# 10 Simple Rules for Developing Ontology-based Data Standards for Pathogen Genomics Contextual Data

## Authors

Emma J. Griffiths[1]*, Charlotte P. Barclay[1]*, Rhiannon Cameron[1], Damion Dooley[1], Nithu Sara John[1a], Anoosha Sehar[1], Ivan Gill[1], Emily Havervold[2], Natalie Knox[2, 3], Andrea D. Tyler[2], Ana T. Duggan[2, 4], Levon Kearney[2], Christopher Townend[2], Suzanne Darling[2], Amber Desrochers[2], Catherine Yoshida[2], Gordon Jolly[2], Morag Graham[2, 3], Gary Van Domselaar[2, 3], John Nash[2], Emil Jurga[2], Eduardo Taboada[2], Gabriel Wajnberg[5], Julie A. Shay[5, 10], Andrew Scott[6], Kirsten Palmier[2], Molly Pratt[2], Jeff Tuff[2], Brian P. Alcock[12], Ed Topp[7] , Lisa A. Johnson[8], James Robertson[2], Justin Schonfeld[2], D. Patrick Bastedo[2], Derek D. N. Smith[9], Jordyn Broadbent[9], Dominic Poulin-Laprade[6], Oliver Lung[5], Sandeep Tamber[10], Finlay Maguire[11], Andrew G. McArthur[12], Richard Reid-Smith[2], Rahat Zaheer[6], Chad R. Laing[5], Catherine D. Carrillo[5], William W.L. Hsiao[1]

## Affiliations

[1] Centre for Infectious Disease Genomics and One Health, Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada
[2] National Microbiology Laboratory, Public Health Agency of Canada, Canada
[3] Department of Medical Microbiology and Infectious Diseases, University of Manitoba, Canada
[4] McMaster Ancient DNA Centre, Department of Anthropology and Department of Biochemistry & Biomedical Sciences, McMaster University
[5] Canadian Food Inspection Agency, Canada
[6] Agriculture and Agri-Food Canada, Canada
[7] Institut National de la Recherche Agronomique (INRAE), Dijon, France
[8] Fisheries and Oceans Canada, Canada
[9] Environment and Climate Change Canada, Canada
[10] Food and Nutrition Directorate, Health Canada, Canada
[11] Dalhousie University, Halifax, NS, Canada
[12] Michael G. DeGroote Institute for Infectious Disease Research and Department of Biochemistry & Biomedical Sciences, McMaster University, Hamilton, ON, Canada
[a] currently at EMBL-EBI, Hinxton, UK
*These authors contributed equally to this work

Corresponding authors: Emma Griffiths (ega12@sfu.ca) and William Hsiao (wwhsiao@sfu.ca)

Keywords: data harmonization, contextual data, pathogen genomics, ontologies, data standards, development methods

# Introduction

Pathogen genomics applications are powerful tools for tracking, characterizing, and controlling infectious disease in clinical and public health programs including One Health surveillance, environmental monitoring, outbreak investigation, and food safety programs. However, interpreting sequence data for pathogen genomics applications requires high quality contextual data. Without context, a sequence is a puzzle piece without the picture on the box — its role is unclear. Contextual data provides that picture by capturing details such as where and when a sample was collected, test results, environmental conditions, and clinical or epidemiological information. This often comes from multiple sources including hospitals and clinics, field teams, epidemiologists and laboratories, each with their own priorities, processes and systems. Consequently, resulting data often has inconsistent formats, structures, and levels of granularity, impeding integration and sharing without prior resource intensive harmonization.

To address these challenges, data specifications - recommended controlled vocabularies, formats, and rules for structuring, storing, transferring, and expressing information - are developed to support particular use cases (e.g., pathogen-specific, target-specific, organization-specific). When broadly adopted by a community, these specifications become data standards, though their implementation can vary. Yet, there is a lack of widely adopted reusable standards for pathogen-genomics contextual data and defining effective interoperable standards remains difficult, especially considering local-to-global needs, without clear methodology.

