The broom of the system: a harmonized contextual data specification for One Health AMR pathogen genomic surveillance

Authors:

Emma J. Griffiths[1], Julie A, Shay[2,3], Rhiannon Cameron[1], Charlie Barclay[1], Anoosha Sehar[1], Damion Dooley[1], Nithu Sara John[1*], Andrew Scott[4], Gabriel Wajnberg[3], Emil Jurga[5], Lisa A. Johnson[6], James Robertson[5], Justin Schonfeld[5], D. Patrick Bastedo[5], Joshua Tang[4], Xianhua Yin[4], Attiq Rehman[4&], Rhiannon Wallace[4], Cheyenne Sargeant[4], Shannon H.C. Eagle[5], Tim McAllister[4], Moussa S. Diarra[4], John H.E. Nash[5], Ed Topp[8], Bryan A. Wee[9], Adrian Muwonge[9], Leonid Chindelevitch[10], Gary Van Domselaar[5], Eduardo N. Taboada[5], Sandeep Tamber[2], Tony Kess[6], Jordyn Broadbent[7], Dominic Poulin-Laprade[4], Derek D. N. Smith[7], Richard Reid-Smith[5], Rahat Zaheer[4], Chad Laing[3], Catherine D. Carrillo[3], William W.L. Hsiao[1]

Affiliations:

[1] Centre for Infectious Disease Genomics and One Health, Faculty of Health Sciences, Simon Fraser University, Burnaby, Canada
[2] Health Canada, Ottawa, Canada
[3] National Centre for Animal Disease, Canadian Food Inspection Agency, Lethbridge, Canada
[4] Agriculture and Agri-Food Canada, Canada
[5] Public Health Agency of Canada, Canada
[6] Fisheries and Oceans Canada, Canada
[7] Environment and Climate Change Canada, Canada
[8] Institut National de la Recherche Agronomique (INRAE), Dijon, France
[9] The Digital One Health Laboratory, The Roslin Institute, University of Edinburgh, Edinburgh, UK
[10] MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, London, UK
*currently European Bioinformatics Institute, Hinxton, UK
&currently Director of Bioscience, Research and Productivity Council, Fredericton NB Canada

Corresponding Author: ega12@sfu.ca (Emma Griffiths)

Keywords

One Health, antimicrobial resistance, data harmonization, data integration, ontology

Abstract

One Health genomics initiatives often involve data streams originating from different sources, institutions, sectors, and information management systems. These are often heterogeneous datasets structured in a variety of ways, posing challenges for data harmonization, integration and meaningful interpretation. The Genomics Research and Development Initiative Shared Priority Projects for AMR (GRDI-AMR) uses a genomics-based approach to understand the prevalence and diversity of antimicrobial resistance determinants associated with food production and different environments that can impact human health, as well as how AMR can evolve, spread, and be mitigated. This work is being carried out by six different federal government departments and agencies, academic institutions, as well as agricultural and environmental networks. To facilitate harmonization of data, a modular, interoperable contextual data (metadata) specification was developed, called the GRDI-AMR One Health specification package. The package consists of an ontology-based data standard, built using semantic best practices and existing standards, and is operationalized in a data curation tool called the DataHarmonizer. This tool automates the transformation of contextual data into NCBI's One Health Enterics BioSample format to support public data sharing. The package also includes different kinds of support materials such as field and term reference guides and a detailed curation protocol highlighting ethical, practical and privacy considerations. Tooling and vocabulary were iteratively improved through multiple rounds of real-world testing. The data standard is continually maintained and version controlled, and has been used to resolve a variety of data harmonization issues experienced throughout numerous collaborative surveillance projects. The standard also encourages the inclusion of prevalence metrics in order to make whole genome sequencing data more useful for risk assessment, and enables communication about data needs between data generators and users. While developed for Canadian surveillance, the GRDI-AMR specification has also been implemented in international harmonization efforts, demonstrating its utility for many types of One Health genomics projects. The specification package is available at (https://github.com/cidgoh/GRDI_AMR_One_Health).

Introduction

Genomic surveillance used in a One Health context is a powerful tool for understanding the impact of pathogens on human, animal and environmental health, as well as their evolution and spread (Djordjevic et al, 2023). Genomic surveillance of pathogens requires high quality sequence data along with well-structured contextual data to enable its interpretation, including but not limited to: sample metadata; laboratory, clinical, epidemiological, and environmental information; methods, provenance, and quality control metrics. The One Health concept recognizes that the health of humans, animals, and plants are closely linked and interdependent within different ecosystems. As such, One Health initiatives often involve data streams originating from different sources, agencies, sectors, and information management systems. These heterogeneous datasets are structured in a variety of ways, posing challenges for data harmonization, integration and meaningful interpretation. The challenges of integrating and re-using non-standardized data are numerous (Sielemann et al, 2020; Pettengill et al, 2021; Gonçalves & Musen, 2019; Mussen et al 2022), as are the benefits of using well-structured and harmonized data (Cernava et al, 2022; Zeb et al, 2021; Damerow et al, 2021; Schriml et al, 2020; Griffiths et al, 2022; Griffiths et al, 2024; Timme et al, 2023; Yilmaz et al, 2011; Field et al, 2008; Dugan et al, 2014).

Structuring contextual data using data standards and ontologies, generates information that is more easily understood and used by both humans and computers, and can be more easily reused for different types of analyses (Pettengill et al, 2021; Lambert et al, 2017; Griffiths et al, 2017). Currently, genomics contextual data standards in public repositories are structured using BioSample packages based on Minimum Information (MIxS) checklists (Courtot et al, 2019). While these are very powerful tools for harmonization in many areas, they are often intentionally general in scope to broaden their applicability. As a result, there can be gaps in applying them to specific pathogens and/or in One Health programs. Ontologies are sets of hierarchical controlled vocabulary, in which terms are linked by logical relationships. The meanings of terms are meant to be universal rather than institution- or project-specific, and are disambiguated using universal identifiers (Internationalized Resource Identifiers (IRIs)) (Smith et al, 2007). With the emphasis on common meaning rather than relying solely on term labels, ontologies incorporate synonyms and database mappings, contributing to interoperability as there is rarely a “one-size fits all" terminology or nomenclature system. While ontologies are largely used to annotate genomic contextual data with standardized terms, the ability to relate entities via axioms and hierarchical groupings better enables standardized classification schemes and the construction of knowledge graphs for more complex queries. These can be employed in different types of artificial intelligence applications and analyses (e.g. machine learning approaches). Communities of practice like the Open Biological and Biomedical Ontology Foundry (OBO Foundry) articulate and implement best principles and practices to enable reuse of terminology across domains and sectors (https://obofoundry.org/). A number of registries and portals promote FAIR (Findable, Accessible, Interoperable, Reusable) ontology development and exploration (e.g. EBI’s Ontology Lookup Service, Ontobee, BioPortal), as well as data modeling languages (LinkML, OWL, RDF) and tools (Protégé, ROBOT, OntoFox) that improve data reuse and interoperability (Whetzel et al, 2011; Ong et al, 2017; Gennari et al, 2003; Jackson et al, 2019).

In support of the Canadian Federal Action Plan for Antimicrobial Resistance (AMR) and Use in Canada (Public Health Agency of Canada, 2015), as well as the Pan-Canadian Action Plan on Antimicrobial Resistance (Public Health Agency of Canada, 2023), the Genomics Research and Development Initiative Shared Priority Projects for AMR (GRDI-AMR and GRDI-AMR-One Health (GAOH)) use a genomics-based approach to understand how food production and environmental processes contribute to the development of AMR of human health concern, and explores strategies for reducing AMR in natural and anthropogenic ecosystems (e.g. oceans and food production environments). To better harmonize GAOH contextual data across public health, agriculture, environment, and food domains, an ontology-based data specification was developed based on ISO standard whole genome sequencing (WGS) requirements (ISO 23418:2022; https://www.iso.org/standard/75509.html). This resource consists of a variety of standardized fields, pick lists of controlled vocabulary, and prescribed formats for the harmonized capture of contextual data. The specification was implemented via a spreadsheet-based collection instrument, and accompanied by a variety of support material (curation SOP, reference guides, curation tools, New Term Request system). The specification was mapped to an NCBI submission format (One Health Enteric BioSample packages) to enable interoperability, and tooling was developed to help automate data transformations for sequence submissions. The specification was iteratively improved through several rounds of testing and incorporation of feedback from users. The specification package is version controlled and available at https://github.com/cidgoh/GRDI_AMR_One_Health. This resource is currently being implemented across GRDI-AMR partners, with a focus on a comprehensive integration within the federal genomics ecosystem. While designed for Canada, the data standard is also readily adaptable for international use, as exemplified by different partnerships.

Results and Methods

Data generation and analysis are key results of GRDI projects, however, just as important is the inter-agency collaboration and communication that they support. Historically, data curation and harmonization have been on an *ad hoc* basis (e.g. Tyson et al, 2019). The GRDI-AMR project began in 2016 and focused on AMR determinants in food production systems. The challenges of data harmonization came to bear during this work, which catalyzed the development of a national standard. The contextual data specification created during the GRDI-AMR project was later expanded during the GRDI-AMR-One-Health project to address samples collected from broader terrestrial and aquatic environments.

Design Principles

The GRDI-AMR established a Metadata Working Group (WG) tasked with enhancing data harmonization by improving contextual data quality and sharing practices. The WG members have over a decade of experience in developing ontologies and data standards, and have contributed to national and international harmonization initiatives, including the development of an ISO standard prescribing requirements and guidance for the use of whole genome sequencing (ISO 23418:2022: Microbiology of the food chain - Whole genome sequencing for typing and genomic characterization of bacteria - General requirements and guidance). The WG adopted a consensus approach focusing on inclusivity of a wide variety of stakeholder needs and preferences, and re-using existing community standards. A set of OBO Foundry best practices were also implemented to ensure the data standard and its associated standardized terms were FAIR (Table 1).

The WG first performed a data needs assessment in consultation with GRDI-AMR participants. Existing standards (e.g. ISO 23418) and contextual data formats, minimum information checklists, OBO Foundry ontologies, public repository requirements, and data collection instruments were also reviewed. These assessments identified the need for additional standardized terminology, clarification about the meanings of existing terms and their appropriate usage, the need for better curation tools and automation between public repositories and local data management formats, and the need for training and guidance in data curation and harmonization. The assessments also highlighted the need for standards for future-proofing data for better internal (organization-specific) reuse, to streamline exchange between trusted partners, and to support public repository submissions. As such, the specification package was scoped for data collection as well as discretional sharing within organizations, within networks and if desired, with public repositories.

Once data needs were assessed, lists of vocabulary were drafted and standardized terms were sourced from existing OBO Foundry ontologies. If no appropriate terms existed, new terms were created and submitted to appropriate ontologies - including ISO recommended ontology terms sourced from the publicly available ISO slim repository (https://github.com/GenEpiO/iso2017). ISO 23418 provides lists and annexes describing fields and terms, while the ontology file provides hierarchical classes of vocabulary. As an overarching organizational structure was missing, a modular, interoperable framework was developed in which standardized terms were grouped into thematic modules (Figure 1). The framework helps to create interoperability as new data standards can be built using the modules as building blocks - new specifications can be created by mixing-and-matching modules and enriching/depleting fields and terms within modules.

This framework was re-used to rapidly develop data specifications during different health crises such as the SARS-CoV-2 pandemic and the MPOX epidemic in 2022 (both specifications are currently implemented by the Public Health Agency of Canada to harmonize Canadian genomic surveillance data), and it is currently supporting the development of a Highly Pathogenic Avian Influenza specification. The framework has also been used to develop a specification for wastewater-based genomic surveillance. To support the collection and sharing of antimicrobial phenotypic data, a module was created based on an ontologized and harmonized version of National Center for Biotechnology Information (NCBI) and European Nucleotide Archive (ENA) antibiograms (Table 2).

