PHA4GE – Wastewater

Contextual Data Template User Guide and SOP 1.0

I. **Purpose**: To structure wastewater contextual data according to the PHA4GE contextual data collection templates in order to better enable harmonization across datasets and systems.
   A. Data providers will curate contextual data according to the steps outlined in the procedure below.
   B. Data providers will populate the appropriate template with information from their datasets using applicable picklists and according to instructions.
   C. Data providers can share the harmonized data according to jurisdictional and organization-specific data sharing policies.

II. **Data**: Contextual data describing repository accession numbers, sample collection and processing, environmental conditions and measurements, sequencing, and bioinformatics and QC metrics, taxonomic identification, AMR detection, lineage/clades, pathogen diagnostic testing, and contributor acknowledgements, as supplied by the data provider.

Note: Different subsets of fields will apply to samples from different contexts (i.e., structured sewage systems v. contaminated surface waters). Fields not pertinent to the sample type of interest need not be filled. Refer to the Appendix B and the Reference Guide for further instructions regarding different sample types.

III. **Procedure**:

	Action
1	Download the file containing the reference guide from the github repository. To use the DataHarmonizer as your data collection instrument follow the instructions for using the DataHarmonizer. You can download the latest wastewater DataHarmonizer templates at the pathogen genomics package.
2	Before you begin to curate your contextual data: ● Review your dataset ● Review the fields and values in the template. ● Review the field definitions and guidance in the template Reference Guide.
3	Confirm the planned mapping of your data fields to those in the PHA4GE collection template with the data steward (e.g. your supervisor). *Note: Confirm the level of granularity of information that can be shared publicly and/or privately, with the data steward and/or your privacy officer. The most detailed information allowable should be included here. Different versions (detailed information vs general information) can be stored.*

4	Populate the collection template with the information from your dataset. ● Fields color-coded yellow are considered mandatory. Fill these in first. ● Fields color-coded purple are strongly recommended. If you have permission, fill these fields in next. ● Fields color-coded white are optional, but still important. If you have permission, fill in these fields. ● Use picklists where provided. ● Ensure the data is stored safely with appropriate encryption. *Note: Sometimes there will be constraints on what information can be shared, other times a field may not be applicable to your sample. Use the null values (controlled vocabulary indicating the reason why information is not provided) in the picklist to report missing data. (an asterisk "*" denotes exceptions between templates)*
5	Use the wastewater contextual data Reference Guide to access field definitions, field-level guidance and examples.

Subsection	Required Fields
Database Identifiers	specimen collector sample ID
Sample Collection and Processing *Note: Consult your supervisor and/or data steward to evaluate whether the specimen collector sample ID is considered identifiable according to your institutional policies. If not considered identifiable, copy the sample ID into the "specimen collector sample ID" field in the collection template. If considered identifiable, provide the alternative sample ID. Be sure to keep a copy of the key in a safe location.*	sample collected by geo_loc_name (country) geo_loc_name (state/province/territory) organism* (not required for AMR template) purpose of sampling sample collection date
Strain and isolate information	isolate ID* (only required for Pathogen Agnostic template)
Sequence Information	purpose of sequencing sequenced by sequenced by contact name sequenced by contact email sequencing instrument
Taxonomic identification information	read mapping software name* (only required for Pathogen Agnostic template) read mapping software version* (only required for Pathogen Agnostic template)

	See **Appendix A** for ethical and privacy considerations of contextual data. See **Appendix B** for examples of how to structure sample descriptions. If a desired term is not present in a picklist, you can request standardized terms using the procedure in **Appendix C**.

IV. Appendix A: Ethical, Practical, and Privacy Considerations

Effective and equitable pathogen surveillance via wastewater sampling requires rapid and sustained international collaboration and data sharing. Many of the contextual data elements described in the PHA4GE wastewater contextual data specification are critical for effective public health surveillance and response. However, many of these same elements have ethical, practical, and privacy issues which must be considered before data can be shared. Data governance policies may vary between data types and jurisdictions, thus users of the specification should consult data stewards and privacy officers regarding organization-specific and jurisdiction-specific policies. Below, we highlight a series of common issues and provide suggestions for ways forward. The PHA4GE Reference Guide should be consulted for field-level guidance.