For over a decade, our team at the Centre for Infectious Disease Genomics and One Health (CIDGOH) has partnered with Canadian federal and provincial agencies to develop and implement ontology-based data standards for public health, food safety, and One Health pathogen genomics, forming the basis for international standards. These standards are based on technical best practices including sourcing terminology from open-source, version-controlled, international semantic resources and structuring specifications with machine-readable modelling languages. Data is “human-readable” when formatted conveniently for human interpretation and “machine-readable” when structured in a way that it can be automatically read and processed by a computer. By linking terminology to ontology-based annotations we avoid organization-specific jargon, enable reuse across domains, and ensure transparent vocabularies, accessible via lookup services. The hierarchy and relationships in these ontologies support standardized classifications and the construction of knowledge graphs for advanced queries.

Here, we describe a 10 step data standards development methodology that can be put into practice by the broader community ~~to make standards~~, ensuring data is structured, consistent and interoperable (Figure 1).

## 1. Identify the need for a new specification (and make sure you have enough money)

Different specifications address different data needs, but new needs in projects or programs don't always require creating a new specification. Some specifications can simply be updated to ensure they are fit-for-purpose, while other situations require a new standard. Before reinventing the wheel, it is worth seeing what standards already exist. Good resources for finding and exploring genomics standards include the Genomics Standards Consortium [1] (https://www.gensc.org), the Public Health Alliance for Genomic Epidemiology GitHub organization (PHA4GE; https://github.com/pha4ge), and the list of International Nucleotide Sequence Database Collaboration (INSDC) attributes webpage (https://www.ncbi.nlm.nih.gov/biosample/docs/attributes).

Critical to building new specifications is sufficient funding. THEY COST MORE THAN YOU THINK. While there is an increasing appetite for the use of data standards in data management systems, their development and implementation are often not accounted for in project planning. Depending on the complexity of the data that needs to be harmonized, costs can vary - with an estimated average of $100-150K CAD (~100K USD) to cover all the work. Additional resources may be required for various implementations and integration into tools and systems. Data standard development checklists should include: scoping, assessing data needs and reviewing existing resources, schema development, mapping vocabularies, building tools, supporting materials for operationalizing standards, testing, and long-term maintenance. All of these various activities take time, resulting in a timeline for standards development that can take several months to years. Funding for standards development and implementation should be built into experimental design and budgets.

## 2. Define the goals and scope of the specification (aspire to having boundaries!)

Once the need for a new specification has been identified, the purpose and objectives should be clearly articulated. As with any project, data specifications can be subject to project creep, resulting in ambiguity and misuse. The role of the specification - standalone or as part of a broader ecosystem - must be clear from the outset. A good specification takes into account the full life cycle of the data, from collection, to reporting to storage in downstream databases or repositories. When part of a larger ecosystem, the specification should be broad enough to accommodate the diversity of data, while remaining focused on the primary purpose. A well-defined scope includes a primary target audience, use cases, and the data types required to support them. Consider the benefits of standardisation for anticipated usage, prioritizing structured formats and controlled vocabularies that balance scientific necessity with practical feasibility. Align the scope of the specification with the level of development effort required to implement and maintain it. Make sure boundaries are well defined and tangible outputs clearly documented.

## 3. Perform a data needs assessment (talk to people who know all the things)

A data needs assessment involves identifying gaps in existing standards and, most importantly, engaging with those collecting the data. It should focus on the requirements to achieve specific data goals, and balance usability and interoperability with real-world data generation, sharing, and access. Consulting stakeholders involved across the data life cycle helps to distinguish essential elements from those that are “nice to have”, ensuring that the specification is fit for purpose while preventing undue burden on data providers and preserving auditability and practicality.

When performing a data needs assessment, choose feedback mechanisms carefully. By providing only structured responses, surveys often miss the nuance and critical insights of diverse data needs that emerge through discussion. In contrast, consultations like interviews or workshops facilitate dynamic conversations, uncovering deeper insight and real-world needs. They help reveal regulatory and data sharing requirements that could influence design and uptake. Equally important, consulting representatives from all communities involved helps address biased data that reinforces stereotypes around gender, ethnicity, and human behaviour.

When defining data requirements, existing data standards or attributes packages can be a valuable starting point, but their use should be carefully considered in context. Standards like ISO 23418, MIxS checklists, and INSDC repository attribute picklists (e.g., NCBI BioSamples, ENA Samples; [2,3]), can provide valuable references, however unclear definitions and mixed concept picklists limit consistent use and universal applicability. An alternative is to review existing databases with similar data; however, they may have built-in conventions that can limit integration and reuse across other datasets or tools. Extracted fields and picklists should therefore undergo structural evaluations. Ultimately, the needs assessment should clarify not only what data is required and available, but also where important elements are missing (i.e., a gap analysis), providing a foundation for schema development.

