Fixing the plumbing: Building interoperability between wastewater genomic surveillance datasets and systems using the PHA4GE contextual data specification

Authors

Jillian S. Paull[1,2]*, Charlotte Barclay[3]*, Rhiannon Cameron[3], Damion Dooley[3], Ivan Gill[3], Dilip Abraham[4], Remigio Arteaga[5], Julian Carrillo-Reyes[6], Barbara Ghiglione[7,8], María Sol Haim[9], John Juma[10], Waqasuddin Khan[11], Iqra Saleh[11], Muhammad Imran Nisar[11], Hongo Koffi Anderson[12], Rintu Kutum[13], Aadya R. Sood[14], Victor V. Mabasa[15], Arsene Djoko Nono[16], Kevine Makuetche[16], Armelle Megueya[16], Michael Owusu[17], Berhanu Yitayew[18], Pierrette Simo Tchuinte[16], Sarmila Tandukar[19], Masood Ur Rehman Kayani[20], Laasya Samhita[14], David Blazes[21], Simon R. Harris[21], Ilene Karsch-Mizrachi[22], Linda M Frisse[22], Anjanette Johnston[22], Emily Clough[22], Christopher O'Sullivan[22], Jonathan Trow[22], J. Rodney Brister[22], Ravinder Pannu[22], Jennifer Ali[23], Peter Woollard[24], Mathew Fisher[23a], Colman O'Cathail[24], Chrystal Landgraff[23], Thomas R. Connor[25], Emily A. Smith[26], Mathew Thomson[27], Douglas G. Manuel[27], Melanie Courtot[28], Ines Mendes[29], Joshua I. Levy[30], Finlay Maguire[31], Ruth E. Timme[32], William W.L. Hsiao[3], Emma J. Griffiths[3].

* These authors contributed equally.
[a] Current affiliation: Canadian Food Inspection Agency, Winnipeg, MB, Canada
Corresponding author: Emma Griffiths (ega12@sfu.ca)

Keywords

Wastewater genomic surveillance, contextual data, metadata, data standards, interoperability, data harmonization

Affiliations

[1] Department of Systems Biology, Harvard Medical School, Massachusetts, United States of America
[2] The Broad Institute of Massachusetts Institute of Technology (MIT) and Harvard, Massachusetts, United States of America
[3] Centre for Infectious Disease Genomics and One Health, Faculty of Health Sciences, Simon Fraser University, British Columbia, Canada
[4] Christian Medical College Vellore, Tamil Nadu, India

[5] Laboratorio para Investigaciones Biomédicas, Escuela Superior Politécnica del Litoral, ESPOL, Guayaquil, Ecuador
[6] Unidad Académica Juriquilla, Instituto de Ingeniería, Universidad Nacional Autónoma de México, Querétaro, Mexico
[7] Laboratorio de Resistencia Bacteriana, Instituto de Bacteriología y Virología Molecular (IBaViM), Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina
[8]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Ciudad Autónoma de Buenos Aires, Argentina
[9] Unidad Operativa Centro Nacional de Genómica y Bioinformática—ANLIS “Dr. Carlos G. Malbrán, Ciudad Autónoma de Buenos Aires, Argentina
[10] International Livestock Research Institute, Nairobi, Kenya
[11] CITRIC Center for Bioinformatics and Computational Biology, Department of Pediatrics and Child Health, The Aga Khan University, Karachi-74800, Pakistan
[12] Universite Nangui Abrogoua, Abidjan, Côte d’Ivoire
[13] Department of Computer Science, Trivedi School of Biosciences, Koita Centre for Digital Health, Ashoka University, Haryana, India
[14] Department of Biology, Ashoka University, Haryana, India
[15] National Institute for Communicable Disease, Gautang, South Africa
[16] Centre Pasteur du Cameroun, Centre Region, Cameroon
[17] Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[18] Armauer Hansen Research Institute and Debre Berhan University, Addis Ababa, Ethiopia
[19] Organization for Public Health and Environment Management, Lalitpur, Nepal
[20] Metagenomics Discovery Lab, School of Interdisciplinary Engineering and Sciences, National University of Sciences and Technology, Islamabad, Pakistan
[21] Gates Foundation, Washington, United States of America
[22] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Maryland, United States of America
[23] Public Health Agency of Canada, Manitoba, Canada
[24] European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridgeshire, United Kingdom
[25] Public Health Wales NHS Trust, Wales, United Kingdom
[26] University of Minnesota Center for Infectious Disease Research and Policy, Minnesota, United States of America
[27] Ottawa Hospital Research Institute, Ontario, Canada
[28] Ontario Institute for Cancer Research and University of Toronto, Ontario, Canada
[29] Theiagen Genomics LLC, Colorado, United States of America
[30] Department of Immunology and Microbiology, The Scripps Research Institute, California, United States of America
[31] Department of Community Health & Epidemiology and Faculty of Computer Science, Dalhousie University, Nova Scotia, Canada
[32] Human Foods Program, US Food and Drug Administration, Maryland, United States of America

Abstract

The evolution of wastewater genomic surveillance (WWGS) has led to the development of many new methodologies, allowing for the broad application of WWGS for detection and monitoring of diverse pathogens and genetic markers. Variability in techniques and approaches creates challenges for data integration and interoperability that hinder analyses necessary for public health insights. Here, the Public Health Alliance for Genomic Epidemiology (PHA4GE) – in collaboration with scientists and stakeholders from over 20 countries, as well as global data repositories – presents a wastewater contextual data specification package relevant for a wide array of public health and research use cases. The PHA4GE wastewater contextual data specification is an ISO-compatible, ontology-based, modular data standard that is implemented by a free, open source data curation and validation tool called the DataHarmonizer. To facilitate interoperability and data sharing, interchange formats and instructions for automated transformations are included among the package's supporting documentation. The specification package is part of a growing library of interoperable pathogen/target-specific standards designed upon a shared framework using semantic best practices. We hope that this standard will not only aid in the implementation of WWS, but also serve as an exemplar for the development of related data standards, such as for other environmental use cases or other metagenomic surveillance efforts.

Introduction

The definition and composition of "wastewater" is highly variable, making it difficult to provide a singular definition. Wastewater has traditionally been defined as sewage, but is increasingly defined as any water that has been affected by domestic, industrial, agricultural and commercial use, for example domestic "grey water", surface runoff, stormwater, sewer inflow or sewer infiltration, etc (Tchobanoglous et al., 2003). Microbiological contamination of water is a phenomenon with impacts on both human and environmental ecosystem health. When pathogens are excreted or shed by humans and animals (e.g. via gastrointestinal tracts, dermal shedding, etc), these microorganisms can often be detected in wastewater (Singh et al., 2024). Given that wastewater surveillance can provide early, unbiased indicators of disease spread at relatively low costs, it has become an important element of pathogen surveillance frameworks (Singh et al., 2024). Polio, for example, was first isolated and analyzed from urban sewage in the 1940s (Paul et al., 1940). Since then, genomic surveillance of wastewater (WWGS) has become an essential keystone in the continuing effort to eradicate poliovirus (Asghar et al., 2014), and has been leveraged to detect other viral pathogens, bacterial pathogens, antimicrobial resistance genes, and intestinal parasites since the 2000s (Harrington et al., 2022; Keshaviah et al., 2023; Maya et al., 2006; Sengupta et al., 2019; Wolfe et al., 2022; Wurtzer et al., 2022).

WWGS was a critical tool for monitoring SARS-CoV-2 introductions, as well as viral evolution, prevalence, and diversity during the COVID-19 pandemic. Since the first wastewater samples were analyzed for the presence of SARS-CoV-2 in early 2020, significant investment has been made in the implementation and development of technologies and software used to collect, process, and analyze wastewater samples (Hayes et al., 2021; Kayikcioglu et al., 2023; Kilaru et al., 2020; Lott et al., 2023; Nemudryi et al., 2020). Currently, over 40 countries monitor SARS-CoV-2 via WWGS programs; these programs monitor on average about 40% of countries' populations (Keshaviah et al., 2023). All these developments and investments in SARS-CoV-2 WWGS can aid the development of broader surveillance strategies for wastewater samples. Just as WWGS of SARS-CoV-2 and polio helped guide policy decisions, surveillance for other pathogens can also meaningfully impact public health (de Jong, 2021; Petros et al., 2022).

Genomic surveillance of pathogens requires not only high quality sequence data, but also well-structured contextual data (e.g., laboratory, clinical, epidemiological, and environmental information) to make sense of analytical results and support decision-making. WWGS often involves data streams originating from different sources and information management systems, posing challenges for data harmonization, integration, and meaningful interpretation (Gonçalves and Musen, 2019; Musen, 2022; Pettengill et al., 2021; Sielemann et al., 2020; Vanrolleghem et al., 2025). Effective analysis of wastewater data is hampered by delays in retrieving and compiling data, poor machine-readability, and limited interoperability between data pipelines and systems, data entry errors, etc. Inconsistent data formatting, for instance variation in dates and measurements, impedes machine readability and introduces inaccuracies, as demonstrated in a set of wastewater samples shared in the National Center for Biotechnology Information (NCBI)'s Sequence Read Archive where surveillance targets were misrecorded as "SARS-CoV-3," "SARS-CoV-4," and so on. While WWGS techniques vary widely by pathogen,key methodological factors influencing pathogen detection (e.g., extraction protocol; size of the contributing population) are often poorly documented, inconsistently shared, or stored in inaccessible locations. Constantly changing laboratory and bioinformatics methods – characteristic of a field still under development – can also lead to inconsistencies within data held by even a single institution or jurisdiction. These issues are compounded when data is shared with external stakeholders or with public databases.

Existing wastewater data harmonization resources cater to specific stakeholders, but fail to address the broader needs of the international WWGS community. The NCBI BioSample package for SARS-CoV-2 wastewater surveillance (version 1.0; https://www.ncbi.nlm.nih.gov/biosample/docs/packages/SARS-CoV-2.wwsurv.1.0/) was designed in collaboration with the CDC's National Wastewater Surveillance System (NWSS; https://www.cdc.gov/nwss/wastewater-surveillance.html), and was developed based on US data needs (Barrett et al., 2012). The European Nucleotide Archive (ENA) currently implements the Genomic Standards Consortium's Minimum Information (MIxS) wastewater/sludge checklist (https://www.ebi.ac.uk/ena/browser/view/ERC000023) as well as the ENA sewage checklist (https://www.ebi.ac.uk/ena/browser/view/ERC000036) - both are narrowly focused and do not address broader wastewater systems and materials that are often sampled for public health surveillance (https://genomicsstandardsconsortium.github.. The Public Health Environmental Surveillance Open Data Model (PHES-ODM; commonly used in Canada and Europe) was originally developed to support PCR-based surveillance and chemical/physical environmental attributes monitoring (Therrien et al., 2024). Additionally, organizations often create their own data structures and data dictionaries for internal management of wastewater contextual data, resulting in limited interoperability across datasets and systems. Applying international data standards and ontologies enhances accessibility, interoperability and reusability of contextual data (Griffiths et al., 2017; Lambert et al., 2017; Pettengill et al., 2021). Standards can also guide the selection of key public health data elements in the context of different data privacy and ethical concerns.

Recognizing these challenges in WWGS and the advantages of standards for genomic surveillance of pathogens, the Public Health Alliance for Genomic Epidemiology (PHA4GE) has identified the need for a fit-for-purpose, open-source WWGS contextual data standard, that structures information consistently and facilitates data sharing with trusted partners and public repositories. PHA4GE is an international coalition that promotes global consensus data standards, documents best practices, and advocates for improvements to make public health bioinformatics more accessible and reproducible. In collaboration with the global WWGS community, PHA4GE has developed a wastewater genomic surveillance contextual data specification package containing a data standard, tooling for easy implementation, exchange formats to enable interchange with public repositories, and supporting reference materials (reference guide, curation protocol, worked examples, new term request system). Designed for broad applicability, the specification supports diverse pathogen targets, wastewater systems and sequencing strategies across both low- and high-resource settings. To ensure broad applicability, the specification package has been piloted in 15 laboratories, databases, and organizations across five continents, and feedback was integrated into the PHA4GE WWGS package and used to strategize future development and tooling.

The specification package can be used by labs i) to structure and future-proof their own data in accordance with their data management policies, and ii) as a common language to facilitate data sharing and communication with partners and networks that may use different standards. Going forward, PHA4GE is also committed to maintaining its resources as WWGS data needs evolve.

Theory and Design Principles

Best Practices for Standards Development

PHA4GE has adopted a set of principles – originally developed by the Centre for Infectious Disease Genomics and One Health (CIDGOH) – that define an ideal international contextual data standard should look like and should achieve (**Box 1**). Ideally, such a standard would support data capture, analysis, sharing and reuse for a wide variety of use cases prioritized by both highly-resourced and lower-resourced laboratories. The standard should include known limitations of datasets that might impact their interpretation, and its development should be based on technical best practices and involve significant community cooperation and communication. Technical best practices for standards development emphasize interoperability, broad utility, extensibility, adaptability, human- and machine-readability, FAIR principles (making data Findable, Accessible, Interoperable, Reusable), and reuse of community semantic resources such as ontologies. Specific practices and their benefits are highlighted in **Table 1**.

Modular, Interoperable Framework based on ISO Standards

The International Organization for Standardization (ISO) develops standards when there is an identified market need, and operates by consensus across its 167 national member countries. PHA4GE contextual data standards for public health pathogen genomic surveillance are structured using a modular, interoperable, customizable framework, adapted by CIDGOH from ISO 23418. Within the framework, standardized fields and terms are grouped into thematic modules (e.g. Database identifiers, Sample collection and processing, Environmental conditions and measurements, etc). This framework helps to create interoperability as new specifications can be created by mixing-and-matching modules – as well as enriching/depleting fields and terms within modules – according to the needs of projects and surveillance programs. The framework has been used previously to structure standardized terminology in the PHA4GE SARS-COV-2 contextual data specification (Griffiths et al., 2022), and is currently being used to structure CIDGOH's data standards for One Health Antimicrobial Resistance (AMR) surveillance (Griffiths et al., 2024) and Mpox (https://github.com/cidgoh/MPox_Contextual_Data_Specification). An example highlighting the framework's modularity across two of these standards is shown in **Figure 1**.