Owing to the wide variety of data that needed to be compared, an aggregative approach to harmonization was used. “Aggregative” fields (higher level fields that combine different ontology classes) were often used to integrate different data elements (e.g. “environmental site” to capture information about many types of built and natural environments). This strategy differs from disaggregative approaches which focus on capturing information in greater detail e.g. instead of using an “environmental site” field, using separate fields for farm types, production facility types, and retail outlets. Different standards are fit for different purposes, and design decisions should reflect the types of questions that are being investigated. These different approaches are both valid, but aim to achieve different goals e.g. data integration for broad comparisons vs narrowly scoped fields for granular analyses. The “aggregative” approach is appropriate when there are many different sample types in a collection, especially when there are relatively few samples of particular types, as it is one strategy for avoiding specification bloating (unwieldy increase in the number of fields). The differences in aggregative and disaggregated approaches are demonstrated in the “Mapping and Interoperability” section. The benefit of using ontologies for structuring and standardizing information using these different approaches is that they provide a hierarchical pathway to linking higher level terms to more granular terms (e.g. production facility type > food production facility type > peanut butter production facility).

Specification Content

The specification contains 775 fields and >1400 standardized pick list values, either sourced from, or contributed to, 18 different OBO Foundry ontologies (see Table 3 for the list of ontologies used). The specification captures information regarding identifiers and accession numbers, enabling tracking of materials across agencies and establishing chains of custody; sample collection and processing (e.g. sample types, sample plans, sample storage, sample properties, context of sampling and presampling events that may impact results, production streams, etc); environmental conditions and measurements (e.g. weather, precipitation); host information including scientific and common names, breeds, ecotypes, etc; strain and isolate information (e.g. typing results); sequencing information (library prep, sampling strategies, sequencing methods); bioinformatics and quality control (e.g. QC methods, read filtering/trimming, dehosting, coverage, reference genomes, etc); taxonomic identification methods (e.g. reference mapping databases and versions, software tools, etc); risk assessment information (e.g. prevalence, stages of production, experimental interventions); public repository information (public repository submissions, reference material); author information (contact information for follow up and attribution, as well as stewardship); and AMR phenotypic profiling (e.g. minimum inhibitory concentration (MIC) measurements, breakpoints, interpretation standards, phenotypes, etc).

As data curation can be burdensome on resource-constrained agencies, a minimal set of contextual data (“required fields") was identified to minimize data entry and prioritize efforts. The set of required fields was scoped for surveillance needs, based on consensus and experience from different genomic surveillance initiatives. The list of required fields is presented in Table 4. Although the specification is large, there are only 16 required fields per record, and six required AMR phenotypic testing fields (per agent tested, per sample). The required fields cover sample and isolate identifiers, sample collection metadata (e.g. geographical location and date of collection), high level sampling strategy information, taxonomic information about the target organism, AMR minimum inhibitory concentration (MIC) values, and different types of contact information for follow up. The complete specification is much larger than the required set of fields in order to accommodate the many different use cases of the GAOH specification. The “enhanced contextual data fields” (fields of information that augment the minimal required set) provide flexibility for many other data types that providers may wish to include for specific studies.

Within the specification, standardized terms are provided in pick lists along with their IRIs, which link project-specific preferred labels to information such as definitions, definition sources, alternative labels used by other communities and organizations (synonyms), as well as logical relationships linking the term to other kinds of information (axioms), and hierarchical relationships enabling different types of classifications and groupings. The identifiers are presented in square brackets and consist of an ontology prefix (identifying the source ontology) and a numerical identifier (provided by the source ontology). An example of this structure for the term “chickpea” would be “Chickpea [FOODON:03306811]”, derived from the Food Ontology. These identifiers are provided not only in order to make annotated data FAIR, but to address the issue of the use of alternative labels. Using ontology identifiers to represent an entity sidesteps challenges and misinterpretations arising from the use of synonyms and alternative spellings as a computer can more easily recognize the same terms without relying on exact text (label) matches (e.g. “Chickpea” vs “Chick pea” vs “Garbanzo bean” are all synonyms of FOODON:03306811).

Ontology Development

The Labour of Minting and Deprecating Terms in Ontologies

Strengths of ontology communities of practice such as the OBO Foundry include their collaborative nature and that their products are community-driven resources. Ontologies are usually created using project-specific funding, and maintenance of these resources are chronically underfunded and efforts short-staffed. As a result, term requests can take a long time to process. While submitting terms to other ontologies, significant labour was involved in temporarily adding these terms to GenEpiO, Simon Fraser University’s Genomic Epidemiology Ontology (GenEpiO) designed to support genomic epidemiology and pathogen genomics (https://genepio.org/).

GenEpiO provides a flexible framework that also allows the creation of temporary mints (allocation of identifiers) for terms while awaiting their acceptance in other ontologies, which is particularly useful for meeting project deadlines. In the GenEpiO ontology, a structured process for creating and managing new terms is used, including the temporary minting of IDs while awaiting requested terms to be integrated in other ontologies - a critical process for ensuring that terms are findable and usable while awaiting formal acceptance in the target ontologies. The temporary IDs were managed through the "Mint/Reservation Protocol," available at https://github.com/GenEpiO/genepio/wiki/Protocol:-Mint-ID-Reservation#id-reservation-protocol-in-progress which distinguishes between reserving an ID for an in-progress term and minting an ID for a fully curated term.

For terms expected to be accepted in other ontologies, a "Planned Obsolescence Protocol" was implemented, available at https://github.com/GenEpiO/genepio/wiki/Protocol:-Planned-Obsolescence, in order to create temporary GenEpiO IDs and track their submission status. Once accepted, these terms were deprecated in GenEpiO and replaced by the permanent IDs from the target ontologies. This protocol facilitated a smooth transition, ensuring that all terms eventually resided in the most appropriate ontology. Throughout this process each step is documented, including tracking reservation dates and editor notes, to maintain traceability and clarity in the curation process.

Ontology Development Process in Response to a New Term Request

Approximately 1,400 terms were incorporated into the GRDI specification through a structured process designed to evaluate the relevance and context of new term requests before integrating them into the existing ontology framework. This process began with mapping vocabulary to existing ontology terms using ontology lookup services such as the European Bioinformatics Institute Ontology Lookup Service (EBI OLS) and Ontobee. These services were frequently used to check if the requested term already existed in any ontology.

When relevant ontology terms were not available, new terms along with their annotations were proposed and submitted to various OBO Foundry ontologies. Formalized, ontological definitions were developed for the terms (in addition to the user-friendly terms provided in the reference guides) for these terms, adhering to OBO Foundry principles. Balancing formal definitions with user perspectives is crucial for effective ontology development and usage, ensuring both precision and relevance in representing domain knowledge.

Communication to Users

Users of the ontology are informed about the deprecation of terms through release notes, documentation updates, or other communication channels. This ensures that users are aware of changes in the ontology and can adjust their usage accordingly. Deprecating terms is a necessary aspect of ontology maintenance, ensuring that the ontology remains current, accurate, and aligned with evolving domain knowledge and best practices.

Publication and Documentation

Once the new term is successfully integrated or accepted, the updated ontology (GenEpiO, FoodOn, and others) is published and made available to users. Additionally, changes are also published as part of an updated version of the GRDI specification. Documentation is provided to inform users about the newly added term, including its definition, usage guidelines, and any associated changes to the curation of developed specification.

Community Feedback and Iteration Development

The ontology development process remains iterative, and feedback from the user community is essential for continuous improvement. Our curators actively solicit feedback, monitor usage patterns, and address user concerns to refine and enhance the vocabulary over time. By following this structured ontology development process, we ensure that new term requests are systematically evaluated, integrated, and documented. The documentation is also vital for managing the tracking of fields and terms and ensuring the ontology remains up-to-date and accurate.

Tools for Implementation

In order to put standards into practice, it is necessary to operationalize them with accessible, easy-to-deploy, easy-to-use tools. The DataHarmonizer is a template-driven spreadsheet application for harmonizing, validating and transforming genomics contextual data into submission-ready formats for public or private repositories. The tool's web browser-based JavaScript environment enables validation and its offline functionality and local installation increases data security. The DataHarmonizer was developed to address the data sharing needs that arose during the COVID-19 pandemic and implements data harmonization specifications as data collection templates, and organizes schemas using LinkML - a modelling language for schemas and data dictionaries that enables specifications to be represented in JSON, YAML csv and other formats (Gill et al, 2023). The GAOH specification was engineered as a template in the DataHarmonizer - a tool that offers a number of curation features, validation, as well as automated transformations - enabling data providers to enter their data once and export it in different formats for downstream applications (e.g. Integrated Rapid Infectious Disease Analysis platform, NCBI) The GAOH template can be found in the Pathogen Genomics Package of pathogen genomics surveillance templates (https://github.com/cidgoh/pathogen-genomics-package/releases). A colour-coding scheme was introduced in which “required” fields were colour-coded in yellow, recommended fields were coloured purple, and optional fields were in white. The DataHarmonizer also provides user tutorials as well as reference guides for fields and terms. More information on LinkML template schemas as well as links to code can be found at https://github.com/cidgoh/DataHarmonizer/wiki/DataHarmonizer-Templates.

Supporting Materials

To enable users to become familiar with the data specification content, proper usage of the DataHarmonizer template, and good curation practices, supporting materials were developed. The support material package included a curation standard operating procedure (SOP) containing instructions for getting started, as well as guidance for applying the standard to different data types and scenarios including ethical, privacy and practical considerations (https://github.com/cidgoh/GRDI_AMR_One_Health/tree/main/SOPs). To provide standardized definitions, guidance for data entry, and examples of use for all fields and terms, reference guides were also developed and made publicly available via GitHub and within the DataHarmonizer (https://github.com/cidgoh/GRDI_AMR_One_Health/tree/main/Reference%20Guide). The curation SOP is periodically updated, and the reference guides are updated upon each new release of the DataHarmonizer template.

Versioning and Availability

The GRDI-AMR-One Health specification is free, publicly available and version controlled on GitHub. All edits and updates are tracked in release notes. (https://github.com/cidgoh/GRDI_AMR_One_Health). Versioning is done in the format of x.y.z.
x = Field level changes
y = Term value / ID level changes
z = Definition, guidance, example, formatting, or other uncategorized changes
Discussions contributing to updates are also tracked on the repository GitHub issue tracker.

Testing, Sustainability, and the Development Cycle

To ensure the specification was fit-for-purpose, three rounds of testing were conducted (2018, 2020, 2023) with different curators from different labs. Workshops were held to provide instruction to volunteer data curators, and the specification package was distributed to all volunteers. Lab-specific datasets were curated and submitted to the Simon Fraser University (SFU) development team for feedback. The development team took note of template areas that were poorly implemented due to lack of appropriate instruction or inappropriate/missing vocabulary, and incorporated feedback received in debriefing sessions.

As more labs implemented the specification, new data types and vocabulary were necessary. Specification developers worked with different ontologies to synchronize resources. Ontology development is described in more detail in the Methods section.

To track data needs over time and enable community discussions about ontology terms, a New Term Request (NTR) System was developed. An NTR template was created and uploaded to the GitHub IssueTracker. Publicly available term requests can be made via the IssueTracker. As ontology maintenance and responsiveness are core principles of the OBO Foundry, the NTR system is a critical part of our sustainability plan, fulfilling community expectations. Templates, tools, and supporting materials are all updated regularly as needed, and version controlled.

Harmonization Challenges and Solutions - Worked Examples

The GAOH contextual data specification was used to solve many harmonization challenges, which have been articulated below. A series of worked examples highlighting specific issues and solutions are also presented. GAOH data is currently mostly private, however there are efforts underway to deposit data in public repositories. As a result, while based on real GAOH scenarios, all of the worked examples involve simulated data.

Null Values

Null values are entities (markers) used to indicate that a piece of information cannot be provided. There can be a variety of reasons that particular data elements cannot be populated e.g., not collected, the information is restricted for privacy concerns, the requested information does not apply to the sample being described, etc. To avoid heterogeneity of null values, the GRDI specification uses standardized null values provided by the INSDC (International Nucleotide Sequence Database Collaboration). These terms have been ontologized and are available in the Genomic Epidemiology Ontology (GenEpiO). Null values consist of the following:

1. **Not Applicable** [GENEPIO:0001619]: A categorical choice recorded when a datum does not apply to a given context.
2. **Not Collected** [GENEPIO:0001620]: A categorical choice recorded when a datum was not measured or collected.
3. **Not Provided** [GENEPIO:0001668]: A categorical choice recorded when a datum was collected but is not currently provided in the information being shared. This value indicates the information may be shared at the later stage.