Note: This guidance is based on the experience of members of the PHA4GE working groups, and is not intended to apply to all situations and use cases. Decisions regarding implementation of the specification must ultimately be made by the user in consultation with data providers and data stewards. If the intended use of the information collected is for research purposes, there will likely be many additional administrative and ethical requirements (e.g. Research Ethics Board (REB) review).

Use Case-specific Templates

Wastewater genomic surveillance data can have many applications, and be used to explore data in different ways in order to answer a wide variety of public health questions. Often, the data that is collected is shaped by these priorities, and the questions that an organization is trying to answer. As such, contextual datasets can vary in content, granularity, and structure. To address this variability in data collection, to provide as much flexibility as possible, and to reduce unnecessary data entry, the PHA4GE Wastewater Specification package contains different templates scoped for different widely used use-cases. These use-cases include wastewater-based SARS-CoV-2 surveillance (SARS-CoV-2 template); wastewater-based antimicrobial resistance detection, surveillance, and research (AMR template); and detection and surveillance of other target pathogens (Pathogen Agnostic template). The templates are scoped for capture of contextual information used in amplicon-based sequencing, single isolate whole genome sequencing, as well as metagenome sequencing. The specification package captures information about sequence reads, consensus sequences and assemblies from cultured isolates, selected results from analyses (e.g. taxonomic identification of pathogens present in a wastewater sample), and the names of reports where further results can be obtained. The content of the different templates overlap, but there are important nuanced differences. **Users should identify the most appropriate template for their data. Users can**

also create their own spreadsheets and templates by recombining selected fields, terms and modules from the PHA4GE specification according to their needs. A JSON file containing all possible vocabulary (the “master wastewater specification”) is also available if the user would prefer to create their own custom template from the master specification - this is encouraged for bespoke applications. More templates may be added over time, and requests can be suggested by contacting datastructures@pha4ge.org.

Identifiers and Repository Accession Numbers

Laboratories world-wide are collecting and analyzing wastewater data; however, while it is common to share clinical sequence and minimal contextual data with public repositories such as GISAID and the INSDC, this is not yet a common practice following wastewater surveillance. Through sharing consensus sequence and raw data, as well as contextual data, with public repositories, we can:

- Detect diverse pathogens from viral, bacterial, and other backgrounds
- Guide vaccination campaigns (ex: polio)
- Develop and improve diagnostic technologies for wastewater samples
- Supplement clinical surveillance in order to understand community rates of infection and predict hospitalization rates
- Inform investment in water sanitation and hygiene

When you share information with a public database, you will receive an accession number (a unique identifier in a database enabling the tracking of multiple versions of the data). If you have shared data with a public database, make sure to capture the accession numbers. GISAID will provide you with a single accession number. Make sure to record it. INSDC members (NCBI, ENA, DDBJ) may provide you with different accession numbers depending on what you share, and how. You can share assemblies and consensus sequences with GenBank (and its equivalents), raw data with Sequence Read Archive (SRA), and contextual data as a BioSample (see reference guide for further information). Information may be organized in BioProjects, and at a higher organizational level, Umbrella BioProjects. Make sure to record all of the applicable accession numbers.

Samples, libraries, sites, sequences (raw, processed, consensus etc) and so on can have many identifiers, especially if there is a division of labor or sharing of information across agencies and organizations. The specification has provided fields to capture many of those that are common, but may not capture all of the IDs you require. **It is essential to track IDs of original materials and information** to establish chain-of-custody and for follow-up, if necessary. It is better to track too many IDs than too few. If you require more fields to capture the IDs you need, add them. Some IDs are considered public health identifiable information (PHII). Make sure to check with the appropriate authorities whether the IDs you plan to share are considered identifiable information. If considered identifiable, you may need to create an alternative set of IDs. If you do, make sure to store the key in a safe and secure place.

Geographical Information

Geographical information (country, province/state/region, city, postal code, latitude/longitude etc) is an essential descriptor of a wastewater site. However, when coupled

with other metadata, this information may incidentally reveal the identities of individuals or a community contributing effluent to a given site. Especially for sites with smaller catchment areas, it may be necessary to abstract geographic information before it can be shared. Before sharing data, especially with public repositories, it is important to ensure the data being submitted complies with the permitted level of granularity. Discuss this with the data steward.

If sharing latitude and longitude coordinates, do not use the center of the region or the location of your agency as a proxy, as this implicates a real location and is misleading. Latitude and longitude should either be shared accurately or kept private.