## 4. Draft a preliminary ontology-based specification schema using a consensus approach (Rule 4 has a lot of other rules)

Across public health, data is often found in heterogeneous formats, reflecting the diversity of studies and investigations [4,5]. A good schema should define key elements by using trusted ontologies to ensure consistency and interoperability. Ontologies are formal descriptions of a knowledge domain, typically functioning as controlled vocabularies that define relationships between concepts (classes), like rulebooks that help turn scattered data into a coherent blueprint (Figure 2). In genomics, they play a crucial role in linking and interpreting data across systems, improving data extraction, retrieval, and connection [6].

Fields and picklist terms should be mapped directly to ontological classes using an Internationalized Resource Identifier (IRI). IRIs disambiguate meanings, supporting machine-readability and linkage to a broader knowledge framework. Consider the completeness and application when selecting ontology classes. Ontologies are always evolving and rarely complete; but if major knowledge gaps exist in an ontology, broad community adoption is unlikely [7]. The Open Biological and Biomedical Ontology (OBO) Foundry is a community led group that develops ontologies relevant to biological research and is a useful foundation to ensure the standard is compatible with the wider ecosystem [8,9].

When new classes are required, following community design patterns ensures consistency. OBO Foundry ontology-based annotations avoid local jargon, facilitating reuse. Classes should have unique, unambiguous, unabbreviated labels; clear, distinct Aristotelian definitions ('An **A** is a **B** differentiated by **C**'); and relationships that reflect the definition content. Logical relationships support standardized classification, knowledge graphs, natural language tools, and automated reasoning without extra interpretation. Standardized terms should be grouped into thematic modules for easy adaptation, ensuring an extensible framework that allows new fields or modules to be added as collection methods and sequencing technologies evolve.

## 5. Develop interchange mappings and tools (there will never be "one standard to rule them all")

Different databases and resources such as analytical platforms, often use varied data structures developed over many years. Ideally, the data needs assessment should identify these downstream destinations and developing interchange mappings is crucial to ensure data can be integrated across systems. Interchange mappings encode the information that transforms the source schema into the target schema. Manual transformations of data into other structures/formats should be avoided as they can be resource intensive and error-prone, ultimately decreasing uptake. Mappings should be included in support materials, while exchange formatting and import/export mapping should be automated into tooling.

Building on ontological foundations, specifications should be provided in machine-readable formats. We recommend the Linked Data Modeling Language (LinkML) framework for defining data models that are both human- and machine-readable, allowing for rich semantic modeling, validation, and easy interoperable format exchange (e.g. JSON, YAML, RDF and OWL) [10]. LinkML has inbuilt linkage to ontology terms, supports regular expression data transformation, and auto-generates documentation. Expressing your standard in LinkML turns a static blueprint into something dynamic. With ontologies acting as rules, standards as the blueprints, and tools as implementation devices, LinkML facilitates adoption by supporting integration and reuse of standards in other systems.

## 6. Develop materials to support operationalizing the specification (by the way, people won't read the manual)

A specification is only valuable if  accessible, easy to adopt, and well-implemented. Operationalising through tooling helps ensure consistent data capture, validation, and sharing. Tools should be simple, portable, and user-friendly, while fitting the workflows and technical capacity of the target community. For instance, several CIDGOH specifications have been implemented in the DataHarmonizer, a template-driven application that provides built-in data validation and transformation while maintaining a familiar interface for users [11].

In addition, clear documentation, such as reference guides and standard operating procedures (SOPs), streamlines use and lowers uptake barriers. At a minimum, SOPs should be provided for curation, tooling, and new term requests. Just as a picture is worth a thousand words, worked examples (using synthetic or open access data) beat pages of documentation by clearly demonstrating specification standardization and harmonization *in practice*.

Where possible, provide skills training to support effective implementation, considering different learning styles and ongoing engagement. Different training formats support complementary implementations - in depth workshops build deep understanding, short videos provide quick overviews, and written guides serve as ongoing reference materials. This ensures users can apply the specification accurately and confidently, improving data quality while offering an additional channel for feedback.

