Putting standards into practice
Ontologizing the Canadian COVID Genomics Network specification

Rhiannon Cameron, Damion Dooley, Emma Griffiths, Sarah Savić Kallesøe, William Hsiao

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

Hsiao Public Health Bioinformatics Laboratory*
Simon Fraser University, BC, Canada

*Soon to be called
The Centre for Infectious Disease Genomics and One Health

INTRODUCTION

The Problem

- Non-standardized information systems across institutions resulting in datasets that are difficult to integrate and compare, as exemplified by the challenge of data harmonization within Canada's decentralized health system.

Canadian COVID-19 Genomics Network (CanCOGeN)

- Canada's national SARS-CoV-2 genomic surveillance response initiative aiming to sequence 150K virus genomes (VirusSeq) and 10K human for (HostSeq).
- Is developing a national framework that can be repurposed to address future outbreaks.

Contextual Data

- Data that tells the essential story surrounding the sequence data.
- Provides critical information for monitoring the origin, spread, and evolution of the SARS-CoV-2 virus.
- Informs public health decision making.

CanCOGeN Contextual Data Specification

bit.ly/3xU6hSx

- Developed collaboratively to harmonize the different provincial contextual data terms and fields before/as datasets arrive at the:
  - National Microbiology Laboratory (NML); where the data is then used to understand how COVID-19 has entered Canada,
  - Canadian VirusSeq Data Portal; a collaborative, open-access data portal where Canadian researchers can track and analyse the SARS-CoV-2 virus.
- Specialized for SARS-CoV-2 / COVID19 related questions.
- Maps to IRIDA, GISAID, BioSample, VirusSeq Portal, CNPHI Laser, and NML LIMS.
- Has been adopted and implement in Canada: CanCOGeN VirusSeq, DNAStack, National Genomic Surveillance Database; and around the world: PHA4GE, NCBI, Austrakka, COV GEN Network, ACEGID, BAOBAB LIMS, SPHERES, TOAST CDC.

DataHarmonizer

github.com/Public-Health-Bioinformatics/DataHarmonizer

- Open-source, template driven (JSON), dynamically generated, spreadsheet application.
- Browser-based application that runs offline to ensure sensitive data can be privately loaded, edited, validated, and saved.
- Customizable import/export templates.
- Utilizes the CanCOGeN specification.
- License: MIT

ONTOLOGIZING

- Ontologies provide data structure semantics with nuanced meaning relations that people comprehend, in a format that computers can logically reason over.
- Provides a controlled vocabulary (e.g., with definitions, synonyms, additional information) that is open, collaborative, and accessible.
- Permanently published vocab that will always lead replacements.

The Open Biological and Biomedical Ontology (OBO) Foundry

- Multiple domain and application ontologies; organized under a Basic Formal Ontology (BFO) to facilitate interoperability.
- **F**indable **A**ccessible **I**nteroperable **R**eusable (FAIR)

GenEpiO

genepio.org

- The Genomic Epidemiology Ontology (GenEpiO).
- Application ontology for infectious disease surveillance and outbreak investigations; houses CanCOGeN specification model:
  - ➢ **140** field specification terms.
  - ➢ **1043** picklist terms and counting (created or imported).
  - ➢ More than **550** in-development or awaiting release through collaboration with other OBOF ontologies.
- Source Code: github.com/GenEpiO/genepio
- License: CC BY 3.0

UTILIZING LinkML

Linkml.github.io

Linked Data Modeling Language

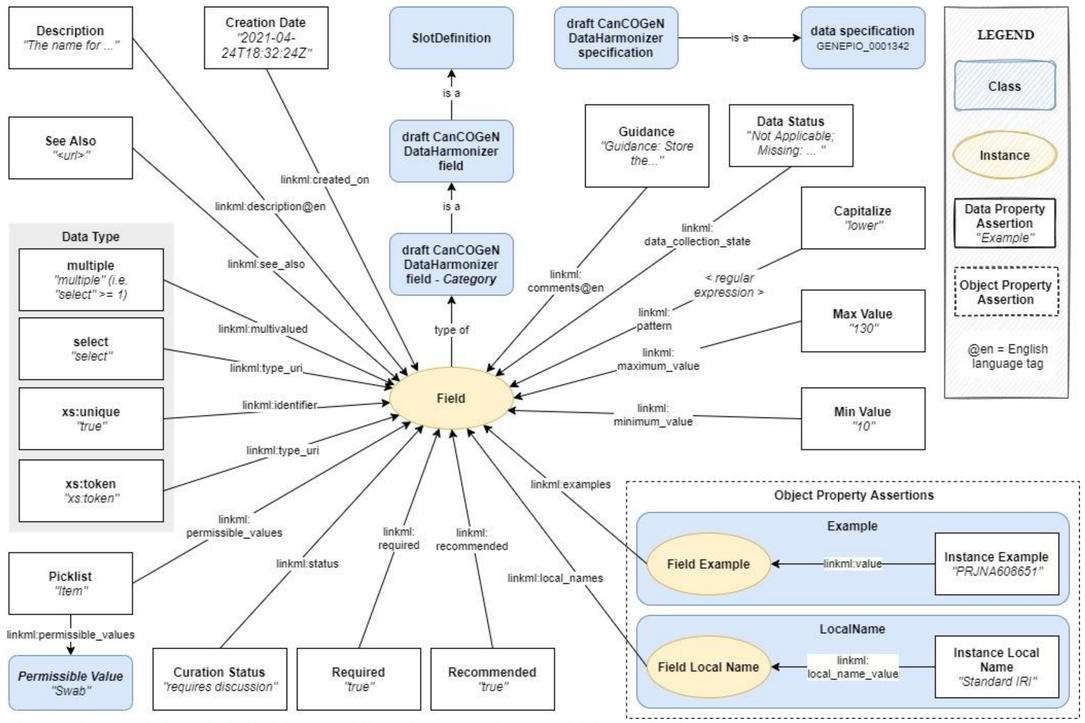
- General purpose, open-source data structure specification language that follows ontological principles (e.g., polymorphism, informative edges between graph nodes).
- Integrates semantics and data processing.

