

PHA4GE Wastewater Contextual Data Specification Package - Testing Instructions

Created February 2024

Overview

PHA4GE Wastewater Contextual Data Specification Package consists of a set of structured terminology that is implemented in different contextual data collection templates designed to capture information scoped for three common wastewater pathogen genomics use cases (SARS-CoV-2 surveillance, AMR detection, identification and characterization of different pathogens found in wastewater). The templates are available in a data curation and validation tool called the DataHarmonizer, which can be downloaded and used locally. The templates are supported by field and term reference guides that provide definitions, guidance for populating templates, as well as examples of expected inputs. Additional ethical, privacy and practical guidance, as well as step-by-step instructions for getting started using the package, are provided in the accompanying curation standard operating procedure (SOP).

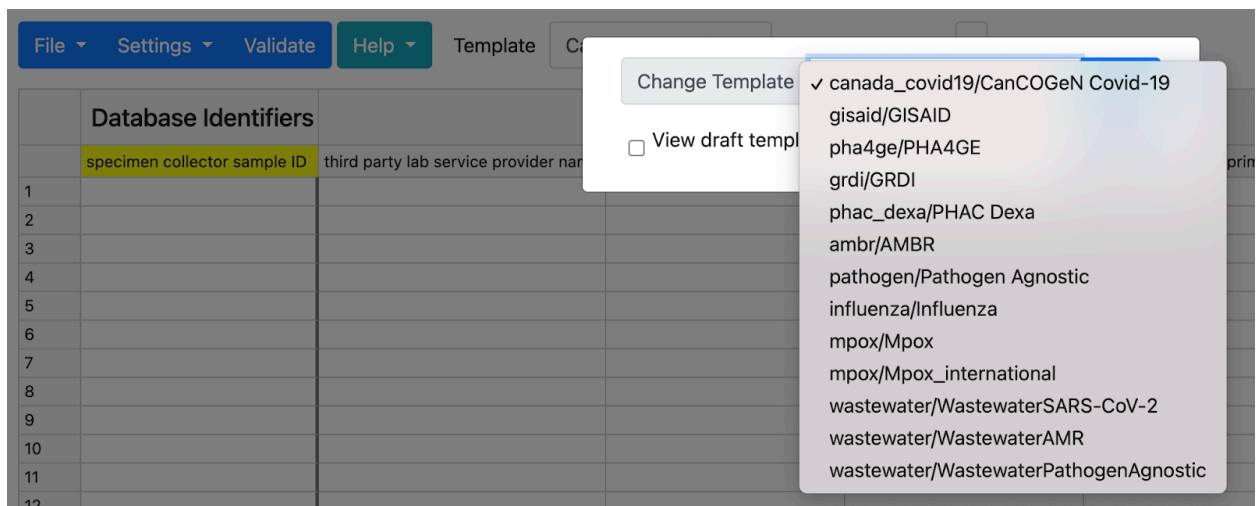
The package was created in collaboration with many public health scientists and wastewater experts from around the world, and is intended for data collection, harmonization, and future-proofing in private (organization-specific) databases, as well as data sharing between trusted partners (within networks), and public sharing. How much, or how little, of the specification is used and/or shared is completely at the discretion of the user, and organization-specific security and sharing policies should be observed when using the specification package.

The PHA4GE standards development workflow includes a testing phase in which specifications are “road-tested” in real-world situations to ensure their utility, applicability to different public health settings, and ease-of-use. The intention of testing is to work with the community to assess that terminology is sufficient and appropriate, that documentation provides the right types of instructions and guidance for users, and that tooling adds value to public health data management processes. These community-based assessments offer opportunities for dialogue and better understanding of public health needs, as well as co-creation, development of best practices, and ultimately transparent improvements of PHA4GE products.

While all feedback is important and appreciated, in this document, we provide guidance for testing in order to streamline the process and highlight areas we are particularly interested in improving. Additional curation training is available upon request. For more information or help, please contact datastructures@pha4ge.org. And if you are reading this as a preamble to testing this specification package - thank you!

Testing Instructions

1. Review the [curation SOP](#).
2. Review the [reference guides](#).
3. Review the [worked examples](#).
4. Review the [DataHarmonizer download and operation instructions](#).
5. Download the [DataHarmonizer](#).
 - a. Download the zip file.
 - b. Unzip and open the main.html file (the DataHarmonizer application will launch in your browser).
6. Choose a template to test from the Template dropdown menu.
(wastewater/WastewaterSars-CoV-2, wastewater/WastewaterAMR, wastewater/WastewaterPathogenAgnostic) e.g.



7. Populate the template using the DataHarmonizer instructions and picklists.
8. Validate your dataset and note any errors that are identified.
9. Note any issues experienced.
10. Try the different DataHarmonizer features (under “File” and “Settings”).
11. Review the instructions for how to request a new field or term ([New Term Request SOP](#)).

12. Complete the [feedback form](#) and submit to PHA4GE (datastructures@pha4ge.org).
(Alternatively, feedback can be provided to PHA4GE by email or directly to Data Structures Working Group members)

Considerations During Testing

Specification Content (Vocabulary)

- Suitability of fields and picklists
 - In general, were the fields and picklist values appropriate for your samples? If not, why not? How could this be improved?
Note: think about describing samples, isolates, diagnostic testing, sequencing methods, bioinformatics and analysis, etc.
 - Were any of the fields/terms that you would need missing? What were they?
 - How easy or difficult was it to locate the right vocabulary within picklists? Were there picklists that you found especially difficult to use? Why?
- Sections of Information:
 - Did the sections of information contain all fields/terms that you needed? If not, what additional sections/categories do you suggest? Which sections, if any, would you be least likely to use?

Suitability of templates

- How easy or difficult was it to fit your data in the template that you tested i.e. was it difficult to translate information in your databases/spreadsheets into the provided formats? If it was difficult, please explain why.
- How easy or difficult was it to navigate to the different sections of the template that you tested? If it was difficult, please explain why.
- Are your data needs covered by one of the templates? If not, how could we make them more fit-for-purpose (e.g. recombination of parts of existing templates)?
- Are there other wastewater pathogen genomics use cases for which we should consider developing a new template?
- Could you see yourself using (implementing) one of the templates for harmonizing and storing your contextual data routinely in your laboratory? If not, why not?
- It's possible that some mixing and matching of parts of different templates might be needed to suit your specific data needs. Would you like instructions for how to build your own customized template (either in the DataHarmonizer or in Excel)?

Ease of use of the DataHarmonizer

- How easy or difficult was it to download and use the DataHarmonizer application?
- Are there any additional instructions that you feel would enable others to use the DataHarmonizer more effectively?

Relevance of guidance documents

- How often did you refer to the reference guide? Was the reference guide clear? Did its definitions and examples make sense? Did the guidance help you populate the template?
- How often did you refer to the worked examples? Did the worked examples help you apply your data to the template that you tested? If not, why not? Are there any other worked examples that you would like included?
- How often did you refer to the curation SOP? Did the curation SOP provide appropriate guidance on the sensitivity of different types of data? What other guidance would you like (e.g. risks and benefits of sharing different types of data; how to map data elements to your system; how to populate bioinformatics and QC fields for different sequencing assays/types of analyses)?
- If you have not already, please read the New Term Request SOP. How easy or difficult do you think it would be to request additional vocabulary using the New Term Request system (NTR) on GitHub?

Training

- If you received curation training, how effective or non-effective was the training? How could training or training materials be improved?

Community Engagement

- How might PHA4GE engage the community better during specification development?

