearable devices.

A challenge to semantic harmonization is the ingestion of data from increasingly diverse sources where vendor specific variations and non-standard representations are common. Some examples include data from wearable devices, different social needs screening tools, and public datasets. Another challenge is the difficulty in interpreting data based on context. Semantic harmonization has to take into account the context in which data are captured as well as the context in which data will be used after transformation. A common data model (CDM) such as Observational Medical Outcomes Partnership (OMOP [1]) helps to integrate data sets coming from multiple sources and transform them for use by researchers; however, most source data are developed for the fit and purpose of specific organizations.

---


*The Eighth Joint Ontology Workshops (JOWO'22), August 15-19, 2022, Jönköping University, Sweden*

✉ [bserdar@cloudprivacylabs.com](mailto:bserdar@cloudprivacylabs.com) (B. Serdar)

🌐 <https://cloudprivacylabs.com/> (B. Serdar)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

A common approach for the semantic harmonization of such data involves developing extract-transform-load (ETL) scripts for each data source. This approach hardwires source-specific variations in data (such as measurement units, coding systems, organizational conventions or extensions to standards) while losing valuable contextual information (e.g. measurement variations for laboratory results, patient-reported vs. physician-entered data). This approach also lacks a standardized and reproducible way to capture and distribute relevant metadata.

The “Layered Schema Architecture” (LSA) [2] is an open-source technology developed by Cloud Privacy Labs to enable semantic interoperability in a data ecosystem where multiple overlapping data standards exist. In a partnership with the DARTNet Institute, we are exploring the use of LSA to semantically harmonize and translate health and health related data captured from disparate sources in different formats into OMOP for research and AI purposes. The specific examples of this project use clinical and social determinants of health (SDoH) related data, but the framework is domain and ontology agnostic and can be applied to various use-cases.

## 2. Layered Schemas and Labeled Property Graphs

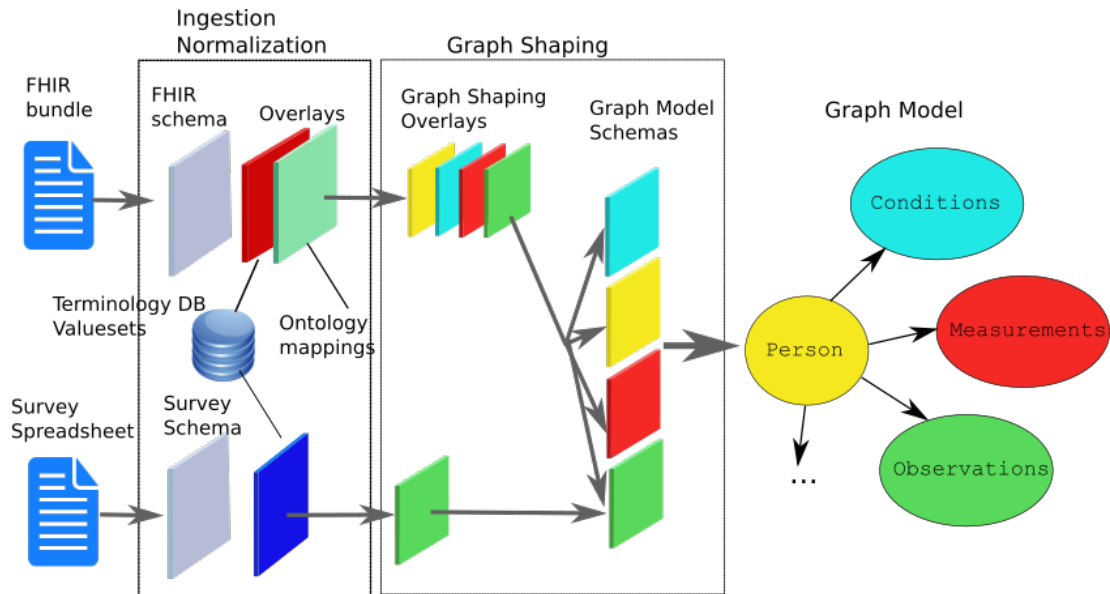
A *schema* is a machine-readable document that describes the structure of data. JSON and XML schemas are widely used to generate executable code from specifications and to check structural validity of data. LSA extends schemas (such as FHIR[3] or OMOP schemas) with layers (overlays) to add semantic annotations. The semantic annotations add ontology mappings, contextual metadata, tags, and processing instructions that control data ingestion and normalization. A *schema variant* is composed of a schema and a set of overlays, and contains the combination of annotations given in the schema and the overlays.

A schema variant is represented using a *labeled property graph* (LPG) that has a node for each data field. An LPG is a directed graph where every node and edge contain a set of *labels* describing its type or class, and a set of *properties* that represent named values. An LPG allows assigning tags that represent different types of metadata to fields. A field may be a simple value, a structured object (e.g a JSON object, array, polymorphic object), or a reference to another schema.

Different schema variants can be used to ingest data that shows variations based on data source. Data variations can be structural (e.g. additional data fields, extensions) or semantic (e.g. measurements in different units, different ontologies or coding systems), and can be due to different vendor implementations, local conventions, or regulations. Ingesting structured data using a schema variant creates an LPG whose nodes combine the annotations from the schema variant and data values from the input.