## 7. Test the specification and gather feedback (make sure the thing works!)

To encourage widespread adoption, the specification needs to be applicable for a variety of use cases. Validate its consistency and usability by applying the schema to existing datasets. In addition to providing a starting point for fields and terms, reviewing existing datasets (such as those in INSDC repositories) can act as a useful indicator of viability and identify descriptors and information types that may be missing.

Additionally, testing through pilot projects representing real-world scenarios is a critical step for validating the practical applicability of a specification and collecting invaluable user feedback to drive meaningful improvements. We recommend engaging underrepresented communities and groups with a diversity of knowledge and data to adequately test effectiveness. Funding incentives (e.g., stipends or subgrants) enable engagement of participants in remote and low resource settings, which is critical for ethical and democratized standards development. Other incentives can include authorship on manuscripts, as well as having a voice and documented role in standards development. As part of a pilot program, training and orientation sessions can help prepare participants, and should be followed by a clear and defined testing period during which users apply the

specification to their own data. This approach fosters global input, supports inclusivity, and generates real-world insights to refine the specification.

## 8. Document real-world implementations (who's actually using it?)

Arguably, the largest challenge for measuring the impact of any data specification is assessing community adoption. To provide evidence-based uptake, gather examples of real-world implementations. Implementations can involve whole specifications or their parts, and can include data sharing agreements (e.g., memoranda of understanding), data management systems (e.g., LIMS), repository schemas and submission requirements, data collection instruments (e.g., spreadsheets), analytical and visualization software (e.g., bioinformatics dashboards), and publications/reports. Seeing how specifications are put into practice for different purposes and in different contexts helps laboratories use them effectively and model success. Communicating about the impacts of standards is also helpful for creating community norms, sharing lessons learned, and setting expectations for quality management.

## 9. Create a maintenance plan for the specification (things change - keep up)

Sustaining a standard over time depends on a coordinated investment in both technical infrastructure and community support. A data specification is not static, it must reflect the current state of a scientific field and evolve as technology, accreditation, regulations, operations and research priorities shift. This evolution requires a clear governance model for revisions, with a defined process for requesting changes, that encourages open contribution, ensures traceability, and maintains alignment with the scope. Identify the individuals or organisations that will be responsible for maintaining the specification, as well as any potential hardware or software integrations.

Hosting the specification on GitHub supports transparent governance, collaboration, and version control -fostering Open Science, building trust, and supporting adoption. Versioning is an important facet of sustainability and reproducibility, offering a clear record of changes, improvements, and deprecated elements. Using semantic versioning (e.g., x.y.z) helps users track structural changes to new fields (x), values and ID (y), and formatting, descriptions, or examples (z). As adoption grows, communities will request new terms and fields, or it may become apparent that further supporting documentation is required. GitHub provides a transparent record of submissions and issues, showing where changes diverge from requests while supporting accountability and engagement. Using standardised templates for new term and field requests helps ensure necessary information is captured for effective review and execution.

### 10. Encourage community participation in standards development and implementation (sharing is caring)

Data standards are successful only if they prove useful and are effectively adopted. As such, there needs to be an interplay between the users, standards development, and how they are implemented (Figure 3). Mechanisms for community participation can include a variety of activities such as membership in standards development working groups, engaging ontology communities, or simply posting questions and comments in GitHub Issue trackers. Encouraging participation is a vital step in development and should be considered during scoping.

Diverse community input from data providers and users (e.g. bioinformaticians, clinicians, epidemiologists, lab technicians, analysts, decision-makers, etc.) is critical to ensuring fit-for-purpose. Participating in the development process can raise awareness of the utility of standards, supporting their routine integration into programs and initiatives. There are many different roles in standards development, including participation in technical aspects, as domain experts, and also in advocacy by building consensus and socialization. Acknowledging these different roles in manuscripts, grant applications, and other benefit sharing activities - incentivizes wider community participation.

### Summary

While data standards are never “internet sensations”, they are critical components of data quality management systems and are worth the investment of time, money, and community capital. The methodology described in this work can help streamline the development of interoperable standards for pathogen genomics activities that help protect lives and economies.

## Acknowledgements

The authors wish to thank all of our community partners who helped to develop, test, and implement ontology-based standards for pathogen genomics. This work was critical to developing the methodology described in this manuscript.

## Figure and Table Legends

Figure 1: Overview of the ten step method for ontology-based pathogen genomics standards development. The 10 steps include 1) Identify the need for a new specification, 2) Define the specification’s goal(s) and scope, 3) Perform a data needs assessment, 4) Develop a draft schema, 5) Develop interchange mappings, 6) Develop supporting materials, 7) Test the specification package, 8) Document real-world implementations, 9) Maintain the specification over time, and 10) Encourage community participation.

