Canadian Bioinformatics Hub Conference

May 28th, 2026

Where Data Meets Meaning: Ontology Integration for Pathogen-Genomics Contextual Data

Rhiannon Cameron[1], Damion Dooley[1], Emma Griffiths[1], William Hsiao[1,2]

Rhiannon Cameron: rcameron@sfu.ca
William Hsiao: wwhsiao@sfu.ca

[1] Faculty of Health Sciences, Simon Fraser University, SFU
[2] Department of Molecular Biology and Biochemistry, SFU

Center for Infectious Disease Genomics and One Health
Simon Fraser University (SFU), Burnaby, BC, Canada

The Problem:

- The SARS-CoV-2 pandemic highlighted to the public health community the need to **collect and compare viral genome contextual data**.
- **Non-standardized information systems** across institutions result in datasets that are difficult to integrate and compare, as exemplified by the challenge of data harmonization within Canada's decentralized health system.
- These variations delay data integration and make consistent analysis difficult or impossible, ultimately hindering our ability to **quantify public health threats and guide mitigation strategies**.

Contextual Data

While a sequence reveals the genetic code, contextual data reveals the circumstances; giving us the essential *who, what, why,* and *how* behind the sample.

Requires...

Harmonization

Where contextual data provides the essential story, data harmonization ensures everyone can read it; giving us the standardized fields, terms and formats required for global comparison.

Guided by...

Specifications

If harmonization brings the language, specifications provide the grammar; giving us the explicit technical framework needed to dictate exactly how data is represented and understood.

Implemented via...

Schemas

While specifications provide the grammar, schemas enforce the syntax; giving us the structured entities, properties, and constraints needed to ensure the data model is machine-actionable.

Anchored by...

Ontologies

If schemas enforce the syntax, ontologies define the meaning; a definitive encyclopedia of controlled vocabulary and logic driving computational reasoning and interoperability.

Why You Should Care:

- **Offload manual curation** to computers, allowing automated validation to handle the heavy lifting of data formatting.
- **Leverage existing data structures** created and vetted by experts, rather than building models from scratch.
- **Streamline integration** across bioinformatics ecosystems via interchange mappings to databases, tools, and alternate models.
- **Facilitate cross-jurisdictional comparisons** by ensuring disparate datasets share the semantic meaning.
- **Build scalable, community-adaptable data infrastructure** that can evolve with dynamic stakeholder needs and emerging threats.