Implementing International Semantics

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)) (The OBI Consortium et al., 2007). IRIs link project- and organization-specific preferred labels to information such as definitions, definition sources, alternative labels used by other communities and organizations (synonyms), logical relationships linking the term to other kinds of information (axioms), and hierarchical relationships enabling different types of classifications and groupings. With an emphasis on common meaning, ontologies incorporate synonyms and database mappings, contributing to interoperability as there is rarely a "one-size fits all" terminology or nomenclature system. 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, such as those employed in different types of artificial intelligence and machine learning applications. PHA4GE data standards implement standardized fields and terms sourced from the Open Biological and Biomedical Ontology (OBO) Foundry library of interoperable ontologies (Jackson et al., 2021; The OBI Consortium et al., 2007).

Communities of practice like the OBO Foundry articulate and implement best principles and practices to enable reuse of terminology across domains and sectors. The OBO Foundry recommends a common overarching organizational structure called the Basic Formal Ontology (top-level ontology) in which all "things" that exist are essentially divided into material entities, processes, and qualities (characteristics). 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 (Gennari et al., 2003; Jackson et al., 2019; Ong et al., 2017; Whetzel et al., 2011).

Standards Development Process

PHA4GE implements CIDGOH's ten step process to develop contextual data standards (Griffiths et al., 2025, 2022). This structured, community-driven approach is anchored in: stakeholder engagement and co-creation, the development of tools and interchange formats to put standards into practice within public health workflows, as well as multiple rounds of testing and feedback. The specification is iteratively tested and reviewed by stakeholders, and examples of real-world implementations are documented. This process encourages community participation through early and consistent engagement. Standards are made publicly available on GitHub, and are updated as data needs evolve over time.

Creating the PHA4GE Wastewater Contextual Data Specification

Community Consultation, Needs Assessment, and Scoping

WWGS programs in highly-resourced countries often operate differently than those in lower-resourced countries or communities, and have different data needs. For example, WWGS programs in highly-resourced settings commonly sample from municipal treatment plants, whereas lower-resourced settings more often sample from alternative systems (e.g. latrines, lagoons, or septic systems more common in rural African or Indigenous-Canadian communities) that require different methodologies (e.g. cold chain maintenance for sample transport from distant regions to central laboratories). The lack of fit-for-purpose data structures for lower-resourced WWGS strategies places additional barriers to effective data collection, harmonization, and integration; when available data structures are difficult to implement in lower-resourced labs, global buy-in and uptake decrease and data silos proliferate.

To ensure that the PHA4GE wastewater contextual data specification was fit for a variety of public health purposes and to ensure broad utility, consultations were conducted during the early development phase of the package with stakeholders representing various interests in WWGS. Consultations were conducted with more than 60 practitioners involved in global wastewater-related public health activities, including: wastewater sanitation, wastewater surveillance of pathogens and resistance, policy-making, disease modeling, and standards development. Groups consulted include: US Centers for Disease Control and Prevention, American Public Health Laboratories, UK Health Security Agency, Public Health Agency of Canada, Africa PGI, World Health Organization, and International Nucleotide Sequence

Database Collaboration (INSDC). Various data dictionaries (e.g. PHES-ODM), public repository requirements (e.g. GISAID, INSDC BioSample packages), as well as a wide range of literature, were also reviewed to understand industry and public health norms and practices (Hendriksen et al., 2019; National Academies of Sciences, Engineering, and Medicine, 2023; Timme et al., 2024; Velkushanova et al., 2019).

Results from consultations and resource reviews revealed a wide range of analytical approaches within the global WWGS community. Surveillance programs use different sequencing assay types including amplicon sequencing, culture and whole genome sequencing (WGS), and metagenomics approaches. A wide variety of pathogens are targeted: polio, SARS-CoV-2, *Salmonella* and other foodborne pathogens, cholera, tuberculosis, antimicrobial resistance determinants, vaccine resistant variants, etc. WWGS programs have been leveraged, in tandem with clinical surveillance, to determine prevalence and emergence of strains/clades/lineages, assess risk, design interventions, or investigate outbreaks. City- or region-wide WWGS programs help establish baseline norms and track pathogen trajectories, whereas other programs survey individual communities or settings (e.g. programs that survey travellers, hospitals or care facilities, remote or indigenous communities, etc.). WWGS data may also need to be integrated with additional water-based genomics data, such as analyses of water impacted by agriculture or other land use.

These consultations also revealed a wide range of data needs. For example, WWS programs sample from a variety of environmental materials (e.g. surface water vs. sewage by-products such as sludge or effluent), sites (e.g. toilets, sewers, ponds), communities (e.g. municipalities, residents of correctional facilities, travellers on airplanes, etc), and systems (e.g. combined vs. sanitary sewer systems). Sample collection approaches vary, as specimens may be: subsampled or recombined; processed prior to sequencing (e.g. culture, enrichment, filtration, etc.); and collected continuously or longitudinally (repeat sampling over time). As such, there is a need to record sampling events, subsamples, and replicates. WWGS can be affected by environmental conditions (e.g. temperature, weather) and by physico-chemical properties (e.g. turbidity, concentration of nitrogen) that impact specific pathogens. Some labs may also need to tag known quality issues for given datasets, which are produced and shared in order to optimize or validate wet lab protocols or bioinformatics software. Upon implementing different laboratory and bioinformatics methods to produce sequence data, some laboratories share processed reads while others share consensus sequences or assemblies.

Based on the data from these consultations, the specification was scoped to address a variety of use cases, targets, sampling strategies, sample types, environmental conditions, detection and sequencing technologies, and bioinformatic approaches. The specification captures contextual data necessary to interpret sequencing data during analyses. While some priority information, such as taxonomic identification, can be recorded, detailed analytical outputs (e.g. SARS-CoV-2 lineage abundances from a Freyja analysis) can be linked to contextual data records using the filename fields. The specification is intended for labs to standardize and future-proof their own contextual data, to streamline sharing of contextual data within trusted networks, and to facilitate sharing of data with public repositories.

Specification Content and Structure

The wastewater specification contains 252 fields divided across 11 modules which include: Database identifiers, Sample collection and processing, Strain and isolate information, Environmental conditions and measurements, Sequence information, Bioinformatics and QC metrics, Taxonomic identification information, AMR detection information, Lineage/clade information, Pathogen diagnostic information, and Contributor acknowledgement. The fields found in each module are illustrated in **Figure 2**. The specification also includes 745 standardized terms distributed across 61 picklists of values. A set of standardized null values sourced from the INSDC (i.e. Missing, Not Applicable, Not Provided, Not Collected, Restricted Access) were also included in every picklist. New ontology terms were developed for any vocabulary that did not already exist in OBO Foundry ontologies using methodology previously described (Griffiths et al., 2024). Standardized terms for vocabulary were sourced from 11 OBO Foundry ontologies which can be found in **Table 2**.

The specification was broad to cover a wide array of use cases; however, adoption and compliance with standards tend to increase when a small, common set of fields are consistently prescribed across data generation projects. As such, a minimal set of contextual data ("required fields") was established in order to decrease the burden of data entry. The set of 16 required fields was selected based on common priorities across different genomic surveillance initiatives. Fields not part of the minimal required set were labeled as either "recommended" or "optional." Required and recommended fields are listed in **Table 3**.

To improve machine-readability, the WWGS specification separates proper nouns from their modifiers and any processes in which they are involved. For example, "environmental material" includes substances such as "Surface water [ENVO:00002042]" and "Sludge [ENVO:00002044]", while "environmental material properties" includes descriptors such as "Fluid (stagnant) [GENEPIO:0101004]" and "Treated [GENEPIO:0101008]". In many fields, multiple descriptors can be included (known as "multi-tagging"). Furthermore, most metrics and measurements are represented by three related fields to separately store the value, unit, and method (ex: "turbidity measurement value", "turbidity measurement unit", and "turbidity measurement method"). Listing the value and units separately increases computability for automated analyses; including a measurement method is relevant for reproducibility as well as interpretation of metrics. Similarly, while many targets in wastewater are identified using PCR, other approaches exist, thus necessitating flexibility in reporting of diagnostic measurements and methods. As such, diagnostic methods and targets are recorded separately, and results may be captured qualitatively (i.e., presence vs. absence) or quantitatively (e.g. "diagnostic measurement value", "diagnostic measurement unit", and "diagnostic measurement method"). All dates in the specification follow a prescribed format according to ISO 8601 (YYYY-MM-DD).

Implementation of the Specification

Implementation and Tooling within the DataHarmonizer

To operationalize the specification within public health workflows, it was implemented as use case-specific templates within a data curation tool called the DataHarmonizer. The DataHarmonizer is a web browser-based spreadsheet application used to curate, validate and transform contextual data into a submission-ready format for different data repositories. This environment is locally installed to increase data security, and provides data validation to confirm proper formatting. The DataHarmonizer organizes schemas using LinkML, a data modelling language for data dictionaries that can represent specifications in JSON, YAML, csv and other formats (Gill et al., 2023). More information can be found at https://github.com/cidgoh/DataHarmonizer/wiki/DataHarmonizer-Templates.

Three DataHarmonizer wastewater templates were developed to support the most common use cases identified through the needs assessment exercise: SARS-CoV-2 surveillance, AMR detection and surveillance, and pathogen-agnostic surveillance (i.e. surveillance of any pathogen). All three templates share a common set of modules including: Database identifiers, Sample collection and processing, Environmental conditions and measurements, Sequence information, Bioinformatics and QC metrics, Pathogen diagnostic testing, and Contributor acknowledgement (**Figure 3, top**). The templates also contain additional modules: Lineage/clade information in the SARS-CoV-2 template; Taxonomic identification information in the AMR and pathogen-agnostic templates; Strain and isolation information in the pathogen-agnostic template; and AMR detection information in the AMR template (**Figure 3, bottom**). The collection of fields and terms within common modules vary slightly across the templates depending on the relevance of the information to each use case (e.g., the SARS-CoV-2 template does not contain the field “genomic target enrichment method” as implementation of amplicon sequencing based strategies is assumed). Worked examples are provided later in the section on “Harmonization Challenges and Solutions” to illustrate how these templates can be used in real-world scenarios. These three wastewater templates can be found online: https://github.com/cidgoh/pathogen-genomics-package/releases. A colour-coding scheme was introduced to demarcate the core set of “required” fields (yellow) vs. recommended (purple) or optional (white) fields. An example of curated data in the DataHarmonizer Pathogen Agnostic template is presented in **Figure 4**.

Mapping, Interchange Formats, and Interoperability

In addition to the PHA4GE wastewater specification, other data standards also support the WWGS community; allowing for conversion between these standards enables interoperability between datasets and systems that implement them. PHA4GE has worked with different community partners to map its data elements to those of different resources and repository requirements, including: NCBI’s SARS-CoV-2 Wastewater Surveillance BioSample package, also used by the DNA Database of Japan (DDBJ); NCBI’s SRA submission form; GSC’s MIxS-based Wastewater/Sludge; ENA’s Sewage BioSample packages; the PHES-ODM;

and GISAID Wastewater submission requirements (**Table 4**). Separate mapping files are provided to demonstrate the overlap and differences between these specifications (**Table 4**). These mappings were used to develop data interchange formats in the DataHarmonizer, so that users can enter data in the PHA4GE format and export data in NCBI BioSample, NCBI SRA, PHES-ODM, and GISAID formats. NCBI exports from the DataHarmonizer are compliant with the INSDC Pathogen Data Object Model: attributes pertaining to a sample are included as user-defined contextual data in BioSample submissions, whereas attributes pertaining to sequence data are included in SRA metadata (Timme et al., 2023). Biosample and SRA adaptations specific for wastewater and environmental surveillance have also been modelled by the FDA's GenomeTrakr network (Timme et al., 2024).

The PHA4GE Wastewater Contextual Data Specification Package is intended to integrate into the international wastewater surveillance ecosystem in a number of ways: i) labs without data standards can adopt the PHA4GE standard; ii) labs implementing other existing standards can use the PHA4GE specification as a common language to facilitate data sharing and communication with other labs, and iii) labs can communicate their needs to PHA4GE in order to contribute to the maintenance and ongoing development of the specification.

Supporting Materials

To better facilitate proper usage of the specification and DataHarmonizer templates, as well as to improve community data curation practices, a variety of supporting materials were developed and included in the package. These supporting materials include: i) DataHarmonizer download and operating instructions, ii) a curation standard operating procedure (SOP) providing guidance on applying the standard to different data types and scenarios as well as ethical, privacy and practical considerations, and iii) Field and Term reference guides providing definitions, guidance for data entry, and examples of use. The curation SOP and the reference guides are periodically updated and aligned with new releases of the DataHarmonizer templates. A summary of resources included in the Wastewater Contextual Data Specification Package is provided in **Table 4**.

Versioning and Availability

The wastewater specification is free, publicly available, and version-controlled on GitHub. All edits and updates are tracked in release notes. (https://github.com/pha4ge/Wastewater_Contextual_Data_Specification/releases). Versioning is done in the format of x.y.z, where x denotes field level changes, y refers to term value / ID level changes, and z includes changes to definitions, guidance, examples, and formatting.

Community discussions contributing to updates are also tracked on the repository GitHub issue tracker. To track data needs over time and enable community discussions about ontology terms, a New Term Request (NTR) System has been provided via GitHub Issues to enable term requests. 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.

Piloting the PHA4GE Wastewater Specification in Real-World Settings

Early user feedback was provided by a number of testing laboratories

Stakeholder engagement and empirical testing were paramount to ensuring the utility of the PHA4GE wastewater specification across diverse settings. To test the specification, pilot projects were launched in labs, databases, and organizations spanning 5 continents and 15 countries (**Figure 5**). To support the involvement of laboratories in low-resourced settings, ten subgrant awards of $5,000 USD were distributed to laboratories by PHA4GE (funding provided by the Gates Foundation). Prior to release of the community-wide call for subgrant applications, an ethics review was performed by the PHA4GE Ethics & Data Sharing team. Applicants who did not receive a subgrant were offered the opportunity to participate in testing in exchange for authorship (funded participants also received authorship). The testing exercise consisted of orientation and training sessions, followed by a four-week testing period during which subgrant awardees were tasked with curating their own data using the wastewater specification package (i.e., DataHarmonizer templates and supporting materials). No contextual data sharing was required of the testing labs - with PHA4GE or otherwise. Testing labs evaluated the specification's performance and coverage of data needs, and reported any gaps and challenges. Participants could troubleshoot issues and ask the PHA4GE development team questions during regular "Office Hours" throughout the testing period.