Figure 2: Ontologies are rule books that define terms and relationships. Data standards are blueprints built from subsets of these rules, guiding what data to collect and how. Tools are the machines that implement standards—supporting data capture, validation, sharing, and analysis. Standards draw from one or more ontologies while tools may use full standards or selected parts.

Figure 3: Diagram summarizing the interaction between standards, the community of users, and implementation tools. Each component influences each other. Standards are influenced by community needs; members within the community use the standard thereby creating norms and streamlining data collection and exchange. Tools implement standards, while standards capture methods and provenance of tools. Standards-based tooling can affect practices and workflows, while feedback about tooling from users can improve functionality and ease of use.

Table 1: Case studies describing how the 10 steps were used in developing a SARS-CoV-2 specification to support pandemic genomic surveillance and responses, as well as a One Health Antimicrobial Resistance (AMR) specification to support genomics-based AMR detection and monitoring in the Canadian environment and food system.

# References


1. Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, Garrity GM, et al. The Genomic Standards Consortium. PLoS Biol. 2011;9: e1001088.

2. Courtot M, Cherubin L, Faulconbridge A, Vaughan D, Green M, Richardson D, et al. BioSamples database: an updated sample metadata hub. Nucleic Acids Res. 2019;47: D1172–D1178.

3. O'Cathail C, Ahamed A, Burgin J, Cummins C, Devaraj R, Gueye K, et al. The European Nucleotide Archive in 2024. Nucleic Acids Res. 2025;53: D49–D55.

4. Schulz S, Stegwee R, Chronaki C. Standards in Healthcare Data. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of Clinical Data Science. Cham (CH): Springer; 2018.

5. Rousseau JF, Oliveira E, Tierney WM, Khurshid A. Methods for development and application of data standards in an ontology-driven information model for measuring, managing, and computing social determinants of health for individuals, households, and communities evaluated through an example of asthma. J Biomed Inform. 2022;136: 104241.

6. Blake JA, Bult CJ. Beyond the data deluge: data integration and bio-ontologies. J Biomed Inform. 2006;39: 314–320.

7. Malone J, Stevens R, Jupp S, Hancocks T, Parkinson H, Brooksbank C. Ten Simple Rules for Selecting a Bio-ontology. PLoS Comput Biol. 2016;12: e1004743.

8. Duncan WD, Diller M, Dooley D, Hogan WR, Beverley J. Concretizing plan

specifications as realizables within the OBO foundry. J Biomed Semantics. 2024;15: 15.

9. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25: 1251–1255.

10. Moxon S, Solbrig H, Unni DR, Jiao D, Bruskiewich R, Balhoff J, et al. The linked data modeling language (LinkML): A general-purpose data modeling framework grounded in machine-readable semantics. ICBO. 2021; 148–151.

11. Gill IS, Griffiths EJ, Dooley D, Cameron R, Savić Kallesøe S, John NS, et al. The DataHarmonizer: a tool for faster data harmonization, validation, aggregation and analysis of pathogen genomics contextual information. Microb Genom. 2023;9. doi:10.1099/mgen.0.000908