4. **Missing** [GENEPIO:0001618]: A categorical choice recorded when a datum is not included for an unknown reason.
5. **Restricted Access** [GENEPIO:0001810]: A categorical choice recorded when a given datum is available but not shared publicly because of information privacy concerns.

Machine Readability

Machine-readability of contextual data has been improved in the specification with certain goals in mind: to facilitate interchange between standards and datasets, to implement different standardized values from domain ontologies, to enrich datasets with information not usually included. Steps to realize these goals include: separating values of measurements from units, separating versions from methods (in most places), including fields for methods in many sections of the specification to encourage providing more methodology, e.g. “water depth” and “water depth units”; “serovar” and “serotyping method”.

Sample Descriptions

Nouns and Modifiers - Avoiding “Word Bombs”

The OBO Foundry implements a common overarching organizational structure called the Basic Formal Ontology (top-level ontology) in which all *things* that exist are divided into three categories - material entities, processes, and qualities (characteristics). In line with this structure, the GRDI specification separates proper nouns from their modifiers and processes, with some exceptions. This structure is simpler, better enables mapping between different dictionaries and schemas, and is more machine-readable. Moreover, this structure helps to avoid “word bombs” that result from composite terms. “Word bombs” are inexhaustible lists that result from combinations of nouns and modifiers that are difficult to update and maintain e.g. Mango; Whole mango; Mango chunks; Bagged mango; Frozen mango; Frozen mango chunks; Frozen, bagged, mango chunks, etc.

Food Products and Properties

There are many different thesauri and food product indexes around the world that have been developed for different purposes - Codex Alimentarius, FoodEx2, IFSAC, LanguaL, FoodData Central, CFSIN, etc to name a few. Food data dictionaries, however, are often project-, system-, or organization-specific, and classify food products and structure food information in different ways. The GAOH specification uses FoodOn, a farm-to-fork food ontology that structures foods and food information according to 14 facets for describing food source plant and animal organisms, food preservation, cooking, packaging, consumer groups, labeling, etc. FoodOn reuses terms from several OBO Foundry ontologies such as environmental terms from ENVO, agriculture terms from AGRO, plant and animal anatomy terms from UBERON, and PO, organisms from NCBITaxon, relations from RO, and nutritional components from CDNO. Conversely, FoodOn terms are reused in a growing list of ontologies such as ENVO, CDNO, ONE, ONS, FIDEO, FOBI, ECTO, and DOID, enhancing interoperability and creating a food semantic ecosystem. FoodOn has also been used by the US FDA’s GenomeTrakr and the USDA’s FoodData Central, and work is underway to create relationships to EFSA’s FoodEx2 system. In the GRDI specification, food information is structured using eight fields, separating food products and food production environments (proper nouns) from food qualities (raw, RTE, organic, frozen etc) and food origins.

Environmental Sites and Materials

Samples can be associated with built, or natural, environments and environmental materials. In the GAOH specification, such samples can be described as being environmental materials (substances and things) and/or associated with particular environmental sites (locations). These fields can be alone, or used in conjunction with other fields (e.g. hosts and foods) when needed.

An additional field, “animal or plant population” helps to characterize the inhabitants of a particular place. For example, farms may grow different kinds of crops and may raise different types of animals. The combinations of animal and plant populations on farms can be difficult to describe if the user needs a specific “type” of farm for every situation and combination of products it produces (e.g. chicken farm, turkey farm, poultry farm, crop farm, turkey and grain farm, poultry and apple farm, fish farm, mollusc farm). This field can be used as a noun modifier, in order to multi-tag the pertinent inhabitants and commodities of an environmental site.

Environmental Conditions and Measurements

Environmental conditions at the time of sampling, as well as prior to sample collection, can impact results. Furthermore, samples may be associated with certain water depths and temperatures, which may influence microbial compositions within those samples. The GAOH template provides different fields for capturing weather conditions before and during sampling, as well as precipitation, temperature and other environmental measurements.

Hosts, Anatomical Parts, and Anatomical Materials vs Body Products

When samples are derived from a living thing (e.g. human, plant, animal), they are captured as “hosts” in the GAOH specification. Hosts can be described using their common names (e.g. “Human”) or their scientific/taxonomic name (e.g. *Homo sapiens*). Similar to environmental samples, anatomical samples from hosts can be captured as anatomical materials (*what* was sampled from the host) and anatomical parts (*where* a material was taken from the body). A third category - body product - is used for materials that are excreted/secreted usually as waste (e.g. feces, urine, vomit) that would not be considered anatomical materials in an anatomy textbook. Hosts can also be described by production type names which are based on the sexual maturity and/or age/weight of an animal. Common production animal terms are provided in the “host (food production name)” field e.g. Cow, Freemartin cow, Heifer, Steer, Feeder cow, Finisher cow, Milker cow, Stocker cow, Weanling cow. Other fields are also available for capture of information regarding breeds and ecotypes (i.e. “host (ecotype)” and “host (breed)”).

Temporal Sample Collection

Samples - most often environmental samples - can be collected at discrete points in time (e.g. particular time or one a particular day), or they can be collected continuously over a period of time. The GAOH specification provides an array of fields for capturing temporal sample collection information, such as sample collection start and end dates, sample collection start and end times, the duration of sample collection, and more.

Risk Assessment

To use genomic information for risk assessment, risk assessors require sample sizes and the number of positive and negative samples (i.e. denominators are needed to determine prevalence and risk ratios), in addition to nomenclature and typing information identifying pathogen hazards (e.g. species, strain, clonal complex, sequence type, etc.). The GAOH specification aims to encourage the inclusion of prevalence information in the “risk assessment information” module in which data providers can indicate the types of prevalence metrics that are available, as well as information about stages of production and experimental interventions being tested.

Worked Examples

A variety of scenarios were selected to highlight contextual data development in diverse situations. Partial contextual data records are provided below to illustrate use of different sample collection, environmental conditions and measurements, and risk assessment fields.

Scenario 1: A swab of an egg belt was collected in a chicken hatchery. 16/200 farm samples were positive for *Salmonella.*
original_sample_description: swab of an egg belt in a chicken hatchery
collection_device: Swab [GENEPIO:0100027]
environmental material: Egg belt [AGRO:00000670]
environmental site: Poultry hatchery [ENVO:01001874]
prevalence_metrics: 16/200 farm samples were positive for Salmonella

Scenario 2: A sample of weep fluid was collected from a packaged carcass in an abattoir. 49/200 abattoir samples were positive for *Salmonella.*
original_sample_description: weep fluid from a carcass in an abattoir
anatomical_part: Carcass [UBERON:0008979]
environmental_material: Weep fluid [AGRO_00000692]
environmental_site: Abattoir [ENVO:01000925]
prevalence_metrics: 49/200 abattoir samples were positive for Salmonella

Scenario 3: Retail skinless, boneless chicken thighs were sampled in a grocery store. 3/200 samples were positive for *Salmonella*.
original_sample_description: retail SLBL chicken thighs
food_product: Chicken thigh (skinless, boneless) [FOODON:03000417]
environmental_site: Retail environment [ENVO:01001448]
prevalence_metrics: 3/200 samples were positive for Salmonella

Scenario 4: A survey of stone fruit revealed the presence of *Acinetobacter baumannii* in 3/18 frozen, bagged mango samples.
original_sample_description: retail - frozen mango, bagged

food_product: Mango (whole or parts) [FOODON:03000217]
food_product_properties: Food (frozen) [FOODON:03002148]
food_packaging: bag, sack or pouch [FOODON: 03490197]
environmental_site: Retail environment [ENVO:01001448]
prevalence_metrics: 3/18 samples were positive for *Acinetobacter baumannii*

Scenario 5: An imported cow from the US that had stopped producing milk and had a mastitis infection was sampled in the Canadian Midwest (specifically Alberta, but data needed to be obfuscated due to privacy concerns).
host (scientific name): Bos taurus [NCBITaxon:9913]
host (common name): Cow [NCBITaxon:9913]
host (food production name): Dry cow [FOODON:00004411]
host_origin_geo_loc_name (country): United States of America [GAZ:00002459]
geo_loc name (country): Canada [GAZ:00002560]
geo_loc name (state/province/region): Prairie region (Canada) [wikidata:Q1364746]

Scenario 6: A river water sample was collected on a warm, sunny day (28ºC). However, it rained heavily 24 hours before sample collection which may have contributed to local farm run-off.
sampling_weather_conditions: Sunny/Clear [ENVO:03501421]
air_temperature: 28
air_temperature_units: degree Celsius (C) [UO:0000027]
presampling_weather_conditions: Rain [ENVO:01001564]
precipitation_measurement_value: 25
precipitation_measurement_unit: millimeter (mm) [UO:0000016]

Scenario 7: A sample of wastewater was continuously collected over a 24 hour period from a wastewater treatment plant.
environmental_material: Wastewater [ENVO:00002001]
environmental_site: Wastewater treatment plant [ENVO:00002272]
sample_collection_time_duration_value: 24
sample_collection_time duration_unit: Hour [UO:0000032]

Provenance and Contact Details

The production of genomic sequence data can involve different partners and laboratories responsible for different processes such as sample collection, culturing and characterizing isolates, sequencing and bioinformatic analysis. All of these organizations have contributed time, resources, labour, and intellectual input. Tracking the roles and contributions of different partners is not only good practice for auditability, articulating chains of custody and maintaining contact information for follow-up, but is also good ethical practice for equitable benefit sharing (e.g. tracking contributions to publicly available datasets, utility in analyses and publications and other metrics used to secure funding). The GAOH specification provides fields for provenance tracking for organization and laboratory names, and contact names and emails regarding sample collection, culture and isolation, sequencing, and AMR phenotypic testing. The specification also provides standardized names for organizations (as provided by the organizations) which have been made available in the Genomic Epidemiology Ontology (GenEpiO). The use of generic email addresses that are directed to organization data stewards is encouraged to ensure that contact information remains current and future-proofed, enabling access to more extensive contextual data that may be available at host organizations.

Worked Example

Scenario: A sample was collected by Johnny Bloggs in the Diarra Lab (AAFC), and an isolate was produced from the sample. As part of a collaboration, the isolate was sequenced by a scientist at the Public Health Agency of Canada (Arvinder Singh). The isolate was also assessed for antimicrobial susceptibility by a collaborator at the CFIA. A partial contextual data record for the sample highlighting different provenance fields, is provided below.
sample_collected_by: Agriculture and Agri-Food Canada (AAFC) [GENEPIO:0100553]
sample_collected_by_laboratory_name: Diarra Lab
sample_collector_contact_name: Johnny Bloggs
sample_collector_contact_email: diarralab@aafc.ca
isolated_by: Agriculture and Agri-Food Canada (AAFC) [GENEPIO:0100553]
isolated_by_laboratory_name: Diarra Lab
isolated_by_contact_name: Johnny Bloggs
isolated_by_contact_email: diarralab@aafc.ca
sequenced_by: Public Health Agency of Canada (PHAC) [GENEPIO:0100551]
sequenced_by_laboratory_name: Not Provided [GENEPIO:0001668]
sequenced_by_contact_name: Arvinder Singh
sequenced_by_contact_email: seqlab@phac.ca
AMR_testing_by: Canadian Food Inspection Agency (CFIA) [GENEPIO:0100552]
AMR_testing_by_laboratory_name: The Ottawa Laboratory (Fallowfield)
AMR_testing_by_contact_name: Zhang Wei
AMR_testing_by_contact_email: amrlab@cfia.ca

Structuring and Capturing Experimental Design

Capturing Biases in Sampling and Sequencing Strategies

Sampling strategies include criteria for sample selection. These criteria can create biases within datasets (e.g. baseline surveillance vs outbreak investigation) and so it is important to capture. The “purpose of sampling” field in the specification captures why a sample was originally collected (e.g. environmental monitoring, animal health diagnostics, public health surveillance), while the “purpose of sequencing” field captures why the sample was selected from a collection for sequencing. Sample plans often have many more details and insights pertaining to sampling strategies and so are highly useful to link to in sequence contextual data. Sample plan information can be included using the “sample plan name” and “sample plan ID” fields (e.g. Sample plan name: ESBL-producing bacteria from frozen stone fruit, Sample plan ID: GRDIAMR-WP4.2-Stonefruit).