Funded applications represented a diversity of projects, ensuring that the standard was rigorously tested across a wide range of scenarios. Testing labs' wastewater projects encompassed a variety of sample types (e.g., activated sludge, domestic water, surface water, water delivered via piped systems vs ground transport vehicles, municipal and industrial waste, livestock effluents and agricultural runoff), sampling sites (e.g., wastewater treatment plants; non-aerated lagoons; rivers; industry, residential, correctional, educational and healthcare facilities; sewered vs non-sewered systems; animal slaughter sites; open drains; manholes; drainage channels in slum settlements and by roadsides; municipal vs airport waste systems; septic tanks; pilot-scale reactors), collection methods (e.g. composite and grab sampling, longitudinal sampling, swabbing), and pathogens and phenomena under investigation (SARS-CoV-2, enterics, AMR determinants, emerging pathogens, microbial diversity).

User feedback indicated that the worked examples were the most powerful tool for understanding how the specification could and should be used. As data needs varied extensively across the projects, participants strongly recommended updating the DataHarmonizer to enable users to create and customize their own templates based on the specification. The ability to adapt global data standards to local needs would better enable uptake and interoperability while supporting data sovereignty and local public health priorities.

Participants suggested a number of edits for content enhancement, including: expanding the pathogen diagnostic testing fields to capture multiple PCR targets, adding additional collection methods and devices, and minor revisions to the reference guide and SOP to simplify and clarify. Participants also requested inclusion of picklists of bioinformatics tools to support reproducibility. As fulfilling this request is challenging owing to the expansive range of evolving tools, PHA4GE is currently considering how best to address this need. Examples of user feedback are presented as two case studies that were developed in collaboration with participating labs in Argentina and Ghana. The case studies, specifying wastewater surveillance targets, sampling strategies, challenges, implementation of the specification, and recommendations, are presented in **Box 2**.

Collaboration with public databases to integrate the specification

Public databases enabling sharing of pathogen genomic data have become essential tools for efficient public health decisions. The INSDC is a collaborative network of organizations that manage nucleotide sequence data repositories – including the National Library of Medicine (NLM), NCBI in the United States that hosts SRA and GenBank, the European Molecular Biology Laboratory (EMBL), European Bioinformatics Institute (EBI) that hosts the ENA in the United Kingdom, and the The Research Organization of Information and Systems (ROIS), National Institute of Genetics (NIG) in Japan that hosts the DNA Data Bank of Japan (DDBJ) – that has provided public sharing of sequence data for over 40 years. The INSDC hosts over 300K wastewater sequence records. To support continued sharing of wastewater genomic data and contextual data, PHA4GE is collaborating with the INSDC to integrate parts of this specification into their systems. Specifically, ENA is integrating a subset of PHA4GE fields into their new WWGS package, and NCBI is developing an entirely new WWGS package based on the PHA4GE specification. PHA4GE has also engineered exchange formats into the DataHarmonizer to export data in existing NCBI and ENA formats to facilitate data submissions.

In addition to the INSDC, PHA4GE is also working with Canada's VirusSeq Data Portal to enable open sharing of wastewater genomics data. This portal currently contains over 600K Canadian SARS-CoV-2 genomes, with corresponding contextual data structured using the CanCOGeN SARS-CoV-2 specification (Gill et al., 2024, 2023), the progenitor of the PHA4GE SARS-CoV-2 contextual data standard (Griffiths et al., 2022). Currently, the VirusSeq Data Portal team are expanding the platform's capacity to store and analyze wastewater genomics data, and are working to implement the PHA4GE Wastewater specification as part of its database schema (Gill et al., 2024).

Worked examples demonstrate solutions to harmonization challenges

To illustrate how harmonization challenges are addressed by the specification, a series of use cases and simulated data based on real-world experience were developed, and are presented below. The PHA4GE Wastewater Contextual Data Specification was designed to address key challenges in harmonizing data across diverse wastewater genomics surveillance efforts, such as: establishing chains of custody; properly attributing contributions of different partners to the generation of sequence data; identifier tracking across platforms and databases; structuring complex sample descriptions; capturing nuances of experimental design; and documenting sequencing and bioinformatics methodologies.

The worked examples highlight how different fields and templates can be used to structure information in common public health scenarios such as minimal data collection, identifier and attribution tracking, sequencing of cultured organisms from a wastewater sample, and identifying pathogens via tiling amplicon sequencing approaches. Worked examples include descriptions of scenarios and accompanying fields and terms, colour-coded to highlight required and recommended data elements. DataHarmonizer files are also provided to illustrate how the data should best be entered in different templates (see resource list in **Table 4**). Note: Use of the specification does NOT require public sharing of data, although PHA4GE encourages sharing according to organizational policies. Structured data can be stored for local use and reuse.

Scenario 1: Structuring complex sample descriptions

A linelist containing free text descriptions of wastewater samples was harmonized using key sample collection fields and terms from the specification. Standardized descriptors for the samples are provided in **Table 5**.

Scenario 2: Partial record from a sample collector

This record highlights the minimal "required" contextual data in the template (15 fields), and highlights what to do when not all the information can be supplied.

A technician collected a wastewater sample (sample ID CVW-6758900) at a municipal wastewater treatment plant in the UK on February 28, 2023, as part of the UKHSA's Environmental Wastewater Monitoring programme. The targets of interest were environmental pollutants, however the sample was later used for pathogen surveillance. The sample is awaiting sequencing. The technician provides a partial contextual data record describing sample collection to the sequencing lab, who will provide further details at a later date. The record submitted to the sequencing lab is provided below.

Template: Pathogen-Agnostic
specimen collector sample ID: CVW-6758900
sample collected by: UK Health Security Agency
geo_loc_name (country): United Kingdom [GAZ:00002637]
geo_loc_name (state/province/territory): Not Provided [GENEPIO:0001668]
organism: Not Provided [GENEPIO:0001668]
purpose of sampling: Wastewater chemical surveillance [GENEPIO:0100870]
sample collection date: 2023-02-28
environmental site: Wastewater treatment plant [ENVO:00002272]
environmental material: Wastewater [ENVO:00002001]
isolate ID: Not Provided [GENEPIO:0001668]

purpose of sequencing: Not Provided [GENEPIO:0001668]
sequenced by: Not Provided [GENEPIO:0001668]
sequenced by contact name: Not Provided [GENEPIO:0001668]
sequenced by contact email: Not Provided [GENEPIO:0001668]
sequencing instrument: Not Provided [GENEPIO:0001668]
read mapping software name: Not Provided [GENEPIO:0001668]
read mapping software version: Not Provided [GENEPIO:0001668]

Scenario 3: Sequencing cultured organisms isolated from wastewater

This record highlights identifier and attribution tracking, sampling strategy capture, structuring of sample metadata, high-level quality control annotations, as well as microbiological, assembly and sequencing methods tagging.

A wastewater sample (WD-1234-i9v) was collected as part of a research project examining the utility of wastewater surveillance for identifying and monitoring cholera in the community. The sample was collected on January 3, 2022, from a school latrine (sampling site ID AAA123). Cholera isolates were cultured from the sample (microbiological method as per Mtonga et al, 2018; doi: 10.1371/journal.pntd.0007642), and one of the isolates was sequenced (VC-1234). The sample was sequenced by the Zambia Ministry of Health (project lead: Cheelo Mtonga; mtongam@moh.zambia.org). The library was prepared using an Nextera XT DNA Library Preparation Kit and sequenced using an Illumina MiSeq. Raw reads were quality filtered and primer sequences trimmed using Trimmomatic. Paired-end reads were assembled *de novo* to construct a draft genome using the SPADES v.3.11.1 software. The quality of de novo assemblies was assessed using Quast (v.4.5) and the data passed the research project's quality control processes. The metadata, sequence reads and assembly were then uploaded to NCBI (BioProject PRJNA608678). The associated contextual data record is provided below.

Template: Pathogen-Agnostic
specimen collector sample ID: WD-1234-i9v
BioProject accession: PRJNA608678
sampling site ID: AAA123
sample collected by: Zambia Ministry of Health
geo_loc_name (country): Zambia [GAZ:00001107]
geo_loc_name (state/province/territory): Not Provided [GENEPIO:0001668]
organism: Vibrio cholerae [NCBITaxon:666]
purpose of sampling: Research [GENEPIO:0100003]
sample collection date: 2022-01-03
environmental site: School [ENVO:03501130]
environmental material: Wastewater [ENVO:00002001]
wastewater system type: Latrine [ENVO:01000519]
isolate ID: VC-1234
microbiological method: doi: 10.1371/journal.pntd.0007642
purpose of sequencing: Research [GENEPIO:0100003]

sequencing assay type: Whole genome sequencing assay [OBI:0002117]
library preparation kit: Nextera XT DNA Library Preparation Kit
sequenced by: Zambia Ministry of Health
sequenced by contact name: Cheelo Mtonga
sequenced by contact email: mtongam@moh.zambia.org
sequencing instrument: Illumina MiSeq [OBI:0002003]
raw sequence data processing method: Trimmomatic
sequence assembly software name: SPAdes
sequence assembly software version: 3.11.1
quality control method name: Quast
quality control method version: 4.5
quality control determination: Sequence passed quality control [GENEPIO:0100563]
sequence submitted by: Cheelo Mtonga
sequence submitter contact email: mtongam@moh.zambia.org

Scenario 4: Identifying single pathogens using amplicon approaches

This record highlights the use of experimental replicates, amplicon scheme, dehosting software, and assembly annotations.

The Global Viral Disease Eradication Initiative uses wastewater samples from different parts of the world to identify and characterize pathogens causing priority diseases such as measles, polio and hepatitis. A wastewater sample (CAM-LAG-00123) was collected from a wastewater treatment lagoon in Cambodia on February 15, 2023, as part of a baseline surveillance assessment. The sample was subdivided and the different subsamples were used as technical replicates during method development and optimization. A sequencing library was prepared for one of the replicates (subsample CAM-LAG-00123-3b) using a proprietary enrichment kit (GVDE enrichment kit) and a primer amplicon panel specific for 20 different diseases, including Poliovirus types 1, 2 and 3 (GVDE Viral Surveillance Panel). The library was sequenced using an Illumina NovaSeq 6000 (sequencing lab contact: Dr. Maya Lee; mlee@gvde.org). The reads were filtered, trimmed, dehosted, and a Poliovirus 1 genome was de novo assembled using a suite of tools available in the RAMPART (v1.2.0) platform. The associated contextual data record is provided below.

Template: Pathogen-Agnostic
specimen collector sample ID: CAM-LAG-00123
specimen collector subsample ID: CAM-LAG-00123-3b
sample collected by: The Global Viral Disease Eradication Initiative
geo_loc_name (country): Cambodia [GAZ:00006888]
geo_loc_name (state/province/territory): Not Provided [GENEPIO:0001668]
organism: Poliovirus 1 [NCBITaxon:12080]
purpose of sampling: Wastewater pathogen surveillance [GENEPIO:0100872]
scale of sampling: Community-level surveillance [GENEPIO:0100874]
sample collection date: 2023-02-15

environmental site: Waste stabilization pond (lagoon) [ENVO:03600076]
environmental material: Wastewater [ENVO:00002001]
wastewater system type: Wastewater stabilization pond (lagoon) [ENVO:03600076]
experimental specimen role type: Technical replicate [EFO:0002090]
isolate ID: Not Provided [GENEPIO:0001668]
purpose of sequencing: Baseline surveillance (random sampling) [GENEPIO:0100005]
sequencing assay type: Amplicon sequencing assay [OBI:0002767]
library preparation kit: GVDE enrichment kit
amplicon pcr primer scheme: GVDE Viral Surveillance Panel
sequenced by: The Global Viral Disease Eradication Initiative
sequenced by contact name: Maya Lee
sequenced by contact email: mlee@gvde.org
sequencing instrument: Illumina NovaSeq 6000 [GENEPIO:0100123]
raw sequence data processing method: RAMPART 1.2.0
dehosting method: RAMPART 1.2.0
sequence assembly software name: RAMPART
sequence assembly software version: 1.2.0

Scenario 5: Identifying single pathogens using metagenomic approaches (alternative to Scenario 4)

This record highlights alternative methods to identifying pathogens in complex samples by capturing “organism” information from taxonomic analysis.

The Global Viral Disease Eradication Initiative uses wastewater samples collected in different parts of the world to identify and characterize pathogens causing priority diseases such as measles, polio and hepatitis. A wastewater sample (CAM-LAG-00123) was collected from a wastewater treatment lagoon in Cambodia on February 15 2023 as part of a baseline surveillance assessment. The sample was subdivided and the different subsamples were used as technical replicates during method development and optimization. Nucleic acids were extracted from one of the replicates (subsample CAM-LAG-00123-4a) using a MagMAX Wastewater Ultra Nucleic Acid Isolation Kit, with Virus Enrichment. The library was prepared using an Nextera XT DNA Library Preparation Kit. The library was sequenced using an Illumina NovaSeq 6000 (sequencing lab contact: Dr. Maya Lee; mlee@gvde.org) and the reads were filtered, trimmed, dehosted using a suite of tools available in the RAMPART (v1.2.0) platform. Total filtered and dehosted reads were mapped to a custom reference taxonomic database (GVDEdb 3.4.5) and Poliovirus 1 was identified with 85x coverage across 90% of the Poliovirus 1 reference genome. The associated contextual data record is provided below.

Template: Pathogen-Agnostic
specimen collector sample ID: CAM-LAG-00123
specimen collector subsample ID: CAM-LAG-00123-4a
sample collected by: The Global Viral Disease Eradication Initiative
geo_loc_name (country): Cambodia [GAZ:00006888]

geo_loc_name (state/province/territory): Not Provided [GENEPIO:0001668]
organism: Poliovirus 1 [NCBITaxon:12080]
purpose of sampling: Wastewater pathogen surveillance [GENEPIO:0100872]
scale of sampling: Community-level surveillance [GENEPIO:0100874]
sample collection date: 2023-02-15
environmental site: Waste stabilization pond (lagoon) [ENVO:03600076]
environmental material: Wastewater [ENVO:00002001]
wastewater system type: Wastewater stabilization pond [ENVO:03600076]
experimental specimen role type: Technical replicate [EFO:0002090]
nucleic acid extraction kit: MagMAX Wastewater Ultra Nucleic Acid Isolation Kit, with Virus Enrichment
purpose of sequencing: Baseline surveillance (random sampling) [GENEPIO:0100005]
sequencing assay type: Whole virome sequencing assay [OBI:0002768]
library preparation kit: Nextera XT DNA Library Preparation Kit
sequenced by: The Global Viral Disease Eradication Initiative
sequenced by contact name: Maya Lee
sequenced by contact email: mlee@gvde.org
sequencing instrument: Illumina NovaSeq 6000 [GENEPIO:0100123]
raw sequence data processing method: RAMPART 1.2.0
dehosting 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

Scenario 6: Characterizing AMR in wastewater samples

This record highlights alternative methods to identifying pathogens and AMR determinants in complex samples by capturing information for contig assembly, as well as taxonomic analysis, and AMR detection software and reference databases information.