## 3. Data Processing Pipeline with LSA

A high-level overview of the data processing pipeline to ingest FHIR messages and SDoH surveys is illustrated in Figure 1. FHIR ingestion uses a schema variant that enriches standard FHIR schemas with tags for terminology database lookups and data privacy information. The first overlay contains tags that specify lookups in a terminology database. These tags are used by the data ingestion logic to add OMOP concept ids for codes given in different coding systems. The second



**Figure 1:** Pipeline to ingest FHIR and SDoH survey data into a common graph model

**Table 1**

SDoH questionnaire data (PRAPARE) sample

patient id	question	answer
309613	Social Integration	More than 5 times a week
309613	Material Security - Food	N
415637	Material Security - Food	Y

overlay assigns data privacy vocabulary [4] terms: it adds `privacyClass: "dpv:Patient"` to `Patient`, and `privacyClass: "dpv-pd:Identifying"` to all personally identifying information fields.

The survey data used in our example are collected using Protocol for Responding to & Assessing Patients' Assets, Risks & Experiences (PRAPARE)[5] in spreadsheet form as shown in Table 1. The ingestion schema describes the columns of this spreadsheet. An overlay is used to specify a valueset to map question and answer values to OMOP concept ids (Table 2). This valueset is different for each data source as each organization codes these questions and answers differently.

The second stage converts the ingested data to the database graph model that organizes conditions, measurements, observations, etc. as clusters of nodes linked to a person object. This database graph model provides a convenient representation to perform searches, or to perform further normalizations on data such a de-duplication, or imputations. Conversion of ingested data to the database graph model is again done using schema variants. The overlays for the "graph shaping" stage assign graph queries (using openCypher [6] language) to target schema fields. Graph reshaping operation creates an instance of the target schema using the

**Table 2**

SDoH questionnaire OMOP concept id mappings valueset

question code	question concept id	answer	answer concept id
Housing Status	37020172	Y	45877994
Housing Status	37020172	N	45878245
Material Security - Food	37020774	Y	37079482
Material Security - Clothing	37020774	Y	37079033

A	B	C	D	E
observation_concept_id	observation_date	value_as_number	value_as_string	value_as_concept_id
37020032	2020-03-13	More than 5 times a week	More than 5 times a week	37079490
37020730	2022-03-30	N	N	45878245
46235507	2020-03-13	N	N	45878245
37020730	2020-03-13	N	N	45878245

**Figure 2:** OMOP observations output

nodes selected by the assigned graph queries. The ingested data is stored in a graph database (Neo4j[7]).

The PRAPARE survey data are ingested as Observations using an overlay for that assigns OMOP concept ids for questions and answers to the corresponding `concept_id` and `value_as_concept_id` fields. The overlay also adds the PRAPARE and Survey tags to the Observation labels. This allows us to differentiate between observations coming from clinical data and observations coming from surveys. Such metadata is useful in determining the situations under which data is captured, and in assigning “confidence levels” to data elements.

OMOP uses a relational data model, thus it is necessary to translate data stored in the graph database after the research population is identified. Once research population is identified, each Person graph is exported from the database, shaped to fit to OMOP schemas using another “graph shaping” stage, and exported in tabular format. A sample output for OMOP observations table is given in Figure 2.

## 4. Discussion

Our goal is to develop a reusable and scalable interoperability framework to ingest and semantically harmonize health-related data and metadata from disparate sources in a research data warehouse setting. While our focus is producing OMOP output for researchers, the architecture is flexible enough to accommodate other CDMs. Our efforts so far suggest that incorporating new data sources using LSA is faster compared to ETL scripting, and yields reusable artifacts.

The LSA approach offers a method for the management of relevant metadata that can be important in building high quality data sets. Such metadata may include information about the context in which data was captured (e.g. whether data was entered by a provider or self-reported), processed (e.g. whether data was imputed, or fuzzed for deidentification purposes), or stored (e.g. the source filename for the data point).

The use of labeled property graphs as the core data model offers unique opportunities, such as the ability to use multiple ontologies to tag data. Future work in this area will involve the

use of graph properties for data imputations, evaluating the effects of different graph models for building study populations, and the incorporation of natural language processing tools to tag textual data.

LSA is an open-source project [8]. The toolset, schemas, overlays, and value sets we developed during this project will be available on a GitHub repository.

## Acknowledgments

DARTNet Institute is the primary grantee of this project, and provided the sample datasets and original ETL scripts. We thank DARTNet Institute for their partnership and their valuable insight.

This project is supported by the Office of the National Coordinator for Health Information Technology (ONC) of the U.S. Department of Health and Human Services (HHS) under grant number 90AX0034, Semantic Interoperability for Electronic Health Data Using the Layered Schemas Architecture, total award \$999,990 with 100% financed with federal dollars and 0% financed with non-governmental sources. This information or content and conclusions are those of the author and should not be construed as the official position or policy of, nor should any endorsements be inferred by ONC, HHS, of the U.S. Government.

## References

- [1] O. D. D. Sciences, Informatics, Omop common data model, 2022. URL: <https://www.ohdsi.org/data-standardization/the-common-data-model/>.
- [2] Layered schemas, 2022. URL: <https://layeredschemas.org>.
- [3] Fast healthcare interoperability resources (fhir), 2022. URL: <https://hl7.org/fhir/>.
- [4] Data privacy vocabulary, 2022. URL: <https://w3c.github.io/dpv/dpv/>.
- [5] Protocol for responding to & assessing patients' assets, risks & experiences (prapare), 2022. URL: <https://prapare.org/>.
- [6] opencypher, 2022. URL: <https://opencypher.org/>.
- [7] Neo4j, 2022. URL: <https://neo4j.com>.
- [8] Lsa github, 2022. URL: <https://github.com/cloudprivacylabs/lsa>.