Pre-Sampling Activities and Experimental Interventions

Often, there will be events that occur upstream of sample collection that may impact downstream results (e.g. vaccination, application of fertilizer, changes in sanitization practices, etc). These events can be deliberate, or incidental, but it is useful to record information about them for comparisons and analysis. The GAOH specification provides two fields - “pre-sampling activities” (under Sample collection and processing) and ”experimental interventions” (under Risk assessment information) - to capture activities upstream of collection. Where pre-sampling activities can be more general and incidental (e.g. application of fertilizer on an adjacent farm), experimental interventions are deliberate activities intended to test variable experimental conditions (e.g. mutagenesis, pre-treatment of food/water with medications or antimicrobials). Picklists of common kinds of upstream activities have been provided, along with free text fields (pre-sampling activity details and experimental intervention details) to enable more detailed descriptions of activities if necessary.

Roles of Samples In Experimental Designs

Samples can play different types of roles in an experiment - they can be experimental samples that are being tested, they can be different types of replicates (technical replicates, biological replicates) enabling quality control and methodological comparisons, they can act as controls (positive control, negative control, and different types of reference materials such as EQA panels). Sometimes samples and data are created in a lab for testing different methods, and represent synthetic lab constructs. Samples that are replicates, control, and synthetic constructs can be tagged as such using the “experimental specimen role type” field. Samples not in these special categories do not need to be tagged.

Specimen Processing

Samples can often be processed prior to library preparation in a way that may affect downstream results e.g. pooling of samples. These methods can be tagged in a standardized way using the “specimen processing” and “specimen processing details” fields.

Worked Examples

Scenario 1: Five surface water samples were collected from the same location by Environmental and Climate Change Canada (ECCC) as part of methods optimization for a new aquatic monitoring program (sample plan name: Healthy Oceans 2024; sample plan ID: QR-1234). The five samples are considered biological replicates. The samples were later sequenced and compared. A partial contextual data record for one of the samples highlighting useful fields for describing sampling strategies and roles of samples, is provided below.
purpose_of_sampling: Field experiment [GENEPIO:0100550]
sample_plan_name: Healthy Oceans 2024
sample_plan_ID: QR-1234
experimental_specimen_role_type: Biological replicate [OBI:0000198]
purpose_of_sequencing: Protocol testing experiment [GENEPIO:0100024]

Scenario 2: Five surface water samples were collected from the same place, pooled, and sequenced by Environmental and Climate Change Canada (ECCC) as part of a routine aquatic monitoring program. A partial contextual data record for one of the samples highlighting useful fields for specimen processing, is provided below.
specimen_processing: Samples pooled [OBI:0600016]

Scenario 3: A wastewater sample was collected from a wastewater treatment plant post grit removal. A partial contextual data record for the sample highlighting useful fields for presampling activities, is provided below.
presampling_activity: Wastewater grit removal [GENEPIO:0100882]

Scenario 4: A sterile water sample is included in a sequencing run as a control. A partial contextual data record for the sample highlighting useful fields for tagging controls, is provided below.
experimental_specimen_role_type: Negative experimental control [GENEPIO:0101019]

Scenario 5: A government scientist is testing the effects of conventional farming practices vs organic practices. Fertilizer was spread 3 days before sampling on the conventional farm under study, which is a regular practice. A partial contextual data record for the sample highlighting useful fields for capturing presampling activities and experimental interventions, is provided below.
presampling _activities: Fertilizer pre-treatment [GENEPIO:0100543]
experimental_intervention: Conventional farming practices [GENEPIO:0100895]

Scenario 6: An abattoir is testing multiple methods for reducing the cross-contamination of *Campylobacter* during chicken slaughter. One intervention being tested is the order and timing of the slaughter of positive and negative flocks. A swab of a chicken carcass post-slaughter is collected for sequencing. A partial contextual data record for the sample highlighting useful fields for experimental interventions, is provided below.
experimental_intervention: Logistic slaughter [GENEPIO:0100545]

Available Data Types

The content of different contextual data records can vary depending on the questions being asked by a genomics-based program or project. The GRDI-AMR specification provides fields for the capture of widely used data types. However, other data types may be collected that can be made available by data providers upon request. Often, consultations and sometimes data sharing agreements are needed before they should/can be shared to ensure users are properly informed about acceptable use of these additional data types with regards to methodology, scope and/or limitations of the data. The "available_data_types" field enables data generators to communicate that additional data types exist and could possibly be shared, without actually including the additional data in their datasets. Types of additional data include documentation (e.g. information about experimental parameters, feed or treatment histories, land use details), chemical characterization (e.g. pH, salinity, oxygen measurements), microbiological characterization (organism identification using different instruments and methods, phenotypic characterization), microbiological quantification results (e.g. colony counts, infectivity), and physical measurements and characterizations (e.g. conductivity, size measurements, turbidity). Multiple additional data types can be multi-tagged in this field, as needed. While this list of additional data tags is not exhaustive, it can easily be updated as needed. Lists of available data types can act as catalogues, better exposing and characterizing different collections. The “available_data_type_details” fields enable free text notes about these additional data types to also be included.

Worked Examples

Scenario 1: An Environment and Climate Change Canada (ECCC) scientist is studying the effects of AMR determinants on aquaculture productivity and has mollusc shell length data and dissolved oxygen measurements that are associated with a metagenomic water sample. A partial contextual data record for the sample highlighting useful fields for additional data types, is provided below.
available_data_types: Mollusc shell measurement [GENEPIO:0100744]; Dissolved oxygen measurement [GENEPIO:0100709]

Scenario 2: A Canadian Food Inspection Agency (CFIA) scientist has sequences associated with an outbreak in particular food products. The scientist has PCR marker data and total fecal coliform counts associated with a sample. A partial contextual data record for the sample highlighting useful fields for additional data types, is provided below.
available_data_types: PCR marker detection [GENEPIO:0100735]; Total fecal coliform count [GENEPIO:0100730]

Scenario 3: An Agriculture and Agri-Food Canada (AAFC) scientist is studying chronic mastitis in dairy cows and has treatment and feed history data associated with a sample from a cow. A partial contextual data record for the sample highlighting useful fields for additional data types, is provided below.
available_data_types: Feed history [GENEPIO:0100704]; Therapeutic administration history [GENEPIO:0100706]

AMR Phenotype Testing

AMR phenotypic data is critical for establishing genotype-phenotype relationships. Antimicrobial phenotype testing can be carried out using a variety of methods, instrumentation, antimicrobial agents, interpretation criteria, etc. A MIC (minimal inhibitory concentration) generated using one method, may not be replicated if a different method was used, even using the same isolate. When comparing genotypic data with phenotypic data, it is important to track how the phenotypic data was produced. The INSDC has created and implemented Antibiogram standards that have been used to collect and organize thousands of phenotypic testing datasets. The INSDC standards were ontologized and included in GenEpiO and the ARO ontologies, and provided as a module within the GRDI specification, with a few additions. Fields specifying the provenance of the AMR characterization (e.g. lab name, date of testing, etc), interpretation standard version information, as well as fields for including breakpoint information were included.

Each agent tested has associated fields describing the name of the agent, the MIC value, its units and comparator, phenotypes (resistant, sensitive etc), breakpoint information (upper and lower thresholds), interpretation criteria (standard used to interpret the MIC such as Clinical & Laboratory Standards Institute (CLSI) or European Committee on Antimicrobial Susceptibility Testing (EUCAST)), and the methods and platforms used for testing. In total, 14 fields are associated with each agent being tested. Eight of these fields are associated with pick lists of standardized values. In addition to basic provenance, only MIC value, units and comparator fields (total of three fields) were considered “required”. Populating other fields is, however, strongly encouraged. In total, 43 antimicrobial agents and drug combinations are included in the specification, although more can be added as needed.

Isolates can be associated with many different tests (individual agents or panels). AMR phenotypic testing data represents a one to many data challenge for systems without the architecture for such relationships in backend storage or during upload. While we recommend more relational structures for capturing phenotypic testing data in databases, to fit the tabular structure requirements for storage of GRDI data, the 14 fields used to capture testing data are provided in the specification per drug.

Worked Example

Scenario: A bacterial isolate was tested for sensitivity to amikacin by a scientist at the CFIA using a bioMérieux VITEK 2 instrument. The MIC was interpreted using the CLSI M100 (2023) Standard. For future reference, the scientist wishes to include notes about recent changes to the current version of the interpretation standard being used. A partial contextual data record for the isolate highlighting useful fields for AMR phenotypic testing, is provided below.

AMR_testing_by: Canadian Food Inspection Agency (CFIA) [GENEPIO:0100552]
AMR_testing_by_laboratory_name: The Ottawa Laboratory (Fallowfield)
AMR_testing_by_contact_name: Zhang Wei
AMR_testing_by_contact_email: amrlab@cfia.ca
AMR_testing_date: 2023-03-26
amikacin_resistance_phenotype: Resistant antimicrobial phenotype [ARO:3004301]
amikacin_measurement: 64
amikacin_measurement_units: milligram per litre (mg/L) [UO:0000273]
amikacin_measurement_sign: greater than (>) [GENEPIO:0001006]
amikacin_antimicrobial_laboratory_typing_method: Broth dilution [ARO:3004397]
amikacin_antimicrobial_laboratory_typing_platform: Vitek System [ARO:3004403]
amikacin_antimicrobial_laboratory_typing_platform_version: VITEK 2
amikacin_antimicrobial_vendor_name: bioMérieux [ARO:3004406]
amikacin_antimicrobial_testing_standard: Clinical Laboratory and Standards Institute (CLSI) [ARO:3004366]
amikacin_antimicrobial_testing_standard_version: CLSI M100 (2023)

amikacin_antimicrobial_testing_standard_details: resistant breakpoint changed from 64 to 16 in CLSI 2023
amikacin_antimicrobial_susceptible_breakpoint: 4
amikacin_antimicrobial_intermediate_breakpoint: Not Provided [GENEPIO:0001668]
amikacin_antimicrobial_resistant_breakpoint: 16

Taxonomy Differences

Organisms can be identified using different methods and classified using different criteria which may change over time. The GRDI specification implements NCBITaxon, a taxonomy ontology based on sequence data, as a standard for specifying microbial species names. As the amount of sequence data increases and microbial taxonomy becomes more refined over time, NCBITaxon may be updated and some hierarchies and relationships may change. The GRDI specification is aligned with current nomenclature and will document and version control any taxonomic changes to nomenclature over time.

The specification also supports the tracking of taxonomic identification differences based on sequences vs other assays (e.g. culture and PCR-based methods). Sample records stored using the specification assume that taxonomic designations are sequence-based, but for those samples awaiting sequencing where an initial identification has been made, the organism taxons can be specified and the method can be documented using the "taxonomic_identification_process" field pick list. Records should be updated as organism names change over time, and the identification process should be included. Multi-tagging of methods is permitted.

Worked Example

Scenario: An isolate was taxonomically identified as *Klebsiella pneumoniae* using PCR and culture on differential growth media. After sequencing and bioinformatic analysis, the organism was identified as *Klebsiella pneumoniae* subsp. *pneumoniae*. A partial contextual data record for the isolate highlighting useful fields for tracking taxonomic identification (before and after sequencing), is provided below.

a) Before sequencing

organism: Klebsiella pneumoniae [NCBITaxon:573]
taxonomic_identification_process: PCR assay [OBI:0002740]; Comparative phenotypic assessment [OBI:0001546]

b) After sequencing

organism: Klebsiella pneumoniae subsp. pneumoniae [NCBITaxon:72407]
taxonomic_identification_process: Whole genome sequencing assay [OBI:0002117]

Sequencing and Bioinformatics Methods

Sequencing and bioinformatics methods can impact analytical results. Documenting these methods is critical for troubleshooting, as well as optimization, validation, reproducibility and auditability. The GRDI specification implements two different modules for capturing these types of information. The Sequencing module contains fields pertaining to sequencing assay types, library preparation, library enrichment strategies, sequencing protocols, insert/amplicon sizes, raw sequencing data file names and flow cell versions. The Bioinformatics and QC module contains fields covering quality control methods and outcomes, bioinformatic processing, tool and database names and version numbers, protocols and commonly used metrics.