A global sewage project collects samples from a wide variety of locations around the world in order to establish baselines for antimicrobial resistance (i.e. prevalence, distribution, types of resistance). The national public health laboratory of the UAE collected a wastewater sample from an airport sewer system in Dubai on June 6, 2023 (sample UAErt3-478-0091). The sample was sequenced using an Oxford Nanopore GridION instrument, and the reads are quality filtered and trimmed using BBduk2 v1.5. To assign resistance genes to pathogen species, contigs were assembled using metaSPAdes (v3.13.0) and then characterized via mapping to the CARD resistance database (v3.2.9) using the Resistance Gene Identifier software (v6.0.3). A CARD report summarizing results is stored in the system for longitudinal comparisons (filename: 20230606_XYB-123_results.txt). The associated contextual data record is provided below.

Template: AMR

specimen collector sample ID: UAErt3-478-0091
sample collected by: The National Public Health Laboratory of the United Arab Emirates
geo_loc_name (country): United Arab Emirates [GAZ:00005282]
geo_loc_name (state/province/territory): Emirate of Dubai
purpose of sampling: Wastewater pathogen surveillance [GENEPIO:0100872]
scale of sampling: Institution-level surveillance [GENEPIO:0100875]
sample collection date: 2023-06-06
environmental site: Airport [ENVO:03501122]
environmental material: Wastewater [ENVO:00002001]
purpose of sequencing: Baseline surveillance (random sampling) [GENEPIO:0100005]
sequencing assay type: Whole metagenome sequencing assay [OBI:0002623]
sequenced by: The National Public Health Laboratory of the United Arab Emirates
sequenced by contact name: Not Provided [GENEPIO:0001668]
sequenced by contact email: mlee@gvde.org
sequencing instrument: Illumina NovaSeq 6000 [GENEPIO:0100123]
number of total reads: 21300465
minimum post-trimming read length: 150
number of contigs: 1500201
raw sequence data processing method: RAMPART 1.2.0
dehosting method: RAMPART 1.2.0
sequence assembly software name: Bowtie2
sequence assembly software version: 2.5.3
AMR analysis software name: Resistance Gene Identifier
AMR analysis software version: 6.0.3
AMR reference database name: Comprehensive Antibiotic Resistance Database (CARD)
AMR reference database version: 3.2.9
AMR analysis report filename: 20230606_XYB-123_results.txt

Scenario 7: SARS-CoV-2 surveillance (rich contextual data)

This record highlights how rich contextual data can be captured using the specification, including: catchment details such as geographical coordinates and population ranges, activity upstream of sampling that may affect results, how to record longitudinal sampling events, capture of environmental conditions and measurements, associated laboratory testing results (Ct values), and lineage designations.

Untreated, fast moving, wastewater is continuously collected in a municipal sewer system starting on Nov 1, 2023, for 72hrs. The sewer system, which collects rainwater as well as household and institutional waste, is part of a routine surveillance program for tracking community-level SARS-CoV-2 variants (sewer site ID WWSC2-ABC-b) in order to establish baseline norms. The sewer is located near a hospital and the hospital's effluent is piped into the sewer system. Five Moore swabs from the site of collection are pooled (sample ID BW-WW-12345). It rained the day before sample collection (5cm of rain). The wastewater catchment area serves approx 800,000 people in a suburban area (Mississauga, Ontario, Canada). The ambient air temperature at the time of collection was 15 degrees Celsius. The water was 8 degrees Celsius at the time of collection, and 3 degrees Celsius when it was received by the sequencing lab. The instantaneous flow rate is 3 cubic meter per second (m^3/s), with 8% total suspended solids. The sample was collected by the Region of Peel regional authority, and sequenced by the Public Health Ontario provincial health laboratory (contact: Johnny Bloggs; jbloggs@provlab.ca). A watershed shapefile delineating the geographical coordinates covered by the sewer system is available. The presence of SARS-CoV-2 was first detected using qPCR (N1 gene, Ct value of 22). The amplicon-based sample was sequenced on an Illumina MiSeq on Jan 18, 2024, using the ARTIC V5 400bp primer scheme (artic-v5.3.2_400), and consensus sequences were generated using ViralRecon software v1.23 and lineage assignments were performed using pUShER (v1.2.6). The rich contextual data record for the sequence is provided below. This record is for the public health laboratory's use only, and many details were removed when sharing data according to organization-specific data sharing policies. The associated contextual data record is provided below.

Template: SARS-CoV-2

specimen collector sample ID: BW-WW-12345
sampling site ID: WWSC2-ABC-b
sample collected by: Region of Peel Regional Authority
geo_loc_name (country): Canada [GAZ:00002560]
geo_loc_name (state/province/territory): Ontario [GAZ:00002563]
geo_loc_name (city): Mississauga
watershed shapefile availability: Watershed shapefile available
organism: Severe acute respiratory syndrome coronavirus 2 [NCBITaxon:2697049]
purpose of sampling: Wastewater pathogen surveillance [GENEPIO:0100872]
presamping activity: Healthcare activity [NCIT:C16205]
presampling activity details: hospital effluent piped into sewer system
sample collection date: 2023-11-01
sample collection end date: 2023-11-04
sample collection time duration value: 72
sample collection time duration unit: Hour [UO:0000032]
scale of sampling: Community-level surveillance [GENEPIO:0100874]
specimen processing: Pooling specimens [OBI:0600016]
specimen processing details: 5 Moore swabs pooled from same sewer
environmental site: Sewer [GENEPIO:0102064]
environmental material: Wastewater [ENVO:00002001]
environmental material properties: Untreated [GENEPIO:0101009]; Fluid (fast) [GENEPIO:0101006]
wastewater system type: Combined sewer system [ENVO:03501453]
collection device: Moore swab [GENEPIO:0100949]
collection method: Passive composite sampling [GENEPIO:0100955]
populated area type: Suburban [GSSO:011077]
water catchment area human population measurement value: 800 000

water catchment area human population bin: 100,000 - 1,000,000 people
presampling weather conditions: Rain [ENVO:01001564]
precipitation measurement value: 5
precipitation measurement unit: centimeter (cm) [UO:0000015]
ambient temperature measurement value: 15
ambient temperature measurement unit: degree Celsius (C) [UO:0000027]
instantaneous flow rate measurement value: 3
instantaneous flow rate measurement unit: cubic meter per second (m^3/s)
total suspended solids (TSS) measurement value: 8
total suspended solids (TSS) measurement unit: Percent (%) [UO:0000187]
sample temperature value (at collection): 8
sample temperature unit (at collection): degree Celsius (C) [UO:0000027]
sample temperature value (when received): 3
sample temperature unit (when received): degree Celsius (C) [UO:0000027]
purpose of sequencing: Baseline surveillance (random sampling) [GENEPIO:0100005]
sequencing assay type: Amplicon sequencing assay [OBI:0002767]
sequencing date: 2024-01-18
sequenced_by: Public Health Ontario
sequenced_by_contact_name: Johnny Bloggs
sequenced_by_contact_email: jbloggs@provlab.ca
sequencing instrument: Illumina MiSeq [OBI:0002003]
amplicon pcr primer scheme: artic-v5.3.2_400 [GENEPIO:0100856]
amplicon size: 400
consensus sequence software name: ViralRecon
consensus sequence software version: 1.23
diagnostic measurement method: Quantitative real time polymerase chain reaction (qPCR) [OBI:0000893]
gene name: N gene (orf9) [GENEPIO:0100153]
diagnostic target presence: Diagnostic target present [GENEPIO:0100987]
diagnostic measurement value: 22
diagnostic measurement unit: cycle threshold (Ct) [GENEPIO:0100657]
lineage/clade name: JN.1
lineage/clade analysis software name: pUShER
lineage/clade analysis software version: 1.2.6

Scenario 8: SARS-CoV-2 lineage determination using Freyja Analysis

This record highlights how lineage identifications using Freyja analyses can be captured using the specification, including: results details such as lineage/clade names (multiple lineages can be summarized and separated by a semicolon) and analysis filenames, as well as methods details such as lineage analysis software name and version number.

A wastewater sample (sample ID BW-WW-12345) was collected in a municipal sewer system on Nov 1, 2023, as part of a routine surveillance program for tracking community-level

SARS-CoV-2 variants (sewer site ID WWSC2-ABC-b) in order to establish baseline norms. The sample was collected by the Region of Peel regional authority, and sequenced by the Public Health Ontario provincial health laboratory (contact: Johnny Bloggs; jbloggs@provlab.ca). The presence of SARS-CoV-2 was first detected using qPCR (N1 gene, Ct value of 22). The amplicon-based sample was sequenced on an Illumina NovaSeq 6000 on Jan 18, 2024, using the ARTIC V5 400bp primer scheme (artic-v5.3.2_400). Lineage assignments were performed using Freyja 1.5.0 (full analysis available in file aggregated-WWSC2-ABC-b_1234.tsv). The associated contextual data record is provided below. This record is for the public health laboratory's use only, and many details were removed when sharing data according to organization-specific data sharing policies.

Template: SARS-CoV-2
specimen collector sample ID: BW-WW-12345
sampling site ID: WWSC2-ABC-b
sample collected by: Region of Peel Regional Authority
geo_loc_name (country): Canada [GAZ:00002560]
geo_loc_name (state/province/territory): Ontario [GAZ:00002563]
organism: Severe acute respiratory syndrome coronavirus 2 [NCBITaxon:2697049]
purpose of sampling: Wastewater pathogen surveillance [GENEPIO:0100872]
sample collection date: 2023-11-01
scale of sampling: Community-level surveillance [GENEPIO:0100874]
environmental site: Sewer [GENEPIO:0102064]
environmental material: Wastewater [ENVO:00002001]
populated area type: Suburban [GSSO:011077]
purpose of sequencing: Baseline surveillance (random sampling) [GENEPIO:0100005]
sequencing assay type: Amplicon sequencing assay [OBI:0002767]
sequencing date: 2024-01-18
sequenced_by: Public Health Ontario
sequenced_by_contact_name: Johnny Bloggs
sequenced_by_contact_email: jbloggs@provlab.ca
sequencing instrument: Illumina NovaSeq 6000 [GENEPIO:0100123]
amplicon pcr primer scheme: artic-v5.3.2_400 [GENEPIO:0100856]
amplicon size: 400
diagnostic measurement method: Quantitative real time polymerase chain reaction (qPCR) [OBI:0000893]
gene name: N gene (orf9) [GENEPIO:0100153]
diagnostic target presence: Diagnostic target present [GENEPIO:0100987]
diagnostic measurement value: 22
diagnostic measurement unit: cycle threshold (Ct) [GENEPIO:0100657]
lineage/clade name: B.1.617.2; B.1.2; AY.6; Q.3; EG.5; JN.1
lineage/clade analysis filename: aggregated-WWSC2-ABC-b_1234.tsv
lineage/clade analysis software name: Freyja
lineage/clade analysis software version: 1.5.0
breadth of coverage: 91

depth of coverage: 100

Discussion

As a result of WWGS's utility as an early warning system for pandemics and emerging threats, WWGS is increasingly being adopted around the world to detect and respond rapidly to serious national and international health threats (Adams et al., 2024; Manirambona et al., 2024). Despite its promising role in global surveillance, wastewater genomics-based projects and surveillance initiatives are often siloed due to heterogeneous and incompatible data structures that limit longevity and reusability. This lack of standardisation discourages investment in WWS and hinders public health responses. It also stymies efforts to develop large-scale evaluations of wastewater surveillance programs, limiting evidence for larger adoption at multiple levels. The PHA4GE wastewater specification offers laboratories an international standard with which to future-proof data for their own use, and to more seamlessly facilitate data sharing between trusted partners, within networks, and with public repositories.

Global collaboration on this specification revealed several barriers to implementation of data standards for WWGS, such as: a lack of multilingual resources, inability to customize data collection tools to better fit local needs, and the need for standardized lists of bioinformatic tools. In response to these challenges, the DataHarmonizer will be updated with new functionality to enable users to design their own data collection templates based on global data standards by selecting fields from a larger specification. While CIDGOH is currently engineering a multilingual functionality within the DataHarmonizer, collective community experience shared with PHA4GE has indications in their own efforts to integrate data specifications into new and existing tooling. In addition to the DataHarmonizer, other examples of existing data curation, harmonization, and submission tools include LexMapr (https://github.com/cidgoh/LexMapr), CDC TOSTADAS (Toolkit for Open Sequence Triage, Annotation and DAtabase Submission; https://github.com/CDCgov/tostadas), METAGENOTE (Quiñones et al., 2020), multiSub (https://github.com/maximilianh/multiSub), REDCap (Harris et al., 2019), Data-flo (https://www.data-flo.io/), and the Metadata Harmonisation Tool of DSI-Africa's eLwazi platform (https://github.com/csag-uct/Metadata-Harmonisation-Tool/tree/main).

Despite the growing body of literature exploring legal and social impacts of wastewater surveillance (Bowes et al., 2023; Hrudey et al., 2021; Nainani et al., 2024; Wardi et al., 2024), further guidance is needed in regards to the ethical use of specific data types and data elements. Wastewater contains a wide array of chemicals and genetic material that can be used to characterize communities and identify activities within them, for better or for worse (Rinde, 2023). Wastewater surveillance has been used to detect narcotics production and/or consumption, track contraceptive use, and even to identify socioeconomic disparities (Rinde, 2023). Only recently has the community begun to explore the ethical considerations of this data generation and analysis (Hrudey et al., 2021). Ethical considerations extend not only to data generation and analysis, but also to reporting of results and publicly sharing findings, as well as how to store samples for potential future use. These considerations are critical for building and maintaining trust within communities. PHA4GE has undertaken two projects to address WWGS ethics: a high-level project focusing on the socio-legal aspects of different components of wastewater, and a data-driven project exploring the benefits and potential impacts of WWGS on global public health, vulnerable populations, and remote or indigenous communities. Community partners interested in ethical aspects of genomic surveillance are welcome to join the PHA4GE Ethics and Data Sharing Working Group to explore these issues further.