Date Information

Geographical and temporal information are key elements of infectious disease surveillance programs. Temporal information consists of dates e.g. sample collection date, sample received date, sample sequenced date, etc. Elements such as “sample collection date” are usually held by the institution that collected the original specimen (e.g. performed the diagnostic test). As such, you may require permission to acquire this information, or it may be difficult to attain due to other burdens on the data provider (workload, system access, manual curation requirements). Alternatively, ”received date” may be used as a substitute in the data you share. In such a situation, it is more accurate to provide a null value for “sample collection date” and provide the date the “received date” in the “received date” field to avoid misinterpretation in analyses.

Purpose of Sampling/Purpose of Sequencing

Sampling strategy can create biases in the data. A sample may be collected for one purpose, but sequenced for another (e.g. collected for diagnostic testing, but sequenced for surveillance of circulating lineages and variants). Information about why samples were collected and why they were selected for sequencing (i.e random vs targeted sampling) can help inform epidemiological modeling and analyses. Standardized tags are available in the “purpose of sampling” field (e.g. protocol testing, wastewater drug surveillance) and in the “purpose of sequencing” field (e.g. baseline surveillance (non-random sampling), travel-associated surveillance). Free text fields are also available for providing extra information about sampling and the selection of samples for sequencing called “purpose of sampling details” and “purpose of sequencing details”. A number of standardized phrases are also suggested in the Reference Guide for describing different common surveillance priorities.

Sewershed Information

Sewershed information, when combined with other metadata, can potentially be used to identify a specific community contributing effluent to a site. For example, if a site is described as a “correctional facility,” this could be combined with geographic information to positively identify a specific prison. Especially when a community is smaller or disadvantaged, the potential to identify a specific community should be taken into account when deciding which metadata to share with collaborators or publicly.

Methods Information

Methodological information, such as sampling and experimental design, laboratory procedures, bioinformatic processing, and quality control metrics, are crucial information to understand the context and limitations of analyses. Capturing as much well-structured information regarding your methods, and storing it in a centralized place (or single document) helps to futureproof the data as well as the work that went into collecting, processing, analyzing and interpreting the data. Capturing methodological information also enables better reproducibility, and increases quality control. The specification provides many fields for capturing experimental design, protocols, and scientific metrics. It is strongly recommended that as much of that information be captured and stored as possible. It should be noted that the template, while designed to capture critical information about wastewater surveillance, is not meant to function in the same way as a Laboratory Information Management System (LIMS). Far more detailed methods may be included in LIMS (e.g. concentrations, values, measurements used for preparing libraries). The PHA4GE templates serve as a summary of this information. However, in the absence of a formal LIMS, these templates can be used for tracking and storing methodological information.

It should be noted that some methods fields appear to require the same information (e.g. consensus sequence software name vs sequence assembly software name). These fields do not all need to be filled in. They represent different bioinformatics processes, and as such, only the fields that pertain to your analysis need to be populated (i.e. your consensus sequences should have the consensus software information populated). If you are unclear about which fields to use, contact the data steward for clarification.

Granularity of Metrics - Exact Values vs Ranges

While it is ideal for measurements to be as precise as possible for comparisons across studies, sometimes only ranges of values are available. The templates are designed to enable flexible capture of data across labs that have different public health priorities, capacities, and data needs. Exact values and ranges should be expressed in separate fields, as provided (e.g. water catchment area human population **measurement value** vs water catchment area human population **bin**). Values, units, and methods should also be captured in separate fields. It should be noted that the methods of environmental measurements are important for the interpretation of the data, and so should be provided as much as possible.

Attribution

The generation of genomic sequences from samples requires a lot of work, and can involve contributions form different partners - from the acquisition of funding, to sample collection, to sample processing and sequencing, to analysis and interpretation, to sequence and contextual data management and submission. There are many roles and tasks along the genomic surveillance continuum and it is important to track and acknowledge the contributions of all those involved. Attribution of contributions can be critical for demonstrating productivity for securing future funding, for establishing chains of custody, for ensuring contact information is available for follow-up, and is essential for ethical benefit sharing in partnerships. Discussions regarding how individuals and/or organizations should be attributed in collections (private/public), manuscripts and other published or non-published work, should be discussed prior to beginning any project. The “authors” field in the template can be populated with a list of

contributors as well as short descriptions of their contributions. Data stewards may want to consider the inclusion of data license tags (fields and values denoting data use permissions/restrictions) in their records (note: data license tags not included in the templates).

