Towards interoperable and FAIR water-based genomic surveillance: An ontology driven contextual data specification for environmental genomics.

Charlie Barclay[1], Rhiannon Cameron[1], Nima Tehrani[2], Baofeng Jia[2], Erin EE Gill[2], Justin Richardsson[3], William WL Hsiao[1], Fiona SL Brinkman[2], Mélanie Courtot[3,4,5], Emma Griffiths[1]

[1]Centre for Infectious Disease Genomics and One Health, Faculty of Health Sciences, Simon Fraser University; [2] Brinkman Lab, Department of Molecular Biology and Biochemistry, Simon Fraser University; [3] Ontario Institue for Cancer Research, Toronto; [4] Medical Biophysics Department, University of Toronto; [5] Department of Computer Science, University of Toronto

Contact: charlie_barclay@sfu.ca or ega12@sfu.ca

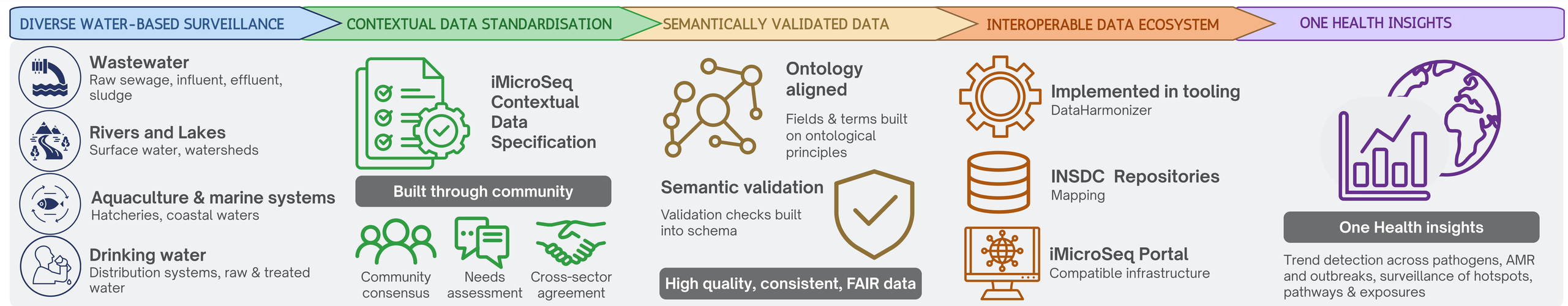


Figure 1: A semantic framework for interoperable water-based microbial genomics. The iMicroSeq contextual data specification harmonizes metadata across environmental genomics workflows using ontology-driven semantic modelling & interoperable tooling to support standardized validation, exchange, reuse, & One Health insights

WHY CONTEXTUAL DATA MATTERS

Genomic sequences require context for interpretation.

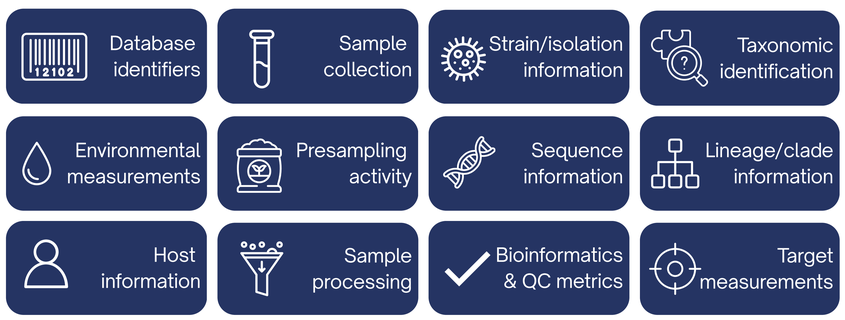
Water-based genomic surveillance generates diverse environmental, laboratory, sequencing, and analytical contextual data that are often difficult to integrate across systems and sectors. Inconsistent terminology and reporting practices limit interoperability, reproducibility, and large-scale data reuse. The iMicroSeq specification harmonizes contextual data to support **F**indable, **A**ccessible, **I**nteroperable and **R**eusable (FAIR) genomic surveillance workflows.

Shared semantics are required for meaningful integration across systems and sectors.

WHAT IS IN THE SPECIFICATION

Modular ontology-based contextual data architecture

155 fields and >900 controlled vocabulary terms across 12 modules



ONTOLOGY DRIVEN SEMANTIC MODELLING

Contextual data are linked entities and processes, not isolated spreadsheet columns, supporting machine-readability, semantic querying and FAIR data exchange.