The lack of best practices and community norms for the development and implementation of microbial genomics standards not only creates a patchwork of non-interoperable tools and systems but also creates uncertainty and friction for practitioners negotiating data governance policies around data management and sharing. The status quo of data standards development usually involves bringing together domain experts for brainstorming and discussions around the aims of a new standard, its scope, and its content. Generally, vocabulary is created anew and is structured according to the conventions of the subject matter area or as a result of the compromises made among the group. As such, standards are often bespoke, and there are few rules and patterns consistently applied. To address these challenges, the PHA4GE Wastewater specification uses rules and design patterns developed by the semantics community to help create consistency, implements a framework for organizing contextual information based on ISO recommendations, and provides a package of support materials to help put the standard into practice based on implementation science principles (e.g. evidence-based practice, identification of uptake and implementation barriers, consideration of individuals involved and context-dependent influences).

The framework and ontology approach implemented in this specification and in other specifications creates a shared and consistent knowledge that streamlines data expectations, enables practitioners to reuse curation skills and data management tools, and greatly reduces the development time of new standards in emergency situations as new pathogens and threats emerge. This shared approach facilitates harmonization of different data intake streams, and empowers data integration across different domains of knowledge to answer a broader array of public health questions (e.g. integrating clinical and wastewater data streams, integrating data across One Health sectors). To better communicate the importance of standards development and implementation of best practices, PHA4GE is partnering with CIDGOH and the International Pathogen Surveillance Network (IPSN) to develop training materials and toolkits, and to coordinate standards development efforts. PHA4GE has also partnered with the Global Alliance for Genomics and Health (GA4GH) to better align and integrate human and pathogen genomics standards.

Just as environmental and anthropogenic wastewater ecosystems are complex, wastewater data ecosystems are also complex: involving different partners, systems, targets, sample types, methods, governance and ethical practices, and more. The PHA4GE wastewater contextual data specification helps streamline and integrate data and practices across the local-to-global data ecosystem. The data standard was co-created with input from public repositories and laboratories from around the world, and we welcome continued feedback so that PHA4GE can update the standard as data needs evolve. To provide feedback, contact datastructures@pha4ge.org or submit an issue on GitHub via https://github.com/pha4ge/Wastewater_Contextual_Data_Specification/issues.

Availability and Requirements

The software used in this study is available on GitHub.
Project name: The PHA4GE Wastewater Contextual Data Specification
Project home page: https://github.com/pha4ge/Wastewater_Contextual_Data_Specification
Operating system: Platform independent
Programming language: Not applicable
Other requirements: None
License: MIT License

List of abbreviations

AMR, antimicrobial resistance; CDC, Center for Disease Control; CIDGOH, Centre for Infectious Disease Genomics and One Health; csv, comma-separated values; COVID, Coronavirus Disease; DDBJ, DNA Database of Japan; GA4GH, Global Alliance for Genomics and Health; EBI OLS, European Bioinformatics Institute Ontology Lookup Service; ENA, European Nucleotide Archive; FAIR, Findable, Accessible, Interoperable, Reusable; GISAID, Global Initiative on Sharing All Influenza Data; GSP, Global Sewage Project; INSDC, International Nucleotide Sequence Database Collaboration; ISO, International Organization for Standardization; JSON, JavaScript Object Notation; LinkML, Linked Open Data Modeling Language; MIxS, Minimum Information about any Sequence; NCBI, National Center for Biotechnology Information; NTR, new term request; OBO, Open Biological and Biomedical Ontology; PCR, polymerase chain reaction; PHA4GE, Public Alliance for Genomic Epidemiology; PHES-ODM, Public Health Environmental Surveillance Open Data Model; QC, quality control; SARS-CoV-2, Severe acute respiratory syndrome coronavirus 2; SOP, standard operating procedure; SRA, Sequence Read Archive; WWS, wastewater surveillance; YAML, Yet Another Markup Language.

Author statements

Authors and contributors

Conceptualization: JSP, DB, SH, EJG;
Data curation: DA, RA, JC-R, BG, MSH, JJ, WK, IS, MIN, HKA, RK, AS, VVM, ADN, KM, AM, MO, BY, PST, ST, MURK;
Funding acquisition: DB, SH, WWLH, EJG;
Methodology: CB, RC, DD, EJG;
Project administration: EJG;
Software: CB, IG, DD, EJG;

Supervision: EJG;
Writing - original draft: JSP, CB, BG, MO, EJG;
Writing - review and editing: JSP, CB, RC, DD, IG, DA, RA, JC-R, BG, MSH, JJ, WK, IS, MIN, HKA, RK, AS, VVM, ADN, KM, AM, MO, BY, PST, ST, MURK, DB, SH, IKM, LMF, AJ, EC, CO, JT, JRB, RP, JA, PW, MF, CC, CL, RW, TRC, EAS, MD, MT, DGM, MC, IM, JIL, FM, RET, WWLH, EJG.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Funding information

This work has been supported by: Public Health Agency of Canada Contribution Agreement # 2223-HQ-000265 to EJG and WH; the Gates Foundation with generous support to PHA4GE and grant 057213 to JIL; the National Science Foundation (NSF) Graduate Research Fellowship Program to JSP (grant no. 1745303) and the Hertz Foundation to JSP; the Canadian Institutes of Health Research (CIHR) Doctoral Research Award to RC (Funding Reference Number: 187519); the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health for IKM, LMF, JRB, RP,CO, AJ,EC, JT; PHA4GE subgrants to ST, BYS, HKA, JJ, PLST, VM, WK, BG, RK, JC-R.

References

Adams C, Bias M, Welsh RM, Webb J, Reese H, Delgado S, et al. The National Wastewater Surveillance System (NWSS): From inception to widespread coverage, 2020-2022, United States. Sci Total Environ 2024;924:171566.

Asghar H, Diop OM, Weldegebriel G, Malik F, Shetty S, El Bassioni L, et al. Environmental surveillance for polioviruses in the Global Polio Eradication Initiative. J Infect Dis 2014;210 Suppl 1:S294–303.

Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Res 2012;40:D57–63.

Bowes DA, Darling A, Driver EM, Kaya D, Maal-Bared R, Lee LM, et al. Structured ethical review for wastewater-based testing in support of public health. Environ Sci Technol 2023;57:12969–80.

Gennari JH, Musen MA, Fergerson RW, Grosso WE, Crubézy M, Eriksson H, et al. The evolution of Protégé: an environment for knowledge-based systems development. Int J Hum Comput Stud 2003;58:89–123.

Gill EE, Jia B, Murall CL, Poujol R, Anwar MZ, John NS, et al. The Canadian VirusSeq Data Portal and Duotang: open resources for SARS-CoV-2 viral sequences and genomic epidemiology. Microb Genom 2024;10. https://doi.org/10.1099/mgen.0.001293.

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

Gonçalves RS, Musen MA. The variable quality of metadata about biological samples used in biomedical experiments. Sci Data 2019;6:190021.

Griffiths EJ, Dooley D, Graham M, Van Domselaar G, Brinkman FSL, Hsiao WWL. Context Is Everything: Harmonization of Critical Food Microbiology Descriptors and Metadata for Improved Food Safety and Surveillance. Front Microbiol 2017;8:1068.

Griffiths EJ, Jurga E, Wajnberg G, Shay JA, Cameron R, Barclay C, et al. Crossing the streams: improving data quality and integration across the One Health genomics continuum with data standards and implementation strategies. Can J Microbiol 2025;71:1–14.

Griffiths EJ, Shay J, Cameron R, Barclay C, Sehar A, Dooley D, et al. The broom of the system: a harmonized contextual data specification for One Health AMR pathogen genomic surveillance 2024. https://doi.org/10.31219/osf.io/xbf4t.

Griffiths EJ, Timme RE, Mendes CI, Page AJ, Alikhan N-F, Fornika D, et al. Future-proofing and maximizing the utility of metadata: The PHA4GE SARS-CoV-2 contextual data specification package. Gigascience 2022;11. https://doi.org/10.1093/gigascience/giac003.

Harrington A, Vo V, Papp K, Tillett RL, Chang C-L, Baker H, et al. Urban monitoring of antimicrobial resistance during a COVID-19 surge through wastewater surveillance. Sci Total Environ 2022;853:158577.

Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform 2019;95:103208.

Hayes EK, Sweeney CL, Anderson LE, Li B, Erjavec GB, Gouthro MT, et al. A novel passive sampling approach for SARS-CoV-2 in wastewater in a Canadian province with low prevalence of COVID-19. Environmental Science: Water Research & Technology 2021;7:1576–86.

Hendriksen RS, Munk P, Njage P, van Bunnik B, McNally L, Lukjancenko O, et al. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nat Commun 2019;10:1–12.

Hrudey SE, Silva DS, Shelley J, Pons W, Isaac-Renton J, Chik AH, et al. Ethics Guidance for Environmental Scientists Engaged in Surveillance of Wastewater for SARS-CoV-2. Environ Sci Technol 2021;55. https://doi.org/10.1021/acs.est.1c00308.

Jackson RC, Balhoff JP, Douglass E, Harris NL, Mungall CJ, Overton JA. ROBOT: A tool for automating ontology workflows. BMC Bioinformatics 2019;20:407.

Jackson RC, Matentzoglu N, Overton JA, Vita R, Balhoff JP, Buttigieg PL, et al. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. Database (Oxford) 2021;2021. https://doi.org/10.1093/database/baab069.

de Jong E. Auckland lockdown to end despite three new cases of Covid-19. The Guardian 2021.

Kayikcioglu T, Amirzadegan J, Rand H, Tesfaldet B, Timme RE, Pettengill JB. Performance of methods for SARS-CoV-2 variant detection and abundance estimation within mixed population samples. PeerJ 2023;11. https://doi.org/10.7717/peerj.14596.

Keshaviah A, Diamond MB, Wade MJ, Scarpino SV, Ahmed W, Amman F, et al. Wastewater monitoring can anchor global disease surveillance systems. The Lancet Global Health 2023;11:e976–81.

Kilaru P, Larsen D, Monk D. Design and utilization of homemade wastewater samplers during the COVID-19 pandemic 2020. https://doi.org/10.31224/osf.io/frbuk.

Lambert D, Pightling A, Griffiths E, Van Domselaar G, Evans P, Berthelet S, et al. Baseline Practices for the Application of Genomic Data Supporting Regulatory Food Safety. Journal of AOAC International 2017;100. https://doi.org/10.5740/jaoacint.16-0269.

Lott MEJ, Norfolk WA, Dailey CA, Foley AM, Melendez-Declet C, Robertson MJ, et al. Direct wastewater extraction as a simple and effective method for SARS-CoV-2 surveillance and COVID-19 community-level monitoring. FEMS Microbes 2023;4:xtad004.

Manirambona E, Lucero-Prisno DE III, Shomuyiwa DO, Denkyira SA, Okesanya OJ, Haruna UA, et al. Harnessing wastewater-based surveillance (WBS) in Africa: a historic turning point towards strengthening the pandemic control. Discov Water 2024;4. https://doi.org/10.1007/s43832-024-00066-0.

Maya C, Jimenez B, Schwartzbrod J. Comparison of Techniques for the Detection of Helminth Ova in Drinking Water and Wastewater. Water Environ Res 2006;78:118–24.

Musen MA. Without appropriate metadata, data-sharing mandates are pointless. Nature 2022;609. https://doi.org/10.1038/d41586-022-02820-7.

Nainani D, Ng WJ, Wuertz S, Thompson JR. Balancing public health and group privacy: Ethics, rights, and obligations for wastewater surveillance systems. Water Res 2024;258. https://doi.org/10.1016/j.watres.2024.121756.

National Academies of Sciences, Engineering, and Medicine. Wastewater-Based Disease Surveillance for Public Health Action. National Academies Press; 2023.

Nemudryi A, Nemudraia A, Wiegand T, Surya K, Buyukyoruk M, Cicha C, et al. Temporal Detection and Phylogenetic Assessment of SARS-CoV-2 in Municipal Wastewater. Cell Rep Med 2020;1:100098.

Ong E, Xiang Z, Zhao B, Liu Y, Lin Y, Zheng J, et al. Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. Nucleic Acids Res 2017;45:D347–52.

Paul JR, Trask JD, Gard S. II. POLIOMYELITIC VIRUS IN URBAN SEWAGE. J Exp Med 1940;71:765–77.

Petros BA, Paull JS, Tomkins-Tinch CH, Loftness BC, DeRuff KC, Nair P, et al. Multimodal surveillance of SARS-CoV-2 at a university enables development of a robust outbreak response framework. Med (N Y) 2022. https://doi.org/10.1016/j.medj.2022.09.003.

Pettengill JB, Beal J, Balkey M, Allard M, Rand H, Timme R. Interpretative Labor and the Bane of Nonstandardized Metadata in Public Health Surveillance and Food Safety. Clin Infect Dis

2021;73:1537–9.

Quiñones M, Liou DT, Shyu C, Kim W, Vujkovic-Cvijin I, Belkaid Y, et al. METAGENOTE: a simplified web platform for metadata annotation of genomic samples and streamlined submission to NCBI's sequence read archive. BMC Bioinformatics 2020;21:378.

Rinde M. The Murky Ethics of Wastewater Surveillance. Science History Institute 2023. https://www.sciencehistory.org/stories/magazine/the-murky-ethics-of-wastewater-surveillance/ (accessed February 10, 2025).

Sengupta ME, Hellström M, Kariuki HC, Olsen A, Thomsen PF, Mejer H, et al. Environmental DNA for improved detection and environmental surveillance of schistosomiasis. Proc Natl Acad Sci U S A 2019;116. https://doi.org/10.1073/pnas.1815046116.

Sielemann K, Hafner A, Pucker B. The reuse of public datasets in the life sciences: potential risks and rewards. PeerJ 2020;8:e9954.

Singh S, Ahmed AI, Almansoori S, Alameri S, Adlan A, Odivilas G, et al. A narrative review of wastewater surveillance: pathogens of concern, applications, detection methods, and challenges. Front Public Health 2024;12:1445961.

Tchobanoglous G, Burton FL, Stensel H, Metcalf. Wastewater engineering : treatment and reuse. Boston: McGraw-Hill; 2003.

The OBI Consortium, Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25:1251–5.

Therrien J-D, Thomson M, Sion E-S, Lee I, Maere T, Nicolaï N, et al. A comprehensive, open-source data model for wastewater-based epidemiology. Water Sci Technol 2024;89:1–19.

Timme RE, Karsch-Mizrachi I, Waheed Z, Arita M, MacCannell D, Maguire F, et al. Putting everything in its place: using the INSDC compliant Pathogen Data Object Model to better structure genomic data submitted for public health applications. Microb Genom 2023;9. https://doi.org/10.1099/mgen.0.001145.

Timme RE, Woods J, Jones JL, Calci KR, Rodriguez R, Barnes C, et al. SARS-CoV-2 wastewater variant surveillance: pandemic response leveraging FDA's GenomeTrakr network. mSystems 2024;9:e0141523.