Worked Example

Scenario: A water sample from coastal waters associated with oyster farming was collected. A metagenomics-based library was prepared and sequenced using a NovaSeq 6000 and the reads were processed (quality checked, filtered, trimmed) using the RAMPART 1.2.0 suite of tools. The reads were mapped to a custom reference taxonomic database (GVDEdb 3.4.5) using Bowtie2 (v2.5.3) to identify the presence of *Vibrio vulnificus*. A partial contextual data record highlighting useful fields for capturing various sequencing and bioinformatic processes, as well as database and tool names and versions, is provided below.
organism: Vibrio vulnificus [NCBITaxon:672]
sequencing assay type: Whole metagenome sequencing assay [OBI:0002623]
sequencing instrument: Illumina NovaSeq 6000 [GENEPIO:0100123]
raw sequence data processing method: RAMPART 1.2.0
read mapping software name: Bowtie2
read mapping software version: 2.5.3
taxonomic reference database name: GVDEdb
taxonomic reference database version: 3.4.5

Collaborative Development, Uptake, and Interoperability

The GRDI-AMR specification is being used to standardize and harmonize a wide variety of data generated by labs from six federal agencies and departments (Public Health Agency of Canada, Canadian Food Inspection Agency, Agriculture and Agri-Food Canada, Environment and Climate Change Canada, Fisheries and Oceans Canada, Health Canada), 15 universities and different agriculture and environmental networks across Canada. Harmonized data shared within the GRDI-AMR network are stored, shared, and analyzed in private projects housed within the IRIDA (Integrated Rapid Infectious Disease Analysis) platform (www.irida.ca). This work will be more fully described in an associated implementation paper in a forthcoming dedicated GRDI-AMR issue of the Canadian Journal for Microbiology (in preparation). Demonstrating the utility of the specification beyond Canadian borders, the standard was also used to strengthen different international data collection and sharing efforts described below.

Extensibility and International Implementation

The specification is being used to structure *Listeria monocytogenes* contextual data for a benchmark dataset in an international data sharing collaboration between Canada and the United Kingdom (Quadram Institute). A subset of the GRDI-AMR specification was used to structure One Health data in a national study involving various labs and agencies across Uganda (Redman-White et al, 2023). Pick lists were updated to reflect sample types, sampling strategies, locations and food types specific for the Ugandan context. A subset of the GRDI specification is also being included as part of gold standard benchmark datasets in the JPIAMR-funded B2B2B AMR Dx initiative (https://www.jpiamr.eu/projects/b2b2b-amrdx/).

Enabling Data Sharing and Interoperability

The GAOH specification is ***a*** solution, not ***the*** solution to contextual data harmonization. As discussed, different standards are created for different purposes - the key is enabling interoperability between them, and by extension, between the datasets they annotate and the systems that contain them. Data shared within the INSDC network must be structured according to prescribed formats and BioSample packages. The US One Health Enterics BioSample checklist developed by GenFS - a consortium of US genomics experts from the CDC, USDA, FDA, PulseNet and NARMS - is meant to support One Health Enterics genomics data sharing. This package is scoped for public sharing (in contrast to the GAOH specification meant for private collections, sharing with networks, as well as public sharing), and uses a more disaggregative approach (i.e. more specific fields) for answering specific questions, usually related to government surveillance of enteric pathogens. This package is currently being used to format GAOH data for NCBI submissions. The US One Health package uses a combination of MIxS fields and recommends different ontologies (and often particular branches within them) as sources of values e.g. FoodOn. A virtual Hackathon was held in 2021 in order to align the Canadian and American One Health standards and to discuss design principles and interoperability. The common use of ontology greatly enabled interoperability. To facilitate automated transformation of GRDI data into the US One Health Enterics format, fields and terms were mapped and an exchange format was developed (Figure 2). A mapping file demonstrating overlap as well as differences between the specifications is provided in Supplementary Table 1. The exchange format was implemented in the DataHarmonizer facilitating conversions of GAOH data into NCBI submission-ready BioSamples.

The GAOH specification also implements PHA4GE’s INSDC-compliant Data Object Model in that attributes pertaining to a sample (an entity collected in the clinical or field) are included in BioSample submissions, whereas attributes pertaining to sequence data are included in SRA metadata (Timme et al, 2023).

Discussion

While there are different data standards and formats that exist for capturing genomics contextual data, our assessments revealed that while existing standards incorporate domain expert knowledge, semantics and “logic-based grammar” was missing - i.e., the emphasis was on vocabulary without sufficient rules/patterns for how the fields/terms are structured and how fields are logically related, which inhibits more complex querying and integration of different types of data. Well structured and linked data is increasingly desirable as data management systems undergo modernization largely catalyzed by the genomics revolution.

The key strengths of this specification include consensus-based development; extensibility and reusability of vocabulary and modules; fluid knowledge exchange with the larger ontology development community; reuse of tools, training, competencies, and data expectations already created by the use of this framework for other pathogen surveillance initiatives. Critically, the content of the specification also creates a channel of communication between data generators and data users. This is accomplished by the inclusion of fields that prompt users to provide pieces of information useful for different analyses that may not have been previously collected due to lack of awareness e.g. sampling strategies, prevalence metrics for establishing denominators for risk assessments, as well as different contact information. As such, what has been created is not just “another metadata standard”, but rather a specification development ecosystem with built-in feedback loops - that both incorporate as well as impact - the user experience, the utility of data annotated by the standard, and the tools used to structure and harmonize the data (Figure 3).

Specification Performance and Ongoing Challenges

Iterative improvements of the specification occurred as it was applied to different kinds of datasets, which enhanced its overall ability to capture, standardize and harmonize information. The content, structure and scope were considered fit-for-purpose as evidenced by its use across federal agencies and international partners. However, there are a number of challenging areas that still need work. We continue to collaborate with our international partners to expand picklists in order to diversify sampling environments, presampling activities and experimental interventions, geographical locations, food products and the range of antimicrobial agents tested to represent more global data. We also continue to expand our fields and terms for different types of isolate characterization (e.g. growth conditions and metabolic phenotypes) and different available data types, as requested. In order to improve risk assessment information, we are currently modelling different types of “food production streams” i.e., all of the environments, inputs/outputs, products, equipment, stages and processes that participate in the production of different food commodities, in order to better articulate risk metrics and enhance risk vocabulary and data linkages.

A critical lesson learned from the development of ISO 23418:2022 (Microbiology of the food chain - Whole genome sequencing for typing and genomic characterization of bacteria - General requirements and guidance) was that for standards to be adopted, users need a clear path to implementation. These clear paths include easy-to-use tools to increase compliance and uptake, but also training and incentives to improve curator proficiency. A specification is only as good as the curator’s ability to use it in the real world. Feedback from DataHarmonizer testing included the need for more one:many relationships between samples and their derivatives. For instance, samples can yield many different types of organisms, there can be multiple isolates of the same organism, different isolates can be tested and characterized using different assays and drug panels, isolates can also be prepped and sequenced using different methods resulting in multiple sequences, etc. To reduce the burden of repetitive data entry (e.g. entering the same sample metadata for different isolates multiple times), one:many functionality is currently being engineered within the DataHarmonizer.

Canadian and international partners have also stressed the need for multilingual data curation and interchange capabilities (e.g. English to French, and vice versa). This functionality is currently under development in the DataHarmonizer, and discussions about how best to support multi-lingual data representations are ongoing within the LinkML and OBO Foundry semantic communities.

Data cleaning and integration can be incredibly time-consuming and resource-intensive processes that are often treated as an after-thought lacking adequate resources. Furthermore, the processes of data standards development and implementation benefit from the expertise and experience of semantics and curation specialists. There are a number of ongoing challenges associated with the lack of engagement of such specialists in routine as well as prioritized genomic surveillance initiatives. One such challenge is the paucity of well articulated best practices for data standards development and the necessary training for the wider community to implement them. As a result of developing the data standards ecosystem described above - for Canada, but also in collaboration with international consortia and partners - we are currently documenting best practices in order to democratize these processes and to build training materials to enable other labs to build their own fit-for-purpose specifications (manuscript in preparation). Putting standards into practice for the GAOH highlighted the importance of formalizing the role of semantics-trained standards developers and curators in collaborative projects. As a result, it is a recommendation by the authors that standards and curation be treated as a (data science) method, and that these activities be consistently incorporated in project design and treated as critical outcomes and deliverables that are appropriately acknowledged and funded.

Conclusions

The GRDI-AMR-One Health specification is currently being implemented by project partners, with a focus on a comprehensive integration within the Canadian federal genomics ecosystem. While designed for Canada, the data standard is compatible with other pathogen surveillance data standards, and is also readily adaptable for international use. As genomic data generation becomes increasingly essential for research, surveillance, and regulatory purposes, advances in sequencing technologies are producing vast amounts of data at lower costs. This data holds the potential for long-term reuse, possibly spanning hundreds of years, making its preservation and accessibility critical. With the advent of advanced AI tools capable of deriving new insights from this extensive data, the importance of robust and meticulously curated contextual information cannot be overstated. Prioritizing the collection and organization of this contextual data–not as an afterthought- is crucial to ensuring the continued utility and relevance of the genomic data over time.

Funding information

EJG, RC, CB, DD, AS, and NSJ were funded by Canadian Institutes of Health Research (CIHR) (PJT-159456) and Genome Canada (286 GET) grants to WWLH, as well as a Public Health Agency of Canada Contribution Agreement # 2223-HQ-000265. This work was also funded by the Government of Canada Shared Priority Projects, Genomics Research and Development Initiative - Antimicrobial Resistance (GRDI-AMR and GRDI-AMR-One-Health). AM & BW were funded by the Royal Society RGS\R1\231137 as part of the digital One Health pilot. LC acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/X020258/1), funded by the UK Medical Research Council (MRC). This UK funded award is carried out in the frame of the Global Health EDCTP3 Joint Undertaking.

Acknowledgements

The authors would like to thank Brittany Banks and Pascal Delaquis for their contributions to the specification and its implementation.

Author contributions

Conceptualization: E.J.G., C.D.C, W.W.L.H; Data curation: A.S., J.T., X.Y., A.R., R.W., C.S., S.H.C.E. B.A.W., A.M., L.C., E.N.T., S.T., T.K., D.D.N.S., R.Z., C.D.C., E.J.G.; Methodology: E.J.G., J.A.S., R.C., C.B., A.S., D.D., N.S.J., A.S., G.W., E.J., L.A.J., J.R., J.S., D.P.B., E.N.T., G.V.D., D.D.N.S., R.R-S., R.Z., C.L., C.D.C., W.W.L.H.; Project administration: E.J.G., C.D.C.; Supervision: E.J.G., C.D.C.; Writing - original draft: E.J.G., A.S.; Writing - review and editing: E.J.G., JA.S., R.C., C.B., A.S., D.D., N.S.J., A.S., G.W., E.J., L.A.J., J.R., J.S., D.P.B., J.T., X.Y., A.R., R.W., C.S., S.H.C.E., T.M., M.S.D., J.H.E.N., E.T., B.A.W., A.M., L.C., G.V.D., E.N.T., S.T., T.K., J.B., D.P-L., D.D.N.S., R.R-S., R.Z., C.L., C.D.C., W.W.L.H.