Multi-Tagging

Users may find that multiple values are necessary in some fields. The templates provide "multi-tagging" capability enabling users to enter multiple values separated by a semi-colon. Users who create their own templates from the PHA4GE master list of vocabulary and enable multi-tagging should also use semicolons as separators as this is the convention preferred by large public repositories.

Null Values

The International Nucleotide Database Collaboration (INSDC) have created standardized missing/null value reporting language to be used where a value of an expected format for sample metadata reporting can not be provided. This controlled vocabulary has been adopted in this specification, and takes into account different types of constraints (i.e. Not Applicable, Missing, Not Collected, Not Provided, Restricted Access). Users are strongly encouraged to always provide as much information as possible in the collection template, however, if missing/null value reporting is required, users are asked to use a term with the finest granularity for their situation. Note: NCBI accepts all null values. ENA will accept any other null value besides "Missing".

V. Appendix B: Describing your sample

Why, how and when samples are collected can impact analyses of sequence data. In determining how a virus spreads, it is critical to track temporal and geographical information. It is also important to capture as much data provenance (who contributed it, where it came from, how it was generated) as possible. Different sampled materials or sampling processes may contain higher viral loads or produce better results, and differences in sampling protocols and practices should be accounted for (e.g. to understand sampling effects on the ability to identify specific pathogens).

A number of recommended and optional fields are provided to capture sampling approach ("purpose of sampling", "scope of sampling"), collection methods ("sample collection time duration value", "sample volume measurement value", "collection device", "collection method"), sample storage ("sample storage medium", "sample storage duration value"), and sewershed details ("environmental site", "environmental material", "wastewater system type"). **Populate only the fields that pertain to your sample**. Provide the most granular information allowable according to your organization's data sharing policies. ***Note: only include facts that are provided in the sample description, do not make assumptions and add details not provided by the data generator.***

e.g., a composite sample made of multiple grab samples should be recorded:

Specimen processing	Collection device	Collection method

Samples pooled	Grab sampler	Manual composite sampling

e.g., a 12-hr 1-liter flow-proportional composite sample should be recorded:

Sample collection time duration value	Sample collection time duration unit	Sample volume measurement value	Sample volume measurement unit	Collection device	Collection method
12	Hour	1	Liter (L)	Automatic flow-proportional sampler	Composite sampling

e.g., a treated sample collected from a WWTP that receives wastewater and run-off should be recorded:

Environmental material	Environmental material properties	Environmental site	Wastewater system type
Wastewater; Run-off (water)	Treated	Wastewater treatment plant	Combined sewer system

e.g., a sample collected from a contaminated lake should be recorded:

Environmental site	Environmental material	Environmental material properties
Lake	Surface water	Contaminated

e.g., a sample collected from a school latrine should be recorded:

Environmental material	Environmental site	Wastewater system type
Wastewater	School	Latrine

VI. Appendix C: New Field/Term Requests

Good data standards are continually maintained, and evolve over time as data needs change. In the event that the template you are using does not contain appropriate vocabulary for your needs, a new term request (NTR) can be made via two different mechanisms.Ideally, an NTR should be made via the PHA4GE Wastewater Specification GitHub IssueTracker. Users

will require a GitHub account, and should follow the instructions provided at XXX. NTR's will require the desired field/term name, suggested definition, definition source (citation or URL for website, article, textbook, or the name of the curator suggesting the definition etc). If possible, the user should suggest the target ontology and parent class of the term (but this is not required). Alternatively, NTRs can be made by emailing datastructures@pha4ge.org.

For more information and/or assistance, contact datastructures@pha4ge.org.

Revision History

Version	Date	Author	Description of Change
0.0	August 15, 2023	Jillian Paull & Emma Griffiths	Created protocol
0.1	December 12 2023	Emma Griffiths	Removed instructions for identifying standardized terms (using OLS)
0.2	January 17 2024	Emma Griffiths	Added new sections (Templates, Attribution, Metrics, added required fields), updated methods section, updated new term request instructions
0.3	February 09 2024	Charlie Barclay	Updated links to DH SOP.
1.0	March 22 2024	Emma Griffiths	Changed field requirements, expanded template descriptions, and added guidance on multi-tagging.