Vanrolleghem PA, Khalil M, Serrao M, Sparks J, Therrien J-D. Machine learning in wastewater: opportunities and challenges - "not everything is a nail!" Curr Opin Biotechnol 2025;93:103271.

Velkushanova K, Strande L, Ronteltap M, Koottatep T, Brdjanovic D, Buckley C. Methods for Faecal Sludge Analysis. IWA Publishing; 2019.

Wardi M, Belmouden A, Aghrouch M, Lotfy A, Idaghdour Y, Lemkhente Z. Wastewater genomic surveillance to track infectious disease-causing pathogens in low-income countries: Advantages, limitations, and perspectives. Environ Int 2024;192:109029.

Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res 2011;39:W541–5.

Wolfe MK, Duong D, Bakker KM, Ammerman M, Mortenson L, Hughes B, et al. Wastewater-based detection of two influenza outbreaks. Environ Sci Technol Lett 2022;9:687–92.

Wurtzer S, Levert M, Dhenain E, Boni M, Tournier JN, Londinsky N, et al. First detection of Monkeypox virus genome in sewersheds in France: The potential of wastewater-based epidemiology for monitoring emerging disease. Environ Sci Technol Lett 2022;9:991–6.

Figures and Tables Legends

Box 1: Characteristics of an equitable, interoperable, machine-readable wastewater pathogen genomics surveillance contextual data standard. PHA4GE standards are meant to be parts of public health systems rather than independent resources, and are intended to help streamline processes and data flows while supporting global equity.

Box 2: Case studies from pilot testing. Participants in the pilot testing provided feedback via a standardized Google form. To illustrate the results and experiences from pilot testing by participants, two case studies are provided. Case study one is provided by the National Scientific and Technical Research Council in Argentina (COCINET). Case study two is provided by the Kwame Nkrumah University of Science and Technology in Ghana. Both case studies report project goals, methods, and surveillance challenges, as well as PHA4GE Wastewater specification implementation, and recommendations for improving the package in the future.

Figure 1: Comparison of data modules across PHA4GE SARS-CoV-2 (blue) and Wastewater contextual data standards (red). Modularity of the standards framework enhances interoperability while enabling customization. Both standards share modules for tracking “Database identifiers”, “Sample collection and processing”, “Sequencing information”, “Bioinformatics and QC information”, “Lineage/clade information”, “Pathogen diagnostic testing”, and “Contributor acknowledgements”. The Wastewater specification also contains modules for “Environmental measurements and conditions”, “Strain and isolate information”, and “Taxonomic identification information”, whereas the SARS-CoV-2 specification contains additional modules for different types of host information (e.g. demographics, exposures, vaccination, reinfection etc). While many of the modules overlap across the specifications, there are variations in module content reflecting different surveillance needs. For example, the Wastewater specification “Database identifiers” module contains additional fields for tracking sampling events and sites (i.e. sampling site ID and sampling event ID) which are not present in the SARS-CoV-2 specification.

Figure 2: Overview of the Wastewater specification’s content. The specification provides 10 different modules (Database identifiers, Sample collection and processing, Environmental conditions and measurements, Sequence information, Bioinformatics and QC metrics, Taxonomic identification information, AMR detection information, Lineage/clade information, Pathogen diagnostic testing, Contributor acknowledgement) containing a wide variety of different standardized fields and terms. Descriptions of data types captured within the different modules are highlighted in the grey text boxes. The figure was created using BioRender.com.

Figure 3: Template-specific module distribution. The Wastewater specification is implemented in data collection templates specific for three common public health use cases - wastewater-based SARS-CoV-2 surveillance (WastewaterSARS-CoV-2 template), wastewater-based AMR surveillance (WastewaterAMR template), and general pathogen surveillance in wastewater (WastewaterPathogenAgnostic template). The templates are available in a data curation and validation tool called the DataHarmonizer. Some modules can be found in all three templates, while others are template-specific. The Database identifiers, Sample collection and processing, Environmental conditions and measurements, Sequence information, Bioinformatics and QC metrics, Pathogen diagnostic testing and Contributor acknowledgement modules are found in all templates. The Taxonomic identification information module is found in the WastewaterAMR template and the WastewaterPathogenAgnostic template; the AMR detection information module is only found in the WastewaterAMR template; and the Lineage/clade information is found only in the WastewaterPathogenAgnostic template. The figure was created using BioRender.com.

Figure 4: Screenshot of the Pathogen Agnostic template offered within the DataHarmonizer. The DataHarmonizer is a data curation, validation, and transformation tool to support pathogen surveillance contextual data management. Target-specific templates are provided in this spreadsheet-style text editor, which provides ontology-based standardized fields and dropdown menus of picklists, as well as widgets for capturing dates in ISO 8601 format. Fields are colour-coded to indicate whether they are considered “required” (yellow), “recommended (purple), or “optional” (white). Standardized null values are also provided in picklists. Curation features can be found under the Settings tab. Operating instructions can be found in the PHA4GE Wastewater specification GitHub repository. More information on the DataHarmonizer is available at https://github.com/cidgoh/DataHarmonizer, and in Gill et al, 2023.

Figure 5: Global distribution of specification pilot projects. The PHA4GE Wastewater specification was tested by labs and organizations in 15 different countries, including: Argentina, Cameroon, Canada, Côte d'Ivoire, Ecuador, Ethiopia, Ghana, India, Kenya, Mexico, Nepal, Pakistan, South Africa, the United Kingdom, and the United States of America. Feedback from these projects was incorporated into different aspects of the specification package such as vocabulary additions as well as protocol and tooling improvements.

Table 1: Best practices in specification design and implementation and their benefits. Many different practices improving machine readability, findability, accessibility, interoperability and reusability, were incorporated into the specification’s design principles. These practices included using open licenses and formats, using ontologies as sources of vocabulary and adhering to conventions within their communities of practice, integrating vocabulary from the specification in other tools such as online look-up services, etc. The culmination of these practices result in a data standard that better enables community input, integration into a wider variety of tools and surveillance programs, and more easily enables linked data.

Table 2: A list of 11 ontologies from which vocabulary used in developing the wastewater specification was sourced. All ontologies are part of the OBO Foundry library, which use a set of common design and implementation principles (https://obofoundry.org/about-OBO-Foundry.html). The OBO Foundry is overseen by an Operations Committee with Editorial, Technical and Outreach working groups.

Table 3: Required and recommended fields. A complete list of the 58 required (colour-coded yellow) and recommended (colour-coded purple) fields from the specification is provided, along with their definitions. The distribution of these fields across the three different data collection templates is also shown. Fields not included in a template are listed as Not Applicable (N/A).

Table 4: PHA4GE Wastewater specification package resource list. The specification package includes the data standard, as well as a field and term reference guides, a curation SOP, tooling (DataHarmonizer templates and operating instructions), worked examples, mapping files, and a new term request form.

Table 5: Harmonized sample descriptions. Free text sample descriptions can be complex and difficult to compare across studies and systems. Keywords from sample descriptions are highlighted, and standardized descriptors are provided in appropriate fields. The PHA4GE Wastewater specification enables separation and standardization of different concepts and data elements to make them more machine readable, and more computable (e.g. environmental materials, sites, properties and wastewater system types). Encoding sample descriptors using ontology terms, with their definitions and identifiers, also improves data sharing and future proofing of data extending its utility and longevity.

Figures and Tables

Box 1

1. Supports data collection globally
   Considering differences in surveillance aims/objectives, resources, cultural practices, sample types, population dynamics, data sensitivity, sharing policies etc.
2. Supports interoperability with other standards
3. Built according to best technical practices
   Defined by data standards communities of practice
4. Provides standardized, clearly defined vocabulary in well-structured, machine-readable formats
5. Enables sufficient capture of information to facilitate analyses pertaining to the most common public health use cases
6. Enables the identification of potential bias and/or limitations of the data
7. Acknowledges the contributions of those involved in data generation
8. Supports different kinds of data sharing
   Including different data sharing models such as sharing within organizations, within trusted networks, with public repositories
9. Is extensible, and supports reuse of data
10. Is implementable in different public health settings
11. Is supported by materials that facilitate implementation
12. Implements FAIR data practices and principles
13. Is adaptable, and is maintained over time
14. Is developed in consultation with domain experts from different countries and settings
15. Is established as a standard by consensus
16. Contributions to the standard's development are acknowledged and attributed via inclusive authorship

Box 2

Case Study 1: National Scientific and Technical Research Council of Argentina (COCINET)

Contact: Dr. Barbara Ghiglione, Universidad de Buenos Aires, Argentina
Pilot Project: Establishing a framework for wastewater-based epidemiology and AMR monitoring from wastewater.
Goals: To determine if wastewater samples accurately represent the circulation of resistant strains within the clinical setting by comparing clones that reach the environment/wastewater treatment plant with those obtained from hospitals.
Methods: 9 filtered sewage samples were cultured on MacConkey agar with and without antibiotics. Gram-negative isolates were selected. MALDI-TOF was used for species identification. 100 *Klebsiella* spp. isolates were tested for resistance via disk diffusion assays and subsequently sequenced on an Illumina NovaSeq 6000 instrument.
Partners: Paediatric Hospital (PH), General Acute Care Hospital, Sewage Treatment Plant (WWTP), National Center for Genomics and Bioinformatics (ANLIS Dr. Carlos G. Malbrán)
Target Pathogen(s): *Klebsiella* spp.
Data types: hospital and WWTP information, sample and sampling method details, AMR phenotypic testing, genomic data (QC and genomic characterization)
Data Management: Contextual data collected in various spreadsheets provided by different partners that must be integrated before analysis.
Major challenges:

1. Lack of standardized methods including selection of most appropriate targets, sampling strategy, and culture conditions
2. Lack of standardized terminology resulting in difficulties integrating data across multiple partners
3. Lack of guidance on prioritizing contextual data useful for broader community
4. Privacy issues

PHA4GE WW Specification Implementation: After participating in testing the specification, the team has proposed to leadership to conduct systematic monthly sampling using the DataHarmonizer as a communication platform among all parties involved. The team expanded the use of the Pathogen-Agnostic data collection template to include additional organisms such as *E. coli, Enterobacter*, and *Acinetobacter*. Template flexibility enables the team to systematically broaden their search for various pathogens from the same wastewater samples.
Benefits of the WWGS specification package:

1. Unified terminology across partners improved data flow
2. Improved data governance; having an international data standard in hand was useful for negotiating broader data collection and sharing practices with leadership

Recommendations:

1. Enable multilingual functionality: in particular, to include Spanish and Portuguese for South American practitioners who do not speak English
2. Enable template customization for different needs

User Feedback Quotes:
"...this framework supports standardized data collection and enables better comparisons across hospitals."

“The practical application of this package in our pilot project improved our ability to document, analyze data and use a unified vocabulary.”

Case Study 2: Kwame Nkrumah University of Science and Technology (KNUST)

Contact: Dr. Michael Owusu, Department of Medical Diagnostics, Kumasi, Ashanti Region, Ghana
Pilot Project: Using environmental surveillance of wastewater as a tool for assessing the burden of typhoid fever in the Asante Akim North District.
Goals: To evaluate the use of Moore and Grab Sampling methods for *Salmonella* Typhi detection in environmental wastewater samples in Ghana.
Methods: Wastewater samples were collected monthly from 40 sites for 12 months using Moore swabs and grab samples. Moore swabs were incubated overnight, and filtered using 0.45 um membrane filters. Grab Samples were collected into sterile wide-mouth bottles using a plastic wastewater collection cup and filtered using similar membrane filters. Both grab and Moore swab samples were extracted using Qiagen QIAamp PowerFecal Pro DNA extraction kit and tested for *Salmonella* Typhi using ttr, tviB and staG primer targets.
Reference: https://verixiv.org/articles/2-2/v1
Partners: Imperial College London; University of Washington, Seattle, USA
Target Pathogen(s): *Salmonella* Typhi
Data types: qRT-PCR, physicochemical parameters (temperature, pH, salinity, seawater specific gravity, dissolved oxygen, turbidity, electrical conductivity, oxidation-reduction potential, water depth, flow rate), and catchment population
Data Management: Contextual data collected in spreadsheets, physicochemical properties collected using RedCap.
Major challenges:

1. Lack of standardized methods to capture environmental surveillance across sites
2. Requirement for Ethics and Data Transfer Agreement for data sharing

PHA4GE WW Specification Implementation: After participating in testing the specification, the KNUST team plans to implement the PHA4GE specification and the DataHarmonizer for data management for wastewater studies and surveillance.
Benefits of the WW specification package:

1. International standard for environmental measurements and metrics
2. Data validation and version control
3. Automated exports in different NCBI formats

Recommendations:

1. Provide a Data Transfer Agreement template
2. Include additional fields for strain identification, microbial organism identity, and lab procedure details

Figure 1

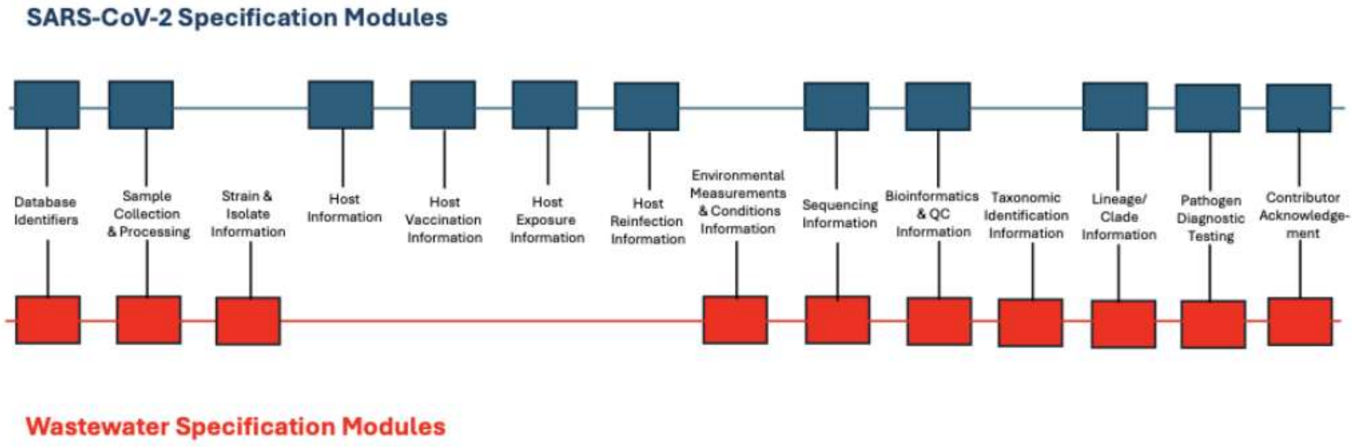

SARS-CoV-2 Specification Modules
Database Identifiers
Sample Collection & Processing
Strain & Isolate Information
Host Information
Host Vaccination Information
Host Exposure Information
Host Reinfection Information
Environmental Measurements & Conditions Information
Sequencing Information
Bioinformatics & QC Information
Taxonomic Identification Information
Lineage/ Clade Information
Pathogen Diagnostic Testing
Contributor Acknowledge-ment
Wastewater Specification Modules