References

1. Cernava T, Rybakova D, Buscot F, Clavel T, McHardy AC, Meyer F, Meyer F, Overmann J, Stecher B, Sessitsch A, Schloter M, Berg G & The MicrobiomeSupport Team (2022). Metadata harmonization–Standards are the key for a better usage of omics data for integrative microbiome analysis. Environmental Microbiome 17:33 https://doi.org/10.1186/s40793-022-00425-1
2. Courtot M, Cherubin L, Faulconbridge A, Vaughan D, Green M, Richardson D, Harrison P, Whetzel PL, Parkinson H, Burdett T (2019). BioSamples database: an updated sample metadata hub. Nucleic Acids Research, 47(D1), D1172–D1178. https://doi.org/10.1093/nar/gky1061

3. Damerow JE, Varadharajan C, Boye K, Brodie EL, Burrus M, Chadwick KD, Crystal-Ornelas R, Elbashandy H, Alves RJE, Ely KS, Goldman AE, Haberman T, Hendrix V, Kakalia Z, Kemner KM, Kersting AB, Merino N, O'Brien F, Perzan Z, Robles E, Sorensen P, Stegen JC, Walls RL, Weisenhorn P, Zavarin M, Agarwal D (2021). Sample Identifiers and Metadata to Support Data Management and Reuse in Multidisciplinary Ecosystem Sciences. Data Science Journal, 20(11), 1–19. DOI: https://doi.org/10.5334/dsj-2021-011
4. Djordjevic SP, Jarocki VM, Seemann T. et al (2024). Genomic surveillance for antimicrobial resistance — a One Health perspective. Nat Rev Genet 25, 142–157 https://doi.org/10.1038/s41576-023-00649-y
5. Dugan VG, Emrich SJ, Giraldo-Calderón GI, et al.(2014). Standardized metadata for human pathogen/vector genomic sequences. PLoS One. 9(6):e99979.
6. EMBL-EBI Ontology Lookup Service (EBI-OLS), https://www.ebi.ac.uk/ols4, accessed February 2024
7. Field D, Garrity G, Gray T, et al. (2008). The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol. 26(5):541–7.
8. Gennari JH, Musen MA, Fergerson RW, Grosso WE, Crubézy M, Eriksson H, Noy NF, Tu SW (2003). The evolution of Protégé: an environment for knowledge-based systems development. International Journal of Human-Computer Studies. 58 (1): 89–123. doi:10.1016/S1071-5819(02)00127-1
9. Gill I, Griffiths EJ, Dooley D, Cameron R, Kallesøe SS, John NS, Sehar A, Gosal G, Alexander D, Chapel M, Croxen MA, Delisle B, Di Tullio R, Gaston D, Duggan A, Guthrie JL, Mark Horsman M, Joshi E, Kearny L, Knox N, Lau L, LeBlanc JJ, Li V, Lyons P, MacKenzie K, McArthur AG, Panousis EM, Palmer J, Prystajecky N, Smith KN, Tanner J, Townend C, Tyler A, Van Domselaar G, Hsiao WWL (2023). The DataHarmonizer: a tool for faster data harmonization, validation, aggregation and analysis of pathogen genomics contextual information. Microbial Genomics, 9 (1), https://doi.org/10.1099/mgen.0.000908
10. Gonçalves RS & Musen MA (2019). The variable quality of metadata about biological samples used in biomedical experiments. Sci. Data 6, 190021.
11. GRDI-AMR1, https://grdi.canada.ca/en/projects/antimicrobial-resistance-amr-project. Accessed July 12 2024
12. GRDI-AMR1, https://grdi.canada.ca/en/projects/antimicrobial-resistance-2-amr2-project. Accessed July 12 2024
13. Griffiths E, Dooley D, Graham M, Van Domselaar G, Brinkman FSL, Hsiao WWL (2017). Context Is Everything: Harmonization of Critical Food Microbiology Descriptors and Metadata for Improved Food Safety and Surveillance. Front. Microbiol., 25(8) https://doi.org/10.3389/fmicb.2017.01068
14. Griffiths EJ, Timme RE, Mendes CI, Page AJ, Alikhan NF, Fornika D, Maguire F, Campos J, Park D, Olawoye IB, Oluniyi PE, Anderson D, Christoffels A, da Silva AG, Cameron R, Dooley D, Katz LS, Black A, Karsch-Mizrachi I, Barrett T, Johnston A, Connor TR, Nicholls SM, Witney AA, Tyson GH, Tausch SH, Raphenya AR, Alcock B, Aanensen DM, Hodcroft E, Hsiao WWL, Vasconcelos ATR, MacCannell DR (2022). Future-proofing and maximizing the utility of metadata: The PHA4GE SARS-CoV-2 contextual data

specification package. Gigascience. 11:giac003. doi: 10.1093/gigascience/giac003. PMID: 35169842; PMCID: PMC8847733.

15. Griffiths EJ, Mendes I, Maguire F, Guthrie JL, Wee BA, Schmedes S, Holt K, Yadav C, Cameron R, Barclay C, Dooley D, MacCannell D, Chindelevitch L, Karsch-Mizrachi I, Waheed Z, Katz L, Petit III R, Dave M, Oluniyi P, Nasar MI, Raphenya A, Hsiao WWL, Timme RE (2024). PHA4GE quality control contextual data tags: standardized annotations for sharing public health sequence datasets with known quality issues to facilitate testing and training. Microb Genom. 10(6). doi: 10.1099/mgen.0.001260.
16. Jackson RC, Balhoff JP, Douglass E, Harris NL, Mungall CJ, Overton JA (2019). ROBOT: A Tool for Automating Ontology Workflows. BMC Bioinformatics 20(407). https://doi.org/10.1186/s12859-019-3002-3
17. Lambert D, Pightling A, Griffiths E, Van Domselaar G, Evans P, Berthelet S, Craig D, Chandry PS, Stones R, Brinkman F, Angers-Loustau A, Kreysa J, Tong W, Blais B (2017). Baseline Practices for the Application of Genomic Data Supporting Regulatory Food Safety. J AOAC Int. May 1;100(3):721-731. doi: 10.5740/jaoacint.16-0269.
18. Mussen, MA (2022). Demand standards to sort FAIR from foul. Nature 609(222)
19. OBO Foundry, https://obofoundry.org/. Accessed July 12 2024
20. Ong E, Xiang Z, Zhao B, Liu Y, Lin Y, Zheng J, Mungall C, Courtot M, Ruttenberg A, He Y (2017). Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. Nucleic Acids Research. 45 (D1), D347–D352, https://doi.org/10.1093/nar/gkw918
21. Pettengill JB, Beal J, Balkey M, Allard M, Rand H, Timme R (2021). Interpretative Labor and the Bane of Nonstandardized Metadata in Public Health Surveillance and Food Safety. Clin Infect Dis.;73(8):1537-1539. doi: 10.1093/cid/ciab615
22. Public Health Agency of Canada (2015). Federal Action Plan On Antimicrobial Resistance And Use In Canada Building On The Federal Framework For Action. https://www.canada.ca/en/public-health/services/publications/drugs-health-products/pan-canadian-action-plan-antimicrobial-resistance.html. Accessed August 15 2024.
23. Public Health Agency of Canada (2023). Pan-Canadian Action Plan on Antimicrobial Resistance.https://www.canada.ca/en/public-health/services/publications/drugs-health-products/pan-canadian-action-plan-antimicrobial-resistance.html. Accessed August 15 2024.
24. Redman-White CJ, Loosli K, Qarkaxhija V, Lee TM, Mboowa G, Wee BA, Muwonge A (2023). A Digital One Health framework to integrate data for public health decision-making. IJID One Health, 1(100012) https://doi.org/10.1016/j.ijidoh.2023.100012.
25. Schriml LM, Chuvochina M, Davies N, Eloe-Fadrosh EA, Finn RD, Hugenholtz P, Hunter CI, Hurwitz BL, Kyrpides NC, Meyer F, Karsch-Mizrachi I, Sansone S-A, Sutton G, Tighe S, Walls R (2020). COVID-19 pandemic reveals the peril of ignoring metadata standards. Scientific Data. 7:188. https://doi.org/10.1038/s41597-020-0524-5
26. Sielemann K, Hafner A, Pucker B (2020). The reuse of public datasets in the life sciences: potential risks and rewards. PeerJ, DOI 10.7717/peerj.9954
27. Smith B, Ashburner M, Rosse C, et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 25(11):1251–5.

28. Timme RE, Karsch-Mizrachi I, Waheed Z, Arita M, MacCannell D, Maguire6 F, Petit R III, Page AJ, Mendes CI, Nasar MI, Oluniyi P, Tyler AD, Raphenya AR, Guthrie JL, Olawoye I, Rinck G, O'Cathail C, Lees J, Cochrane G, Cummins C, J. Brister R, Klimke W, Feldgarden M, Griffiths E (2023). Putting everything in its place: using the INSDC compliant Pathogen Data Object Model to better structure genomic data submitted for public health applications. MGen 9:12. https://doi.org/10.1099/mgen.0.001145
29. Tyson S, Peterson C-L, Olson A, Tyler S,Knox N, Griffiths E, Dooley D, Hsiao W, Cabral J, Johnson RP, Laing C, Gannon V, Lynch T, VanDomselaar G, Brinkman F, Graham M (2019). Eleven high-quality reference genome sequences and 360 draft assemblies of Shigatoxin-producing Escherichia coli isolates from human, food, animal, and environmental sources in Canada. Microbiol Resour Announc 8:e00625-19. https://doi.org/10.1128/MRA.00625-19.
30. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen MA (2011). BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Research. 39 (2), W541–W545, https://doi.org/10.1093/nar/gkr469
31. Yilmaz P, Kottmann R, Field D, et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. 29(5):415–20.
32. Zeb A, Soininen J-P, Sozer N (2021). Data harmonisation as a key to enable digitalisation of the food sector: A review. Food and BioProducts Processing. 127, 360–370

Figure Legends

Figure 1: Structure of the GRDI-AMR specification. The GRDI-AMR specification is structured according to an ISO-based framework in which standardized fields are grouped together into thematic, interoperable modules. Modules can be reused in different combinations to create new specifications - and modules can be enriched/depleted with different fields and picklists to enable customization. Modules are populated using vocabulary sourced from OBO Foundry ontologies and existing standards. Thematic modules included in the specification are listed.

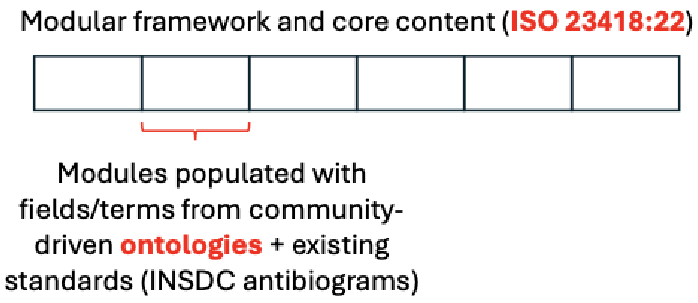
Figure 2: Enabling interoperability between data specifications. While Canada and the US implement different approaches to structuring their One Health contextual data standards, the common use of ontologies and collaborative development during a joint hackathon facilitated mapping and creation of a data interchange format. An example of mapping illustrating the relationship between environmental site fields is presented.

Figure 3: Data curation ecosystem. A well-established data curation ecosystem formalizes data standards development and the role of curation in genomics projects and initiatives. It also operationalizes standards through the development of software and tools that are well maintained. Curators using the tools should be provided with training, and also given opportunities to inform further standards and tool development.