# Figures and Tables

Figure 1

Figure 2

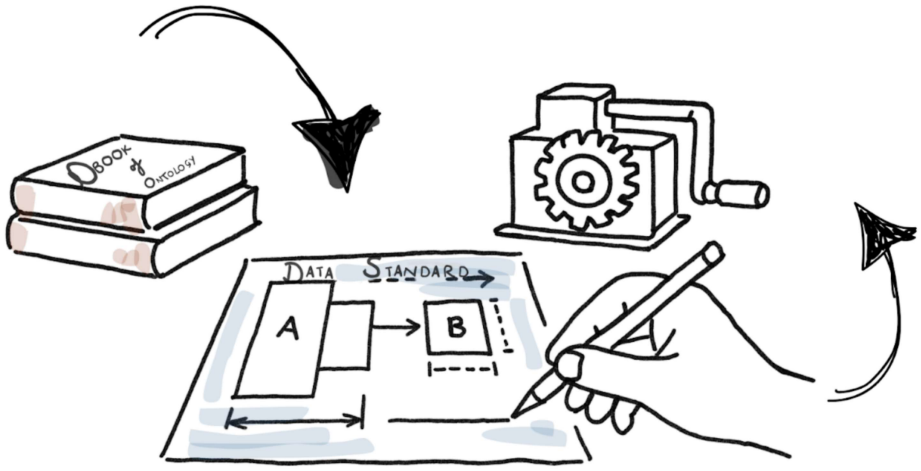

Figure 3

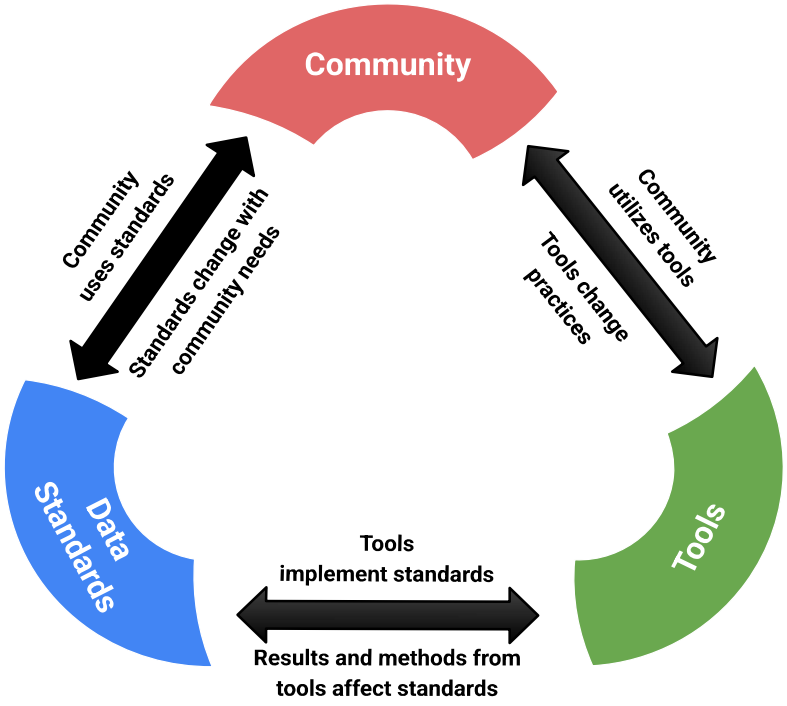
Community
Community uses standards
Standards change with community needs
Community utilizes tools
Tools change practices
Data Standards
Tools
Tools implement standards
Results and methods from tools affect standards

Table 1

| **Specification** | **SARS-CoV-2** | **One Health AMR** |
|---|---|---|
| **Specification Overview** | A data specification for harmonizing One Health Canadian COVID-19 pathogen genomics contextual data. The specification was designed to enhance the sharing and interoperability of genomic and clinical data, and to support surveillance by the Canadian COVID-19 Genomics Network during the SARS-CoV-2 pandemic. | A data specification for harmonizing One Health AMR pathogen genomics contextual data. Developed to support the GRDI-AMR and GRDI-AMR-OH initiatives, as part of the Canadian Federal Action Plan for AMR and Use in Canada, and the Pan-Canadian Action Plan on AMR. |
| **Step 1:** *Identify the need for a new specification* | At the beginning of the COVID-19 pandemic, pathogen genomic surveillance was identified as a critical tool. Data sharing across Canada was necessary to identify incursions, monitor viral evolution, and understand community health impacts. No standard for SARS-CoV-2 existed. | Previous data integration efforts across sectors, laboratories and domains were challenging due to heterogeneity of contextual data. No standard existed that addressed the breadth of data diversity. Attempts to use Biosample packages for data harmonization had failed in pilot tests. |
| **Step 2:** *Define the goal and scope* | Pathogen target: SARS-CoV-2.<br>Data types: sample metadata, epidemiological data, clinical information, diagnostic results, methods (e.g. sample processing, sequencing, bioinformatics, sampling strategies, etc.), provenance details. | Genetic target(s): AMR determinants (mutations, genes, mobile elements, etc.)<br>Pathogen target: many<br>Data types: sample metadata (hosts, environments, materials, food, etc.), AMR phenotypic testing results, methods (e.g. sample processing, sequencing, bioinformatics, sampling strategies, etc.), risk assessment metrics, experimental conditions, provenance details. |
| **Step 3:** *Perform a data needs assessment* | **Stakeholders:** provincial/territorial/federal laboratories; epidemiologists, clinicians, bioinformaticians, analysts, decision makers.<br>**Resources:** sample datasets, case report forms, public | **Stakeholders:**<br>Federal laboratories (public health, human health, animal health, agriculture, food regulation, environmental monitoring); researchers, risk assessors, analysts, decision |