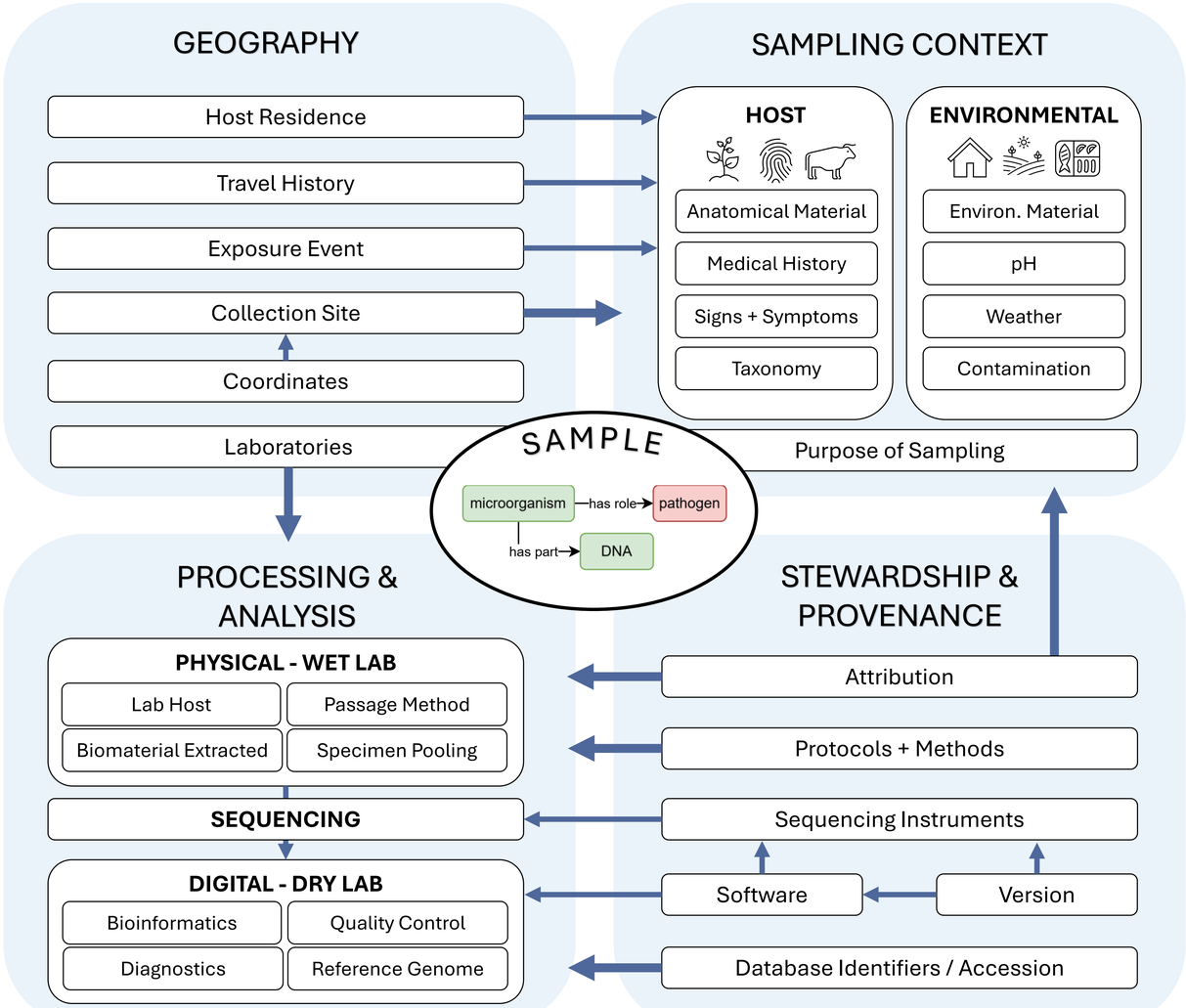

Figure 1: Sample-Centric Specification. An abstracted contextual map tracking the transformation of a physical sample from its environmental or clinical origin into a digital record.

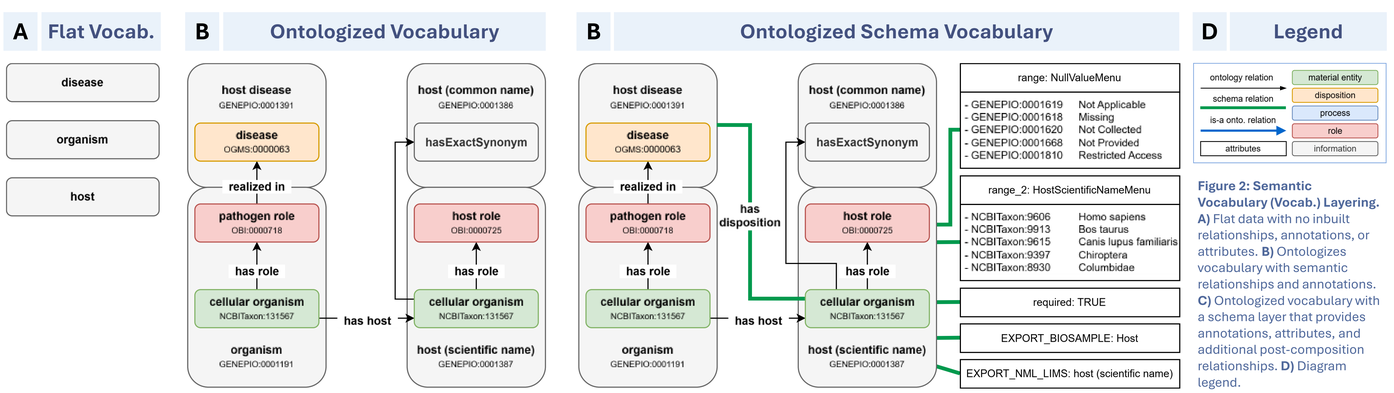

Figure 2: Semantic Vocabulary (Vocab.) Layering. **A)** Flat data with no inbuilt relationships, annotations, or attributes. **B)** Ontologizes vocabulary with semantic relationships and annotations. **C)** Ontologized vocabulary with a schema layer that provides annotations, attributes, and additional post-composition relationships. **D)** Diagram legend.