Figure 2

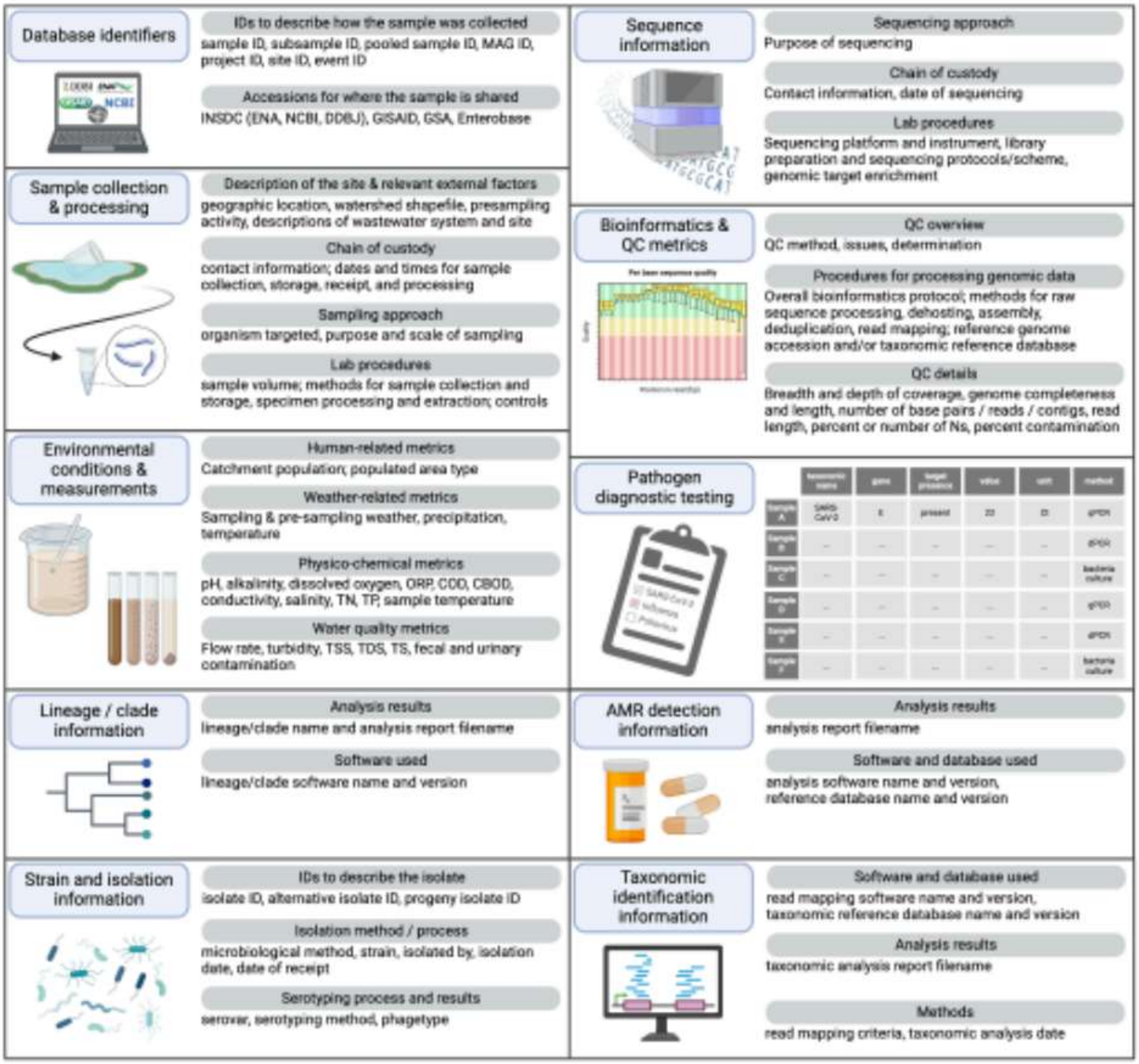

Database identifiers
IDs to describe how the sample was collected
sample ID, subsample ID, pooled sample ID, MAG ID, project ID, site ID, event ID
Accessions for where the sample is shared
INSDC (ENA, NCBI, DDBJ), GISAID, GSA, Enterobase
Sample collection & processing
Description of the site & relevant external factors
geographic location, watershed shapefile, presampling activity, descriptions of wastewater system and site
Chain of custody
contact information; dates and times for sample collection, storage, receipt, and processing
Sampling approach
organism targeted, purpose and scale of sampling
Lab procedures
sample volume; methods for sample collection and storage, specimen processing and extraction; controls
Environmental conditions & measurements
Human-related metrics
Catchment population; populated area type
Weather-related metrics
Sampling & pre-sampling weather, precipitation, temperature
Physico-chemical metrics
pH, alkalinity, dissolved oxygen, ORP, COD, CBOD, conductivity, salinity, TN, TP, sample temperature
Water quality metrics
Flow rate, turbidity, TSS, TDS, TS, fecal and urinary contamination
Lineage / clade information
Analysis results
lineage/clade name and analysis report filename
Software used
lineage/clade software name and version
Strain and isolation information
IDs to describe the isolate
isolate ID, alternative isolate ID, progeny isolate ID
Isolation method / process
microbiological method, strain, isolated by, isolation date, date of receipt
Serotyping process and results
serovar, serotyping method, phagetype
Sequence information
Sequencing approach
Purpose of sequencing
Chain of custody
Contact information, date of sequencing
Lab procedures
Sequencing platform and instrument, library preparation and sequencing protocols/scheme, genomic target enrichment
Bioinformatics & QC metrics
Per base sequence quality
QC overview
QC method, issues, determination
Procedures for processing genomic data
Overall bioinformatics protocol; methods for raw sequence processing, dehosting, assembly, deduplication, read mapping; reference genome accession and/or taxonomic reference database
QC details
Breadth and depth of coverage, genome completeness and length, number of base pairs / reads / contigs, read length, percent or number of Ns, percent contamination
Pathogen diagnostic testing
SARS-CoV-2
Influenza
Poliovirus
Sample A
SARS-CoV-2
present
qPCR
bacteria culture
AMR detection information
Analysis results
analysis report filename
Software and database used
analysis software name and version, reference database name and version
Taxonomic identification information
Software and database used
read mapping software name and version, taxonomic reference database name and version
Analysis results
taxonomic analysis report filename
Methods
read mapping criteria, taxonomic analysis date

Figure 3

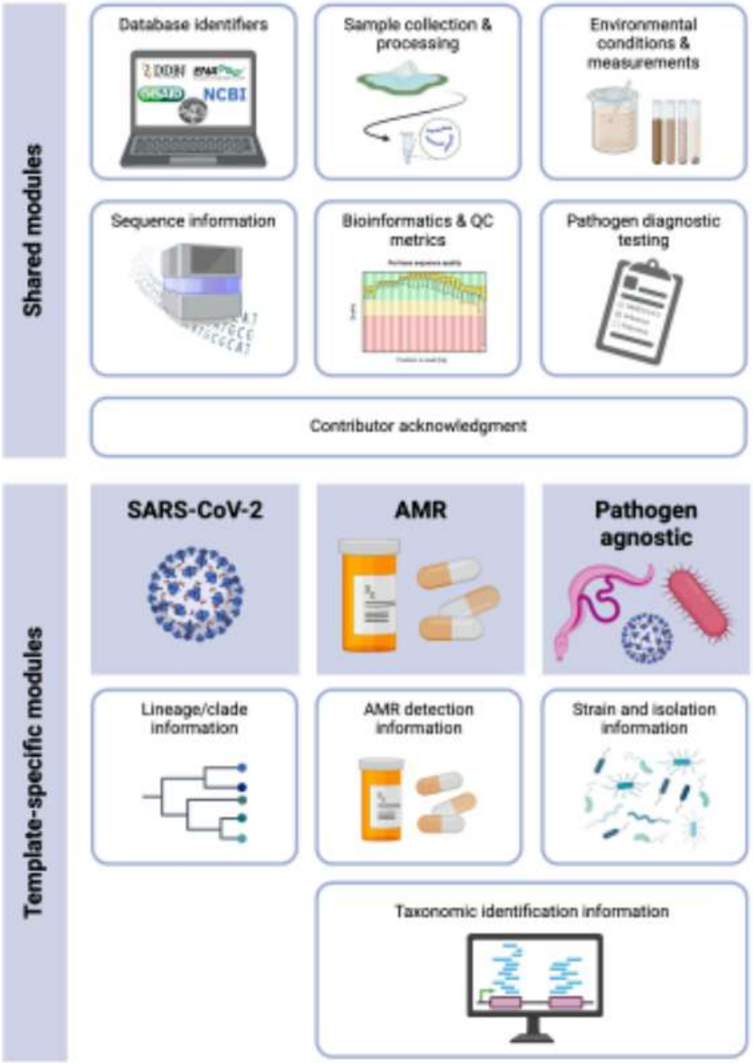
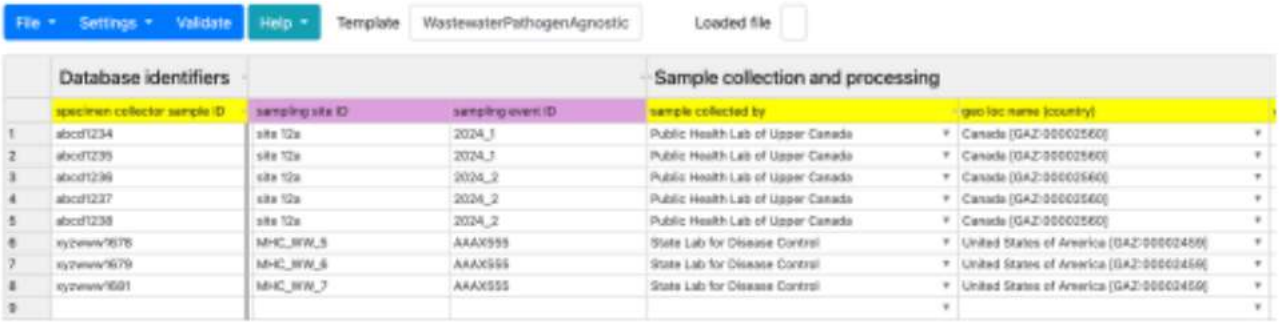

Shared modules
Database identifiers
DDBJ
ENA
GISAID
NCBI
Sample collection & processing
Environmental conditions & measurements
Sequence information
Bioinformatics & QC metrics
Pathogen diagnostic testing
Contributor acknowledgment
Template-specific modules
SARS-CoV-2
AMR
Pathogen agnostic
Lineage/clade information
AMR detection information
Strain and isolation information
Taxonomic identification information

Figure 4



Template: WastewaterPathogenAgnostic

Loaded file

	Database identifiers			Sample collection and processing	
	specimen collector sample ID	sampling site ID	sampling event ID	sample collected by	geo loc name (country)
1	abcd1234	site 12a	2024_1	Public Health Lab of Upper Canada	Canada [GAZ:00002560]
2	abcd1235	site 12a	2024_1	Public Health Lab of Upper Canada	Canada [GAZ:00002560]
3	abcd1236	site 12a	2024_2	Public Health Lab of Upper Canada	Canada [GAZ:00002560]
4	abcd1237	site 12a	2024_2	Public Health Lab of Upper Canada	Canada [GAZ:00002560]
5	abcd1238	site 12a	2024_2	Public Health Lab of Upper Canada	Canada [GAZ:00002560]
6	xyzwww1678	MHC_WW_5	AAAX555	State Lab for Disease Control	United States of America [GAZ:00002459]
7	xyzwww1679	MHC_WW_6	AAAX555	State Lab for Disease Control	United States of America [GAZ:00002459]
8	xyzwww1681	MHC_WW_7	AAAX555	State Lab for Disease Control	United States of America [GAZ:00002459]
9					