Why?

- LinkML is currently in the process of encoding other standards (e.g., GSC MIxS); by LinkML will therefore facilitate the mapping of said standards to the CanCOGeN specification.
- Facilitates information exchange between different model representations (e.g., RDF, JSON-LD, OWL, YAML...).
- Being spearheaded by several American agencies.

CONCLUSION

- Ontologizing the CanCOGeN specification will facilitate the export of comparable and interoperable datasets, converging from different sources, while providing a controlled vocabulary and improved mapping to other standards via LinkML.



Overview of CanCOGeN specification's OBO and LinkML data structures

With the specification and broad field categories being **classes** (e.g., a genome sequencer) and the individual fields being **instances**/individuals (e.g., said sequencer, but unique and existing in the lab). The **data property assertions** (e.g., short-read data) connect instances with literal data values, while **object property assertions** connect instances to other instances (e.g., the serial number) – in these cases via a class intermediaries. In this case, the instances are of the specification while utilizing the CanCOGeN DataHarmonizer application.

specimen collector sample ID	NML submitted specimen type	NML related specimen relationship type	anatomical material	anatomical part	body product
sample_01	Swab	Acute	Saliva	Oropharynx (OP)	Not Applicable
sample_02	Swab	Acute	Saliva	Oropharynx (OP)	Not Applicable
sample_03	Swab	Acute	Saliva	Nasopharynx (NP)	Not Applicable
sample_04	RNA	Follow-up	Saliva	Nasopharynx (NP)	Not Applicable
sample_05	RNA	Follow-up	Saliva	Nasopharynx (NP)	Not Applicable

GENEPIO_0001123	GENEPIO_0001204	GENEPIO_0001206	GENEPIO_0001211	GENEPIO_0001214	GENEPIO_0001216
sample_01	OBI_0002600	HP_0011009	UBERON_0001836	UBERON_0001729	GENEPIO:0001619
sample_02	OBI_0002600	HP_0011009	UBERON_0001836	UBERON_0001729	GENEPIO:0001619
sample_03	OBI_0002600	HP_0011009	UBERON_0001836	UBERON_0001728	GENEPIO:0001619
sample_04	OBI_0000880	EFO_0009642	UBERON_0001836	UBERON_0001728	GENEPIO:0001619
sample_05	OBI_0000880	EFO_0009642	UBERON_0001836	UBERON_0001728	GENEPIO:0001619

Tabular OWL export

Example of export feature, under development for the CanCOGeN DataHarmonizer metadata tool, that will enable the export of curated data that corresponds to the specification permissible values into Ontology IDs. All Ontology IDs have permanent URLs (PURLs) which are accessible online and contain additionally information (e.g., definition, synonyms); avoiding semantic ambiguity and information mismatching by providing a controlled vocabulary to users.

Acknowledgements: We would like to acknowledge our provincial and national collaborators who continue to provide valuable feedback, our funding agencies who we couldn't have done this without, along with other project collaborators and alumni:

Ivan Gill, Gurinder Gosal, Anoosha Sehar, Lauren Tindale, Matthew Croxen, David Alexander, and the **Public Health Alliance for Genomic Epidemiology (PHAGE).**

Public Health Alliance for Genomic Epidemiology

Funding Agencies:
National Collaborating Centre for Infectious Diseases
Public Health Agency of Canada
Agence de la santé publique du Canada
CanCOGeN
VirusSeq
GenomeCanada
20YEARS GenomeBritishColumbia
Michael Smith Foundation for Health Research
CIHR IRSC
NSERC CRSNG