Figures and Tables

Figures

Figure 1



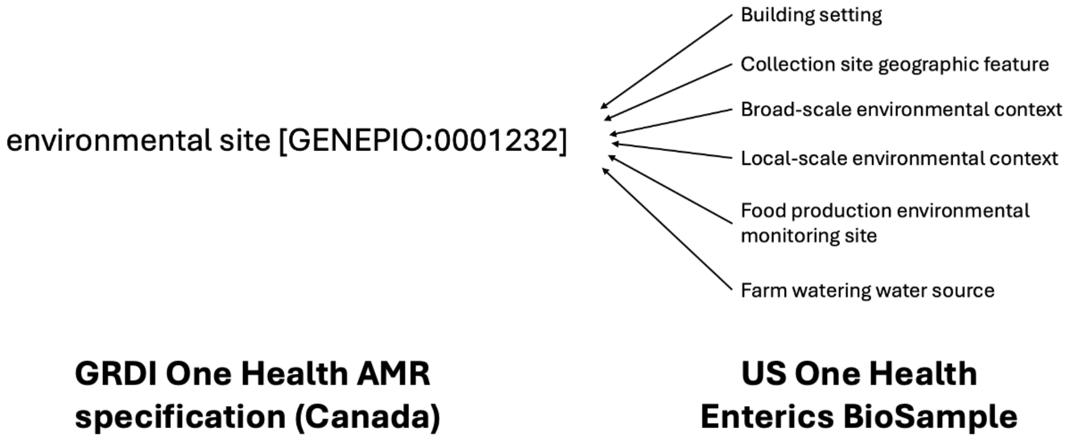
Modular framework and core content (**ISO 23418:22**)

Modules populated with fields/terms from community-driven **ontologies** + existing standards (INSDC antibiograms)

Thematic Modules

- Sample collection and processing
- Host information
- Strain and isolate information
- Environmental conditions & measurements
- Sequencing methods
- Bioinformatics and quality control metrics
- Taxonomic identification information
- AMR phenotypic testing
- Risk assessment information
- Public repository information

Figure 2



environmental site [GENEPIO:0001232]

Building setting

Collection site geographic feature

Broad-scale environmental context

Local-scale environmental context

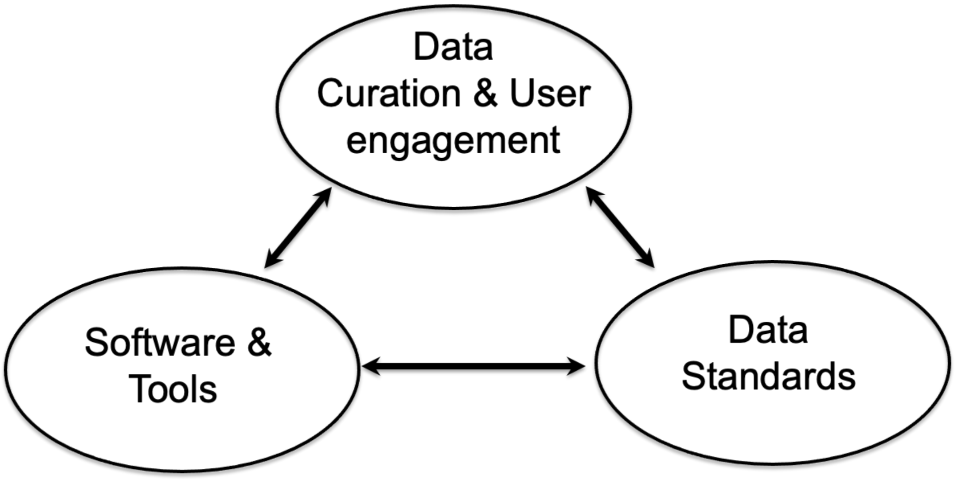
Food production environmental monitoring site

Farm watering water source

GRDI One Health AMR specification (Canada)

US One Health Enterics BioSample

Figure 3



Tables

Table 1: OBO Foundry best practices implemented in GRDI specification development

Best Practice	**Benefit**
Openness	openly available to the community for use
Common Format	common formal language in an accepted concrete syntax
Universal and Unique Identifier (URI) Space	use unique IRIs in the form of an OBO Foundry permanent URL (PURL)
Versioning	developers provide documented procedures for versioning the ontology, different versions of ontology are marked, stored, and officially released
Defined Scope	has a clearly specified scope and content that adheres to that scope
Textual Definitions	every term is accompanied by a definition

Consistent logical relations	logical axioms based on prescribed logical relations from the Relations Ontology (RO)
Documentation	owners of an ontology should strive to provide as much documentation as possible
Plurality of Users	documentation indicating use by multiple independent people or organizations indicating reuse and consensus
Commitment To Collaboration	development, in common with many other standards-oriented scientific activities, should be carried out in a collaborative fashion
Locus of Authority	designated representative responsible for communications between the community, other ontology developers, and the Foundry, to ensure feedback is incorporated as well as scientific advancement
Naming Conventions	labels for fields/terms must be unique, and intelligible to scientists and amenable to natural language processing
Maintenance	ontologies should be maintained and evolve over time to reflect changes in scientific consensus
Responsiveness	ontology developers MUST offer channels for community participation and SHOULD be responsive to requests

Table 2: INSDC antibiogram mappings and ontology equivalents used in the GRDI-AMR specification. Required fields are highlighted in yellow.

Ontologized Antibiogram Equivalents	**ENA Antibiogram Fields**	**NCBI Antibiogram Fields**
biosample_accession [GENEPIO:0001139]	bioSample_ID	bisample_accession
antimicrobial_agent_name [GENEPIO:0100521]	antibiotic_name	antibiotic
AMR_phenotype [GENEPIO:0100525]	resistance_phenotype	resistance_phenotype
AMR_measurement_sign [GENEPIO:0100524]	measurement_sign	measurement_sign
AMR_measurement [GENEPIO:0100522]	measurement	measurement
AMR_measurement_units [GENEPIO:0100523]	measurement_units	measurement_units
AMR_laboratory_typing_method [GENEPIO:0100526]	laboratory_typing_method	laboratory_typing_method

AMR_laboratory_typing_platform [GENEPIO:0100527]	platform	laboratory_typing_platform
AMR_laboratory_typing_platform_version [GENEPIO:0100528]	No equivalent	laboratory_typing_method_version_or_reagent
AMR_testing_standard [GENEPIO:0100530]	ast_standard	testing_standard
AMR_testing_standard_version [GENEPIO:0100531]	No equivalent	No equivalent
AMR_testing_standard_details [GENEPIO:0100520]	No equivalent	No equivalent
AMR_testing_susceptible_breakpoint [GENEPIO:0100516]	No equivalent	No equivalent
AMR_testing_intermediate_breakpoint [GENEPIO:0100517]	No equivalent	No equivalent
AMR_testing_resistant_breakpoint [GENEPIO:0100518]	No equivalent	No equivalent
No equivalent	breakpoint_version	No equivalent
organism [GENEPIO:0001191]	species	In BioSample
AMR_vendor_name [GENEPIO:0100529]	No equivalent	vendor
AMR_testing_by [GENEPIO:0100511]	No equivalent	No equivalent
AMR_testing_by_laboratory_name [GENEPIO:0100512]	No equivalent	No equivalent
AMR_testing_by_contact_name [GENEPIO:0100513]	No equivalent	No equivalent
AMR_testing_by_contact_email [GENEPIO:0100514]	No equivalent	No equivalent
AMR_testing_date [GENEPIO:0100515]	No equivalent	No equivalent

Table 3: List of OBO Foundry Ontologies implemented in the GRDI specification.

Ontology Name	**Ontology Scope**
Genomic Epidemiology Ontology (GenEpiO)	describes the genomics, laboratory, clinical and epidemiological contextual information required to support data sharing and integration for foodborne infectious disease surveillance and outbreak investigations

Food Ontology (FoodOn)	describes parts of animals, plants, and fungi which can bear a food role for humans and domesticated animals, as well as derived food products and the processes used to make them
Ontology of Biological Investigations (OBI)	describes experimental designs, protocols, instrumentation, materials, generated data, and types of analysis
Environment Ontology (ENVO)	describes built and natural environments
Antibiotic Resistance Ontology (ARO)	describes antibiotic resistance genes and mutations, their products, mechanisms, and associated phenotypes, as well as antibiotics and their molecular targets
Chemical Entities of Biological Interest Ontology (ChEBI)	describes small chemical compounds
Agriculture Ontology (AGRO)	describes agronomic practices, techniques, and variables used in agronomic experiments
Health Surveillance Ontology (HSO)	describes surveillance system level data including outputs from surveillance activities and aims to support One-Health surveillance, covering animal health, public health and food safety surveillance
Environment Ontology for Livestock (EOL)	describes the feeding modalities, the environment, and the structure of livestock farms and rearing systems
NCBI Taxonomy Ontology (NCBITaxon)	describes curated classifications and nomenclature for all organisms in the NCBI public sequence repository
Uber Anatomy Ontology (UBERON)	describes cross-species anatomical parts and materials
Gender, Sex, and Sex Orientation Ontology (GSSO)	describes gender identity, gender expression, romantic identity, sexual identity, sexual orientation, sexual behavior, sexual abuse, and related topics
Biological Spatial Ontology (BSPO)	describes biological spatial concepts, anatomical axes, gradients, regions, planes, sides and surfaces
Units Ontology (UO)	describes units of measurement
Phenotype Ontology (PATO)	describes biological phenotypes such as qualities, properties, attributes or characteristics
Gazetteer Ontology (GAZ)	describes geographical locations and entities
Ontology for Biobanking (OBIB)	describes activities, contents, and administration of a biobank

National Cancer Institute Thesaurus (NCIT)	describes cancer related diseases, findings and abnormalities; anatomy; agents, drugs and chemicals; genes and gene products

Table 4: Required fields for samples, isolates and sequencing methods. Field labels, as well as ontology IDs, definitions, curator guidance, and examples of expected values are provided.

Field	Ontology ID	Definition	Guidance	Example Value
sample_collector_sample_ID	GENEPIO:0001123	The user-defined name for the sample.	The sample_ID should represent the identifier assigned to the sample at time of collection, for which all the descriptive information applies. If the original sample_ID is unknown or cannot be provided, leave blank or provide a null value.	ABCD123
alternative_sample_ID	GENEPIO:0100427	An alternative sample_ID assigned to the sample by another organization.	Alternative identifiers assigned to the sample should be tracked along with original IDs to establish chain of custody. Alternative sample IDs should be provided in the in a prescribed format which consists of the ID followed by square brackets (no space in between the ID and bracket) containing the short form of ID provider's agency name i.e. ID[short organization code]. Agency short forms include the following: Public Health Agency of Canada: PHAC Canadian Food Inspection Agency: CFIA Agriculture and Agri-Food Canada: AAFC Fisheries and Oceans Canada: DFO Environment and Climate Change Canada: ECCC Health Canada: HC Multiple identifiers can be provided and separated by semicolons. If the information is unknown or cannot be provided, leave blank or provide a null value.	ABCD1234[PHAC]; 12345rev[CFIA]
sample_collected_by	GENEPIO:0001153	The name of the agency, organization or institution with which the sample collector is affiliated.	Provide the name of the agency, organization or institution that collected the sample in full (avoid abbreviations). If the information is unknown or cannot be provided, leave blank or provide a null value.	Public Health Agency of Canada (PHAC) [GENEPIO:0100551]
sample_collector_contact_email	GENEPIO:0001156	The email address of the contact responsible for follow-up regarding the sample.	Provide the email associated with the listed contact. As personnel turnover may render an individual's email obsolete, it is more preferable to provide an address for a position or lab, to ensure accuracy of information and institutional memory. If the information is unknown or cannot be provided, leave blank or provide a null value.	johnnyblogs@lab.ca
purpose_of_sampling	GENEPIO:0001198	The reason that the sample was collected.	The reason a sample was collected may provide information about potential biases in sampling strategy. Provide the purpose of sampling from the picklist in the template. Most likely, the sample was	Surveillance [GENEPIO:0100004]

			collected for Diagnostic testing. The reason why a sample was originally collected may differ from the reason why it was selected for sequencing, which should be indicated in the "purpose of sequencing" field.	
geo_loc_name (country)	GENEPIO:0001181	The country of origin of the sample.	Provide the name of the province/state/region where the sample was collected. If the information is unknown or cannot be provided, provide a null value.	Canada [GAZ:00002560]
geo_loc_name (state/province/region)	GENEPIO:0001185	The state/province/territory of origin of the sample.	Provide the name of the province/state/region where the sample was collected. If the information is unknown or cannot be provided, provide a null value.	British Columbia [GAZ:00002562]
sample_collection_date	GENEPIO:0001174	The date on which the sample was collected.	Provide the date according to the ISO 8601 standard "YYYY-MM-DD", "YYYY-MM" or "YYYY".	2020-10-30
sample_collection_date_precision	GENEPIO:0001177	The precision to which the "sample collection date" was provided.	Provide the precision of granularity to the "day", "month", or "year" for the date provided in the "sample collection date" field. The "sample collection date" will be truncated to the precision specified upon export; "day" for "YYYY-MM-DD", "month" for "YYYY-MM", or "year" for "YYYY".	day [UO:0000033]
isolate_ID	GENEPIO:0100456	The user-defined identifier for the isolate, as provided by the laboratory that originally isolated the isolate.	Provide the isolate_ID created by the lab that first isolated the isolate (i.e. the original isolate ID). If the information is unknown or cannot be provided, leave blank or provide a null value. If only an alternate isolate ID is known (e.g. the ID from your lab, if your lab did not isolate the isolate from the original sample), make asure to include it in the alternative_isolate_ID field.	SA20131043
IRIDA_isolate_ID*	GENEPIO:0100459	The identifier of the isolate in the IRIDA platform.	Provide the "sample ID" used to track information linked to the isolate in IRIDA. IRIDA sample IDs should be unqiue to avoid ID clash. This is very important in large Projects, especially when samples are shared from different organizations. Download the IRIDA sample ID and add it to the sample data in your spreadsheet as part of good data management practices.	GRDI_LL_12345
IRIDA_project_ID*	GENEPIO:0100460	The identifier of the Project in the iRIDA platform.	Provide the IRIDA "project ID".	666
organism	GENEPIO:0001191	Taxonomic name of the organism.	Put the genus and species (and subspecies if applicable) of the bacteria, if known. The standardized term can be sourced from this look-up service: https://www.ebi.ac.uk/ols/ontologies/ncbitaxon.	Salmonella enterica subsp. enterica [NCBITaxon:59201]
sequenced_by	GENEPIO:0100416	The name of the agency,	Provide the name of the agency, organization or institution that performed the sequencing in full	Public Health Agency of Canada