Figure 5

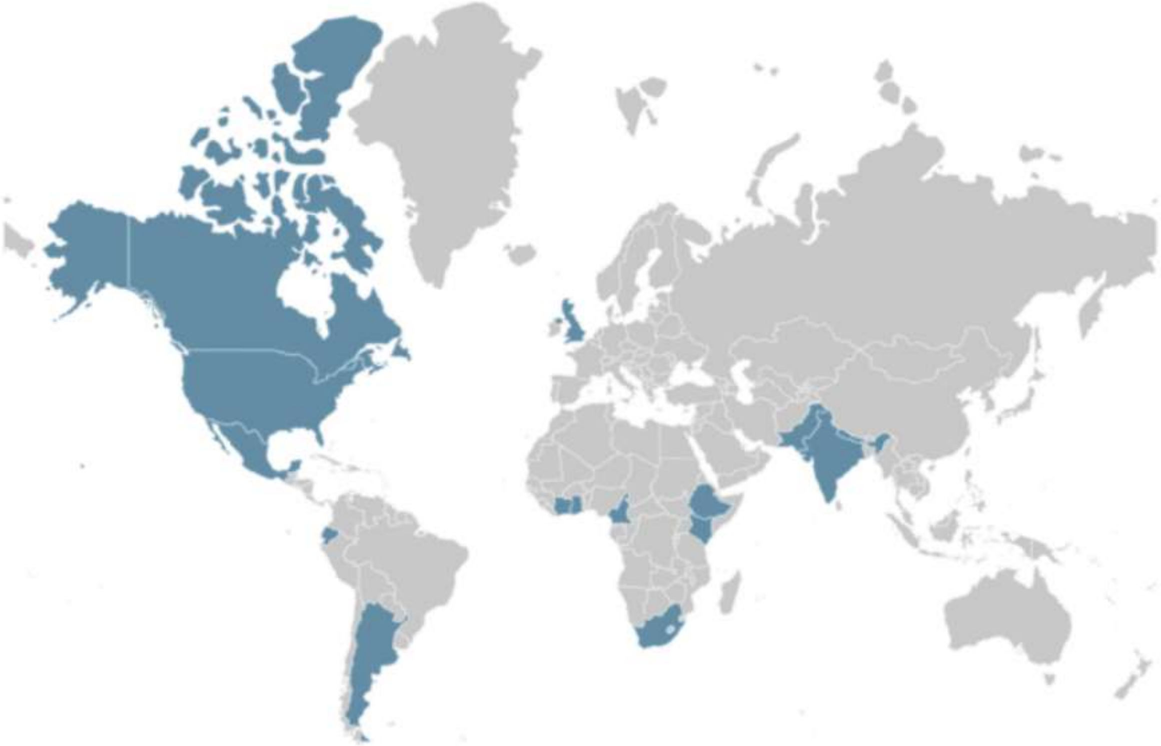


Table 1

Best Practices	**Benefits**
Specifications are freely available on the Web using an open license.	● Supports equitable access to the standard.
Make machine-readable specifications available in different formats, e.g. JSON, YAML, csv (e.g. via LinkML).	● The standard is available in open (rather than proprietary) formats. ● Specifications can be more easily integrated into systems and tools. ● More standards-based data management tools are needed. Machine-amenable specifications facilitate the development of such tools.
Source terminology from open source, version-controlled, international, semantic resources e.g. OBO Foundry ontologies.	● Vocabulary is transparently developed, made publicly available, searchable via different indexed portals and look-up services (e.g. EBI-OLS, BioPortal, Ontobee). ● Vocabulary is updated as needed over time, and contact information for developers is provided for community input. ● Vocabulary is well annotated with database cross-references, curation notes, synonyms, etc. ● Enables linked data (e.g. ontologies support knowledge graphs to enable more complex queries, inferencing, and broader data integration).
Disambiguate meanings of terminology using universal identifiers (e.g. IRIs).	● Vocabulary is better defined and the meaning of annotated data is clearer. ● The use and inclusion of identifiers makes data more machine-readable compared to data that is annotated with term labels alone, which may vary according to institutional preferences.
When developing new vocabulary: ● Adhere to established naming conventions / semantic patterns. ● Avoid pre-composed terms (separate nouns and descriptors). ● Avoid abbreviations in the label. ● Avoid questions and boolean values in data collection (Y/N, T/F). ● Use technical (not colloquial) terms.	● Data elements are easier to index, catalogue, and compare across studies and surveillance initiatives.
Facilitate reusability by considering use cases beyond immediate scope during	● Data elements are as universal as possible but as specific as necessary. (Note: an attribute

development of terminology.	may appear to be highly generalizable within a particular use case, but once other use cases are considered, language may need to be more specific to disambiguate different meanings and usage). - Avoiding organization-specific usage of terminology and can enable data to be reused across different domains of knowledge and across different projects.
Use tags (attributes) to richly annotate data with methods and provenance information.	- Data is more comparable across projects. - Data can be more accurately interpreted because limitations of data can be better understood.

Table 2

Ontologies	Domain	URL
EFO	Experimental factor	https://www.ebi.ac.uk/efo/
ENVO	Environments	https://obofoundry.org/ontology/envo.html
GAZ	Geography	https://obofoundry.org/ontology/gaz.html
GENEPIO	Genomic epidemiology	https://obofoundry.org/ontology/genepio.html
GSSO	Gender, sex, sexual orientation	https://obofoundry.org/ontology/gsso.html
IDO	Infectious disease	https://obofoundry.org/ontology/ido.html
NCBITaxon	Taxonomy	https://obofoundry.org/ontology/ncbitaxon.html
NCIT	National Cancer Institute Thesaurus	https://obofoundry.org/ontology/ncit.html
OBI	Biomedical investigations and assays	https://obofoundry.org/ontology/obi.html
PATO	Phenotypic qualities	https://obofoundry.org/ontology/pato.html
UO	Units of measurement	https://obofoundry.org/ontology/uo.html

Table 3

Field	Template			Definition
	SARS-CoV-2	AMR	Pathogen agnostic	
Database Identifiers				
specimen collector sample ID	🟨	🟨	🟨	The user-defined name for the sample.
BioSample accession	🟪	🟪	🟪	The identifier assigned to a BioSample in INSDC (i.e., ENA, NCBI, or DDBJ) archives.
sampling site ID	🟪	🟪	🟪	The user-defined identifier assigned to a specific location from which samples are taken.
sampling event ID	🟪	🟪	🟪	The user-defined identifier assigned to a specific event during which one or more samples are taken, from one or more sites.
Sample collection and processing				
sample collected by	🟨	🟨	🟨	The name of the organization with which the sample collector is affiliated.
geo loc name (country)	🟨	🟨	🟨	The country of origin of the sample.
geo loc name (state/province/territory)	🟨	🟨	🟨	The state/province/territory of origin of the sample.
organism	N/A	🟪	🟨	Taxonomic name of the organism.
purpose of sampling	🟨	🟨	🟨	The reason that the sample was collected.
scale of sampling	🟪	🟪	🟪	The range of locations or entities sampled expressed in general terms.
sample collection start date	🟨	🟨	🟨	The date on which the sample was collected, or sampling began for a continuous sample.
sample collection end date	🟪	🟪	🟪	The date on which sample collection ended for a continuous sample.
sample collection start time	🟪	🟪	🟪	The time at which sample collection began.
sample collection end time	🟪	🟪	🟪	The time at which sample collection ended.
sample collection time duration value	🟪	🟪	🟪	The amount of time over which the sample was collected.
sample collection time duration unit	🟪	🟪	🟪	The units of the time duration measurement of sample collection.
environmental site	🟪	🟪	🟪	An environmental location may describe a site in the natural or built environment e.g. contact surface, metal can, hospital, wet market, bat cave.
environmental material	🟪	🟪	🟪	A substance obtained from the natural or man-made environment e.g. soil, water, sewage.
environmental material properties	🟪	🟪	🟪	The properties, characteristics and qualities of a substance obtained from the natural or man-made environment.
wastewater system type	🟪	🟪	🟪	The type or classification of a wastewater system e.g. sanitary sewer, combined sewer, latrine
collection method	🟪	🟪	🟪	The process used to collect the sample.
endogenous control details	🟪	🟪	🟪	The description of the endogenous controls included when extracting a sample.
Strain and isolation information				
microbiological method	N/A	N/A	🟪	The laboratory method used to grow, prepare, and/or isolate the microbial isolate.
isolate ID	N/A	N/A	🟨	The user-defined identifier for the isolate, as provided by the laboratory that originally isolated the isolate.
alternative isolate ID	N/A	N/A	🟪	An alternative isolate_ID assigned to the isolate by another organization.
serovar	N/A	N/A	🟪	The serovar of the organism.

serotyping method	N/A	N/A	🟪	The method used to determine the serovar.
Environmental conditions and measurements				
water catchment area human population measurement value	🟪	🟪	🟪	The numerical value of the human population measurement that contributes to the composition of water in a catchment area.
precipitation measurement value	🟪	🟪	🟪	The amount of water which has fallen during a precipitation process.
precipitation measurement unit	🟪	🟪	🟪	The units of measurement for the amount of water which has fallen during a precipitation process.
turbidity measurement value	🟪	🟪	🟪	The numerical value of a measurement of turbidity.
turbidity measurement unit	🟪	🟪	🟪	The units of a measurement of turbidity.
fecal contamination indicator	🟪	🟪	🟪	A gene, virus, bacteria, or substance used to measure the sanitary quality of water in regards to fecal contamination.
fecal contamination value	🟪	🟪	🟪	The numerical value of a measurement of fecal contamination.
fecal contamination unit	🟪	🟪	🟪	The units of a measurement of fecal contamination.
Sequence Information				
purpose of sequencing	🟨	🟨	🟨	The reason that the sample was sequenced.
sequenced by	🟨	🟨	🟨	The name of the agency, organization or institution responsible for sequencing the isolate's genome.
sequenced by contact name	🟨	🟨	🟨	The name or title of the contact responsible for follow-up regarding the sequence.
sequenced by contact email	🟨	🟨	🟨	The email address of the contact responsible for follow-up regarding the sequence.
sequence submitted by	🟪	🟪	🟪	The name of the agency that submitted the sequence to a database.
sequence submitter contact email	🟪	🟪	🟪	The email address of the contact responsible for follow-up regarding the sequence.
sequencing instrument	🟨	🟨	🟨	The model of the sequencing instrument used.
sequencing assay type	🟪	🟪	🟪	The overarching sequencing methodology that was used to determine the sequence of a biomaterial.
sequencing protocol	🟪	🟪	🟪	The protocol or method used for sequencing.
genomic target enrichment method	N/A	🟪	🟪	The molecular technique used to selectively capture and amplify specific regions of interest from a genome.
amplicon pcr primer scheme	🟪	N/A	🟪	The specifications of the primers (primer sequences, binding positions, fragment size generated etc) used to generate the amplicons to be sequenced.
Bioinformatics and QC metrics				
raw sequence data processing method	🟪	🟪	🟪	The method used for raw data processing such as removing barcodes, adapter trimming, filtering etc.
dehosting method	🟪	🟪	🟪	The method used to remove host reads from the pathogen sequence.
Taxonomic identification information				
read mapping software name	N/A	🟪	🟨	The name of the software used to map sequence reads to a reference genome or set of reference genes.
read mapping software version	N/A	🟪	🟨	The version of the software used to map sequence reads to a reference genome or set of reference genes.
taxonomic reference database name	N/A	🟪	🟪	The name of the taxonomic reference database used to identify the organism.
taxonomic reference database version	N/A	🟪	🟪	The version of the taxonomic reference database used to identify the organism.
AMR detection information				
AMR analysis software name	N/A	🟪	N/A	The name of the software used to perform an in silico antimicrobial resistance determinant identification/analysis.
AMR analysis software version	N/A	🟪	N/A	The version number of the software used to perform an in silico antimicrobial resistance determinant identification/analysis.

AMR reference database name	N/A		N/A	The name of the reference database used to perform an in silico antimicrobial resistance determinant identification/analysis.
AMR reference database version	N/A		N/A	The version number of the reference database used to perform an in silico antimicrobial resistance determinant identification/analysis.
AMR analysis report filename	N/A		N/A	The filename of the report containing the results of an in silico antimicrobial resistance analysis.
Contributor acknowledgement				
authors				Names of individuals contributing to the processes of sample collection, sequence generation, analysis, and data submission.

Table 4

Resource	**Description**	**Directory URL**
DataHarmonizer data management templates (PathogenAgnostic, AMR, SARS-CoV-2)	Spreadsheet-based contextual data collection forms containing controlled vocabulary and prescribed formats. Fields are colour-coded to indicate required (yellow), recommended (purple), or optional status (white).	https://github.com/cidgoh/pathogen-genomics-package/releases
DataHarmonizer Operating instructions	Step-by-step instructions for downloading and using the DataHarmonizer	https://github.com/pha4ge/Wastewater_Contextual_Data_Specification/tree/main/SOPs
Curation protocol	Data curation and guidance for populating the data collection forms. Ethical, practical, and privacy considerations are highlighted.	https://github.com/pha4ge/Wastewater_Contextual_Data_Specification/tree/main/SOPs
Field and term reference guides	Definitions of controlled vocabulary (fields and terms), as well as examples and additional guidance for structuring data.	https://github.com/pha4ge/Wastewater_Contextual_Data_Specification/tree/main/Reference%20Guide
Directory of mapping files of the PHA4GE Wastewater specification to other attribute packages and public repository requirements	Mappings of PHA4GE fields to equivalents in NCBI's wastewater BioSample package (SARS-CoV-2.wwsurv.1.0), ENA's sludge (ERC000023) and sewage (ERC000036) BioSample packages, the PHES-ODM (https://docs.phes-odm.org/), and GISAID Wastewater submission requirements.	https://github.com/pha4ge/Wastewater_Contextual_Data_Specification/tree/main/Mappings
Worked examples and DataHarmonizer representations	Data from worked examples structured using standardized fields and terms in DataHarmonizer templates. Three files are provided - one for examples curated using each template (i.e. PathogenAgnostic, SARS-CoV-2, AMR). Mock data has been used in these examples.	https://github.com/pha4ge/Wastewater_Contextual_Data_Specification/tree/main/Testing%20and%20Validation
New Term Request form	Templates in the GitHub Issuetracker enabling single and bulk field or term requests. Community requests for new terminology are welcome in order to adapt the standard for evolving data needs.	https://github.com/pha4ge/Wastewater_Contextual_Data_Specification/tree/main/SOPs

INSDC data submission protocols	PHA4GE protocols for setting up INSDC repository accounts, setting up BioProjects, preparing BioSample records, and sequence deposition for NCBI (BioSample and SRA) and ENA (BioSample and Experimental data) can be found at Protocols.io. Protocols can be adapted from SARS-CoV-2 to Wastewater by simply choosing the appropriate BioSample package when submitting data. The DataHarmonizer can be used to help prepare files in the appropriate BioSample formats.	https://www.protocols.io/workspaces/pha4ge/publications

Table 5

Original Sample Description	Environmental site	Environmental material	Environmental material properties	Wastewater system type
The treated effluent from a wastewater treatment plant, to which piped systems transport a combination of wastewater and run-off	Wastewater treatment plant [ENVO:00002272]	Wastewater effluent [ENVO:03501457]	Treated [GENEPIO:0101008]	Combined sewer system [ENVO:03501453]
A latrine outside of a school building	School [ENVO:03501130]	Wastewater [ENVO:00002001]		Latrine [ENVO:01000519]
A lake known to be contaminated with fecal matter	Lake [ENVO:00000020]	Surface water [ENVO:00002042]	Fecal-contaminated [GENEPIO:0101010]	
Sludge from a wastewater stabilization lagoon that services a local community		Sludge [ENVO:00002044]		Waste stabilization pond [ENVO:03600076]
Wastewater from an airplane's lavatory	Airplane [ENVO:03501349]; Toilet [ENVO:01000516]	Wastewater [ENVO:00002001]		
A manhole, within a piped wastewater-only sewage system, located at the receiving point of slow-flowing effluent from a homeless shelter	Homeless shelter [ENVO:03501133]	Wastewater [ENVO:00002001]	Fluid (slow) [GENEPIO:0101005]	Sanitary sewer system [ENVO:03501454]
A composting toilet located at the convergence of multiple farms	Farm [ENVO:00000078]	Wastewater [ENVO:00002001]	Semi-solid [NCIT:C149895]	Composting toilet [ENVO:01000550]
An open channel, in which people are known to defecate, that runs through a refugee camp	Refugee camp [NCIT:C85867]	Surface water [ENVO:00002042]	Fecal-contaminated [GENEPIO:0101010]	Drainage channel