TOWARDS A PLATFORM FOR SCHEMA-LEVEL HARMONIZATION: DATA CATALOGUE, SCHEMA GENERATION, AND AUTOMATED MAPPING

Ivan Sohrab Gill (ivan_gill@sfu.ca), **Jun Duan (jun_duan@sfu.ca)**, Damion Dooley (damion_dooley@sfu.ca), and William Hsiao (wwhsiao@sfu.ca)

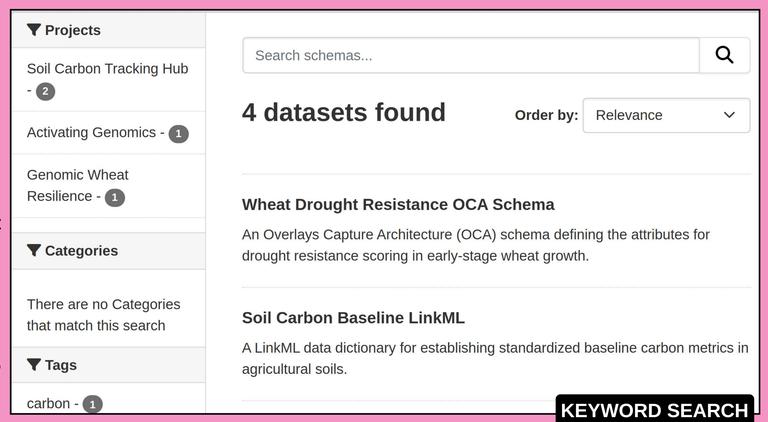
SIMON FRASER UNIVERSITY

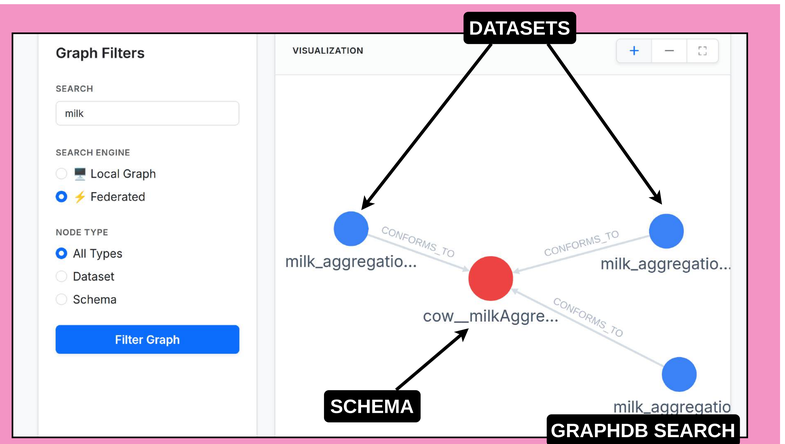
CENTRE FOR INFECTIOUS DISEASE GENOMICS AND ONE HEALTH

INTRODUCTION

Biological and biomedical datasets are increasingly generated independently across institutions, yet most lack the interoperability needed for cross-dataset integration and reuse. Data schemas address this by enforcing structured vocabularies and consistent field definitions, improving both machine readability and dataset discoverability. However, the infrastructure for schema-driven data practices remains fragmented: existing data portals prioritize dataset discovery over schema reuse and harmonization, and mapping data to schemas is largely manual, inconsistent, and costly. What is needed is not just individual schemas, but an ecosystem: shared tools and conventions that make schema adoption, editing, and harmonization a tractable part of the data lifecycle. Here we present our development of a suite of tools for schema discovery, schema editing, and automated within-schema harmonization, lowering the barrier to producing datasets that are interoperable by design.

Our schema catalogue integrates with DataHarmonizer to let users easily search, discover, and compare complex metadata schemas. It acts as a centralized backend that allows users to pull international standards directly into their spreadsheet GUI, simplifying cross-platform interoperability.





Before data collection
(design / planning)

- Browse schema library to find similar schemas
- Reuse controlled vocabularies and schema components
- Create or extend schemas for data collection

Find & reuse

Schema discovery
(schema catalogue)

- Search schemas by fields, data types, and controlled vocabularies
- Compare schemas
- Explore reusable components (picklists, enums, slots)

Find compatible schemas & terms

After / during data collection
(harmonization)

- Have existing/ad hoc data and schema
- Map values to standard terms and controlled vocabularies
- Harmonize fields and data types to standard schema

Leverag schemas & controlled vocabularies

Schema editing
(LinkML editor)

- Create new schemas
- Edit and extend existing schemas
- Manage fields, data types, and controlled vocabularies

Harmonized schema / data
(interoperable & reusable)

- Consistent schemas and field types
- Controlled vocabularies applied
- Data interoperable across systems

LLM harmonization
(within-schema mapping)

- Map data values to controlled vocabulary terms
- Harmonize field names and data types
- Suggest standard terms and resolutions

Our DataHarmonizer schema editor extension enables users to load, edit, and create LinkML schemas directly within the spreadsheet GUI, simplifying reuse of existing standards without familiarity with LinkML syntax.

Schema	Table	Table section	Field	Picklist
key				
Schema ID		Enum ID		
PathogenAgnostic		AnatomicalPartMenu		
PathogenAgnostic		BodyProductMenu		

Schema	Table	Table section	Field	Picklist	Picklist choices
key					
Schema ID		Picklist ID		Code	
PathogenAgnostic		AnatomicalPartMenu			Oropharynx (OP)
PathogenAgnostic		BodyProductMenu			Breast Milk
PathogenAgnostic		BodyProductMenu			Feces

schema.yaml (snippet)

```
body_product:
  slot_uri: GENEPIO:0001216
  any_of:
  - range: OrganismMenu
BodyProductMenu:
  permissible_values:
    Breast Milk:
      meaning: UBERON:0001913
    Feces:
      meaning: UBERON:0001988
```

PICKLIST

PICKLIST CHOICES

FUTURE WORK

- Integrate in-development `MenuManager` with **schema editor**, to auto-fetch external picklist data
- Publish **LLM mapper** to npm, and integrate directly into the DataHarmonizer web interface
- Integrate **schema catalogue** cross-cluster search directly into the DataHarmonizer, enabling users to seamlessly query and retrieve metadata schemas from distributed partner nodes

ACKNOWLEDGMENTS

CIHR IRSC Canadian Institutes of Health Research / Instituts de recherche en santé du Canada

GenomeCanada

CS-DCC Climate-Smart Data Collaboration Centre

Sample results **Schema:** canada-covid19 **Slot:** exposure_event

Input term	1st candidate	2nd candidate	3rd candidate
Social gatthering	Social Gathering	Community Event	Family Gathering
rec	Recreational Event	Community Event	Sporting Event
conference	Convention	Community Event	Concert
Family Runion	Family Gathering	Family Reunion	Community Event
Nosocomial infection	n/a	n/a	n/a

In our hybrid embeddings–LLM workflow, the embeddings stage generates consistent candidate mappings, while the LLM assigns confidence scores to each potential match. Combining embedding retrieval, with averaged scores across multiple LLM calls, produced consistent candidate rankings across runs. Batching candidate mappings and parallelizing LLM requests reduced runtime for ~1000 input terms from ~10 minutes to ~3 minutes post hoc.