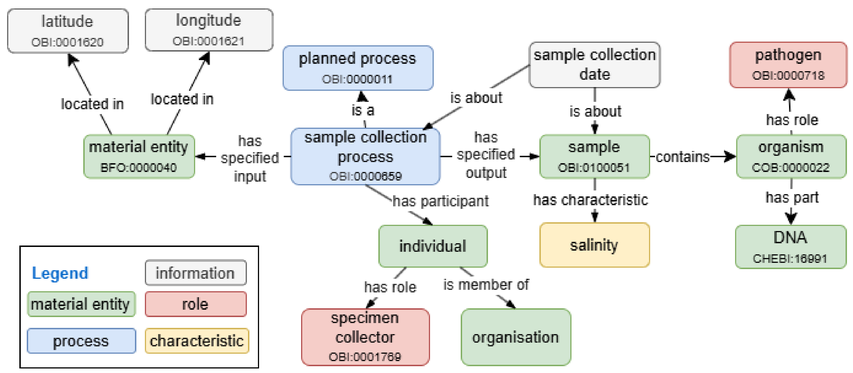


Figure 2. Ontology-driven semantic model for sample collection provenance. Semantic relationships between samples, collection processes, metadata, individuals, and organizations support interoperable and FAIR environmental genomics workflows.

SPECIFICATION DEVELOPMENT

Community driven semantic standards development

The specification was developed through user engagement and needs assessments, review of existing standards, and cross domain gap analyses.

Built using OBO Foundry ontologies and semantic modelling; the framework extends the international **PHA4GE wastewater standard** to support marine, freshwater, drinking water, and built environment surveillance contexts.

IMPLEMENTATION ARCHITECTURE

Ontology driven semantics operationalised through LinkML

The specification is implemented in **LinkML** and operationalized through **DataHarmonizer** templates, to support community uptake and interoperable data exchange. Ontologies provide semantic relationships between entities and processes, while LinkML operationalises these semantics into **validated**, **machine-readable schemas** supporting structured data entry, repository mappings, interoperable exchange formats, and downstream analytical pipelines.

The framework functions as the structural backbone of the iMicroSeq Data Portal (https://imicroseq-dataportal.ca).

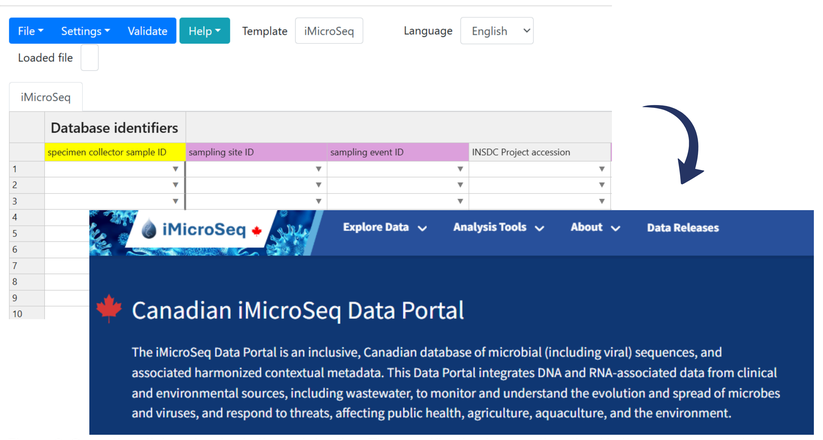


Figure 3. Operationalising standards through tooling. The specification is operationalised in the spreadsheet style validation applincation the DataHarmonizer with automated exports to downstream repositories including the iMicroSeq Portal.

IMPACT:

✓ Harmonized contextual data enabling cross domain water-based genomic surveillance
✓ Ontology-driven semantic modelling supporting **interoperable** and **machine-readable** data exchange
✓ Structured validation and **automated transformation** into repositories and analytical workflows
✓ Direct implementation in the **iMicroSeq Data Portal**

The authors respectfully acknowledges the unceded traditional territories of the Coast Salish peoples, including the səlilwətaɬ (Tsleil-Waututh), kʷikʷəƛ̓əm (Kwikwetlem), Sḵwx̱wú7mesh Úxwumixw (Squamish) and xʷməθkʷəy̓əm (Musqueam) Nations, on which Simon Fraser University Burnaby is located.

With thanks to all the contributors to the iMicroSeq contextual data specification and wider iMicroSeq project, in particular the Ontario Institute for Cancer Research and Public Health Agency of Canada as well as our funders/supporters Genome BC, Genome Canada, Koonkie, CANUE, McMaster University, Simon Fraser University, McGill University, DNAstack, the Gates Foundation and the Canadian Institutes of Health Research