| | | |
|---|---|---|
| | repositories requirements (e.g. INSDC Biosample packages, GISAID), National Microbiology Laboratory (NML) LIMS, literature, nomenclature systems (ICTV, Pango lineages, Nextstrain clades), OBO Foundry ontologies, MIxS checklists, resources created by other standards bodies. | makers.<br>**Resources:** sample datasets, literature, public repositories requirements (e.g. INSDC AST and Biosample packages), ISO 23418, OBO Foundry ontologies (e.g. ARO), MIxS checklists, resources created by other standards bodies (e.g. EUCAST, CLSI). |
| **Step 4:** *Develop a draft schema* | **Specification:** https://github.com/cidgoh/CanCOGeN_Contextual_Data_Specification<br>**Machine readable templates (yaml/.json/.tsv):** https://github.com/cidgoh/pathogen-genomics-package/tree/main/templates | **Specification:** https://github.com/cidgoh/GRDI_AMR_One_Health<br>**Machine readable templates (yaml/.json/.tsv):** https://github.com/cidgoh/DataHarmonizer/tree/master/web/templates/grdi |
| **Step 5:** *Develop supporting materials* | **Tooling:** DataHarmonizer (https://github.com/cidgoh/pathogen-genomics-package/releases)<br>**Reference Guides:**https://github.com/cidgoh/CanCOGeN_Contextual_Data_Specification/tree/main/Reference%20Guide<br>**Curation Protocol & New Term Request Template:** https://github.com/cidgoh/CanCOGeN_Contextual_Data_Specification/tree/main/SOPs | **Tooling:** DataHarmonizer (https://github.com/cidgoh/pathogen-genomics-package/releases)<br>**Reference Guides:** https://github.com/cidgoh/GRDI_AMR_One_Health/tree/main/Reference%20Guide<br>**Curation Protocol & New Term Request Template:** https://github.com/cidgoh/GRDI_AMR_One_Health/tree/main/SOPs |
| **Step 6:** *Develop interchange formats* | Mapping and interchange formats available in the DataHarmonizer as exports (NCBI SARS-CoV-2 Biosample, NCBI, GISAID, NML LIMS, VirusSeq Data Portal). | Mapping and interchange formats available in the DataHarmonizer as exports (NCBI AST, NCBI One Health Biosample, NCBI SRA, DEXA). |
| **Step 7:** *Testing* | Extensive testing by Canadian Public Health Laboratory Network (CPHLN)<br>members at the beginning of the pandemic before integration into routine data sharing workflows between jurisdictions. | Three curation pilot tests (2018, 2020, 2023) involving participants from different labs and agencies. Iterative improvements based on evolving needs and user feedback. |
| **Step 8:** *Document real-world implementations* | Implemented by the CPHLN (provincial and federal partners). https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000908 | Implemented across the Canadian federal genomics ecosystem.<br>Also implemented in Canada's Virtual Microbial Resource. |

| | | |
|---|---|---|
| | | https://cdnsciencepub.com/doi/10.1139/cjm-2024-0203; https://osf.io/preprints/osf/xbf4t_v1 |
| **Step 9:** *Maintain and evolve* | **Current version:** 3.0.0<br>Version notes describe many updates based on user requests and changing data needs. | **Current version:** 14.5.5<br>Version notes describe many updates based on user requests and changing data needs. |
| **Step 10:** *Encourage community participation* | Ongoing requests transparently shared and monitored on GitHub. Academic representatives attend CPHLN meetings, and meet regularly with NML (national reference lab) partners to discuss new data needs. Methodology also implemented by the PHA4GE (https://github.com/pha4ge/SARS-CoV-2-Contextual-Data-Specification). | Ongoing requests transparently shared and monitored on GitHub. Academic representatives meet regularly with federal scientists to discuss new data needs. Specification also being used in other initiatives e.g. One Health surveillance in Uganda; B2B2B JPIAMR initiative (https://github.com/cidgoh/B2B2B2_Contextual_Data_Specification) |