		organization or institution responsible for sequencing the isolate's genome.	(avoid abbreviations). If the information is unknown or cannot be provided, leave blank or provide a null value.	(PHAC) [GENEPIO:0100551]
sequenced_by_contact_email	GENEPIO:0100422	The email address of the contact responsible for follow-up regarding the sequence.	Provide the email associated with the listed contact. As personnel turnover may render an individual's email obsolete, it is more preferable to provide an address for a position or lab, to ensure accuracy of information and institutional memory. If the information is unknown or cannot be provided, leave blank or provide a null value.	enterics@lab.ca
purpose_of_sequencing	GENEPIO:0001445	The reason that the sample was sequenced.	Provide the reason for sequencing by selecting a value from the following pick list: Diagnostic testing, Surveillance, Monitoring, Clinical trial, Field experiment, Environmental testing. If the information is unknown or cannot be provided, leave blank or provide a null value.	Research [GENEPIO:0100003]
AMR_testing_by	GENEPIO:0100511	The name of the organization that performed the antimicrobial resistance testing.	Provide the name of the agency, organization or institution that performed the AMR testing, in full (avoid abbreviations). If the information is unknown or cannot be provided, leave blank or provide a null value.	Canadian Food Inspection Agency (CFIA) [GENEPIO:0100552]
AMR_testing_by_contact_name	GENEPIO:0100513	The name of the individual or the individual's role in the organization that performed the antimicrobial resistance testing.	Provide the name of an individual or their job title. As personnel turnover may render the contact's name obsolete, it is more preferable to provide a job title for ensuring accuracy of information and institutional memory. If the information is unknown or cannot be provided, leave blank or provide a null value.	Enterics Lab Manager
AMR_testing_by_contact_email	GENEPIO:0100514	The email of the individual or the individual's role in the organization that performed the antimicrobial resistance testing.	Provide the email associated with the listed contact. As personnel turnover may render an individual's email obsolete, it is more preferable to provide an address for a position or lab, to ensure accuracy of information and institutional memory. If the information is unknown or cannot be provided, leave blank or provide a null value.	johnnyblogs@lab.ca
AMR_measurement	GENEPIO:0100522	The measured value of amikacin resistance.	This field should only contain a number (either an integer or a number with decimals).	4
AMR_measurement_units	GENEPIO:0100523	The units of the antimicrobial resistance measurement.	Select the units from the pick list provided. Use the Term Request System to request the addition of other units if necessary.	ug/mL [UO:0000274]
AMR_measu	GENEPIO:0100	The qualifier	Select the comparator sign from the pick list	greater than (>)

rement_sign	524	associated with the antimicrobial resistance measurement	provided. Use the Term Request System to request the addition of other signs if necessary.	[GENEPIO:0001006]

Supplementary Table 1

US One Health Enteric Package Field	**GRDI-AMR Specification Field**
Sample Identifiers	
sample_name	sample_collector_sample_ID
sample_title	No match
bioproject_accession	bioproject_accession
strain	strain; isolate_ID
isolate_name_alias	alternative_isolate_ID
culture_collection	No match
reference_material	No match
Sample/isolate collection information	
organism	organism
collected_by	sample_collected_by
collection_date	sample_collection_date
cult_isol_date	isolation_date
geo_loc_name	geo_loc_name (country):geo_loc_name (state/province/region)
isolation_source	host (scientific name); environmental_site; environmental_material; anatomical_material; body_product; anatomical_part; food_product; food_product_properties; collection_device; collection_method No match (required so user must add a descriptor from this list)

source_type	IF Host (scientific name) is Homo Sapiens, THEN Human
	IF Food_product, THEN Food
	IF host (scientific name) is NOT Homo Sapiens, THEN

	Animal IF host (scientific name) is Homo Sapiens, THEN Human
samp_collect_device	collection_device
purpose_of_sampling	purpose_of_sampling
project_name	sample_collection_project_name
ifsac_category	No match
lat_lon	geo_loc latitude | geo_loc longitude
serotype	No match
serovar	serovar
sequenced_by	sequenced_by
description	No match
Human/animal host	
host	host (scientific name) OR host (common name) OR host (food production name)
host_sex	No match
host_age	host_age_bin
host_disease	host_disease
host_subject_id	No match
animal_env	environmental_site
host_tissue_sampled	anatomical_material
host_body_product	body_product
host_variety	host (ecotype)
host_animal_breed	host (breed)
upstream_intervention	experimental_intervention
host_am	IF presampling_activity has some: Antimicrobial pre-treatment [GENEPIO:0100537])"; THEN presampling_activity_details

host_group_size	No match
host_housing	IF environmental site, and has some: Animal cage [ENVO:01000922], Aquarium [ENVO:00002196], Building [ENVO:00000073], Barn [ENVO:03501257], Breeder barn [ENVO:03501383], Broiler barn [ENVO:03501386], Sheep barn [ENVO:03501385], Pigsty [ENVO:03501413], Animal pen [ENVO:03501387], Stall [EOL:0001903], Poultry hatchery [ENVO:01001874], Roost (bird) [ENVO:03501439], Crate [ENVO:03501372]
Food samples	
food_origin	food_product_origin_geo_loc_name (country)
intended_consumer	No match
spec_intended_cons	No match
food_source	animal_source_of_food
food_processing_method	food_product_properties: Food (canned) [FOODON:00002418], Food (cooked) [FOODON:00001181], Food (cut) [FOODON:00004291], Food (chopped) [FOODON:00002777], Food (chunks) [FOODON:00004555], Food (cubed) [FOODON:00004278], Food (diced) [FOODON:00004549], Food (grated) [FOODON:00004552], Food (sliced) [FOODON:00002455], Food (shredded) [FOODON:00004553], Food (fresh) [FOODON:00002457], Food (pulped) [FOODON:00004554], Food (raw) [FOODON:03311126], Food (unseasoned) [FOODON:00004287], Meat (boneless) [FOODON:00003467], Meat (skinless) [FOODON:00003468], Meat (with bone) [FOODON:02010116], Meat (with skin) [FOODON:02010111]
food_preserv_proc	food_product_properties: Food (canned), Food (dried) [FOODON:03307539], Food (frozen) [FOODON:03302148]
food_prod	food_product_production_stream
label_claims	label_claim
food_product_type	food_product

food_industry_code	No match
food_industry_class	No match
food_additive	No match
food_contact_surf	No match
food_contain_wrap	food_packaging
food_pack_medium	No match
food_pack_integrity	No match
food_quality_date	food_quality_date
food_prod_synonym	No match
Food facility, built environment	
facility_type	IF environmental site, and has some: Abattoir [ENVO:01000925], Dairy [ENVO:00003862], Farm [ENVO:00000078], Hatchery [ENVO:01001873], Retail environment [ENVO:01001448], Shop [ENVO:00002221], Butcher shop [ENVO:03501396], Supermarket [ENVO:01000984], Manure digester facility [ENVO:03501422]
building_setting	No match
coll_site_geo_feat	IF environmental material, and has some: Animal transportation equipment [AGRO:00000671], Dead haul trailer [GENEPIO:0100896], Dead haul truck [AGRO:00000673], Live haul trailer [GENEPIO:0100897], Live haul truck [AGRO:00000674], Bulk tank [ENVO:03501379], Animal feeding equipment [AGRO:00000675], Animal feeder [AGRO:00000679], Animal drinker [AGRO:00000680], Feed pan [AGRO:00000676], Watering bowl [AGRO:00000677], "Belt [NCIT:C49844], Boot [GSSO:012935], Boot cover [OBI:0002806], Broom [ENVO:03501431], Bulk tank [ENVO:03501379], Chick box [AGRO:00000678], Chick pad [AGRO:00000672], Cleaning equipment [ENVO:03501430], Dumpster [ENVO:03501400], Egg belt [AGRO:00000670], Fan [NCIT:C49947], Freezer [ENVO:03501415], Freezer handle [ENVO:03501414], Plucking belt [AGRO:00000669]
food_type_processed	No match

location_in_facility	No match
env_monitoring_zone	No match
indoor_surf	No match
indoor_surf_subpart	No match
surf_material	No match
material_condition	No match
surface_orientation	No match
surf_temp	No match
biocide_used	No match
animal_intrusion	No match
Environment (farm/water/natural env)	
env_broad_scale	No match
env_local_scale	IF environmental site, and has some: Agricultural Field [ENVO:00000114], Alluvial fan [ENVO:00000314], Artificial wetland [ENVO:03501406], Breeding ground [ENVO:03501441], Creek [ENVO:03501405], Farm [ENVO:00000078], Beef farm [ENVO:03501443], Breeder farm [ENVO:03501384], Dairy farm [ENVO:03501416], Feedlot [ENVO:01000627], Beef cattle feedlot [ENVO:03501444], Fish farm [ENVO:00000294], Research farm [ENVO:03501417], Freshwater environment [ENVO:01000306], Hatchery [ENVO:01001873], Poultry hatchery [ENVO:01001874], Lake [ENVO:00000020], Manure lagoon (Anaerobic lagoon) [ENVO:03501423], Manure pit [ENVO:01001872], Marine environment [ENVO:01000320], Benthic zone [ENVO:03501440], Pelagic zone [ENVO:00000208], Park [ENVO:00000562], Pond [ENVO:00000033], Reservoir [ENVO:00000025], Irrigation reservoir [ENVO:00000450], River [ENVO:00000022], Roost (bird) [ENVO:03501439], Rural area [ENVO:01000772], Slough [ENVO:03501438], Stream [ENVO:00000023], Tributary [ENVO:00000495], Water surface [ENVO:01001191], Woodland area [ENVO:00000109]
env_medium	environmental material EXCEPT: Animal transportation equipment [AGRO:00000671], Dead haul trailer [GENEPIO:0100896], Dead haul truck

	[AGRO:00000673], Live haul trailer [GENEPIO:0100897], Live haul truck [AGRO:00000674], Bulk tank [ENVO:03501379], Animal feeding equipment [AGRO:00000675], Animal feeder [AGRO:00000679], Animal drinker [AGRO:00000680], Feed pan [AGRO:00000676], Watering bowl [AGRO:00000677], "Belt [NCIT:C49844], Boot [GSSO:012935], Boot cover [OBI:0002806], Broom [ENVO:03501431], Bulk tank [ENVO:03501379], Chick box [AGRO:00000678], Chick pad [AGRO:00000672], Cleaning equipment [ENVO:03501430], Dumpster [ENVO:03501400], Egg belt [AGRO:00000670], Fan [NCIT:C49947], Freezer [ENVO:03501415], Freezer handle [ENVO:03501414], Plucking belt [AGRO:00000669]
plant_growth_med	No match
plant_water_method	No match
rel_location	No match
soil_type	No match
farm_water_source	No match
fertilizer_admin	IF presampling_activity has some "Fertilizer pre-treatment [GENEPIO:0100543]"; THEN presampling_activity_details
food_clean_proc	No match
sanitizer_used_postharvest	No match
farm_equip	No match
extr_weather_event	No match
mechanical_damage	No match