Schema vs Ontology:

Feature	Schema	Ontology
Primary Role	Defines the structure	Defines the meaning
Validation	"Does the data fit the shape?"	"Does the data make logical sense?"
Key Components	Classes, slots, enumerations, values, and data properties	Classes, instances, URIs, semantic triplets, logic heritance
Implementation	LinkML[1]	OBO Foundry[2] GENEPIO[3]
Tooling Example	DataHarmonizer Templates[4]	Knowledge Graph Database

1 Linked Modelling Language (linkml.io/linkml)
2 Open Biological and Biomedical Ontology Foundry (http://www.obofoundry.org/)
3 Genomic Epidemiology Ontology (https://genepio.org/)
4 The DataHarmonizer (https://github.com/cidgoh/DataHarmonizer)

Value of a LinkML Schema:

Ontologies can express schema-like elements but are designed for open-world semantic reasoning (deriving truths) rather than full schema definition (checking for completeness or validity). By contrast, **LinkML is a schema language** for defining structured data, validation rules, and interoperable models. Offloading schema concerns to LinkML provides several advantages:

- **Structured data modeling**; including non-relational style patterns used by traditional databases and applications.
- **Schema-driven generation**; source for generation artifacts (documentation, code, even ontology patterns).
- **Multi-target conversions**; can be converted into multiple representations (e.g., JSON, SQL, OWL, RDF, etc.).
- **Explicit validation;** offers sophisticated validation rules including patterns, ranges, and dynamic enumerations.
- **Post-composition**; enables composite assertions that go beyond ontology axioms.

Logic of an Ontology:

The Semantic Triplet (The Logic)

Figure: RDF triple. A machine-readable statement that can unambiguously reasoned over and queried. Subject/Object = Classes, Predicate = Relational Axiom.

Ontology Inheritance (The Hierarchy)

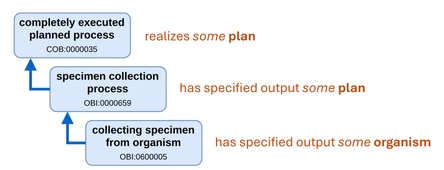

Figure: Inheritance. The "is a" hierarchy allows subclasses to automatically inherit parent axioms. For example, "collecting specimen from organism" inherits its parent's axioms, eliminating the need to recode logic across classes.

Design Principles:

- **Linked Data** (Berners-Lee, 2006; 2009); technical foundation for publishing structured, interlinked data on the Web.
- **FAIR** (Wilkinson *et al.*, 2016); maximizing the utility of research assets by making them **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable.
- **Open Biological and Biomedical Ontology** (Jackson *et al.* 2021) operational principles for biomedical ontology development, emphasizing governance, collaboration, and ontological rigor.
- **CARE** (Carroll *et al.* 2022) Indigenous data governance addressing gaps in FAIR regarding **C**ollective Benefit, **A**uthority to Control, **R**esponsibility, and **E**thics.
- **OCAP®** (FNIGC, 2014); First Nations data **O**wnership, **C**ontrol, **A**ccess, and **P**ossession.
- **Integrated Knowledge Mobilization (iKMb);** Integrated Knowledge Mobilization where knowledge users are involved from research design through dissemination, operating within CARE and OCAP® governance structures.

ACKNOWLEDGEMENTS: The authors respectfully acknowledges the unceded traditional territories of the Coast Salish peoples, including the səlilwətaɬ (**Tsleil-Waututh**), kʷikʷəƛəm (**Kwikwetlem**), Sḵwx̱wú7mesh Úxwumixw (**Squamish**) and xʷməθkʷəy̓əm (**Musqueam**) Nations, on which SFU Burnaby is located.

The authors also acknowledge our provincial and national collaborators who continue to provide valuable feedback, along with other project collaborators and alumni: **Charlie Barclay, Anoosha Sehar, Madeline Iseminger** and the Public Health Alliance for Genomic Epidemiology (**PHA4GE**).

FUNDING AGENCIES: