Wastewater DataHarmonizer Download and Operation Instructions

I. **Purpose:** To harmonize contextual data across wastewater genomics data generators.
   a. Data providers will extract and curate lab-specific contextual data according to the steps outlined in the procedure below.
   b. Laboratories will populate the harmonized template with information from their datasets using the *DataHarmonizer* application.

II. **Data:** The contextual data describing identifiers and accession numbers, sample collection and processing, environmental conditions and measurements, strain and isolate information,, sequencing, bioinformatics and QC metrics, AMR profiling data, taxonomic identification information, lineage/clade information, and pathogen diagnostic testing information as supplied by the data provider.

III. **Procedure:**

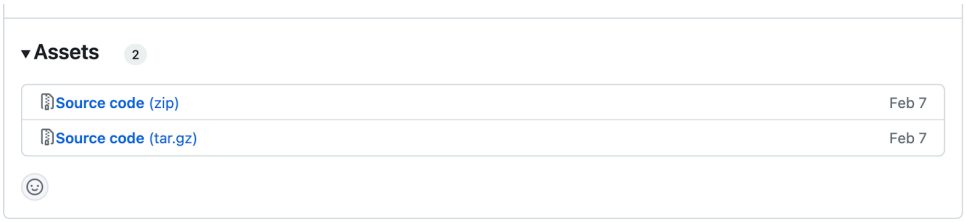
	Action	Related docs
1	Download the zip file ("Source code (zip))" containing The DataHarmonizer application from the following link: https://github.com/cidgoh/pathogen-genomics-package/releases	

Pathogen Genomics Package 5.0.1 Latest

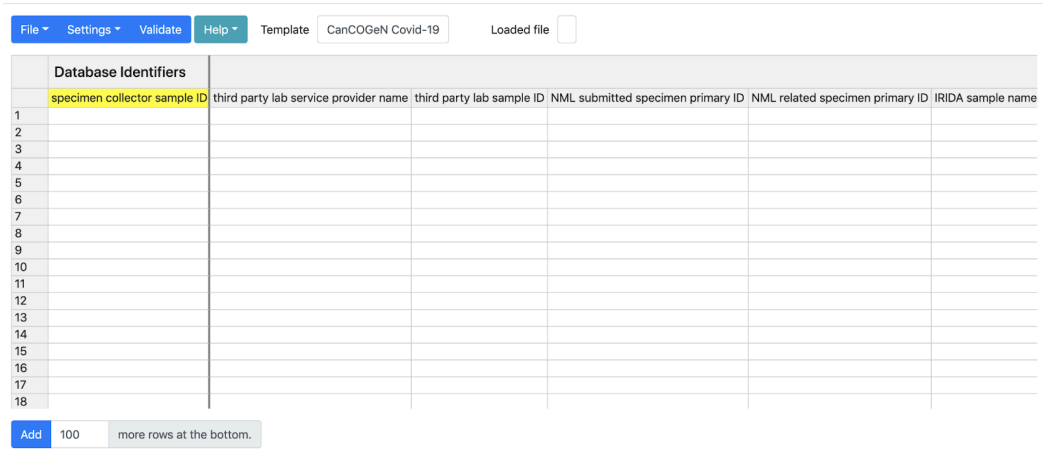
A fix to PGP5.0.0 involving restore of accidental missing CanCOGeN enumeration lists.

DataHarmonizer version 1.6.5 and CanCoGeN 2.3.4 and GRDI v9.0.0 and Wastewater 1.0.0 templates

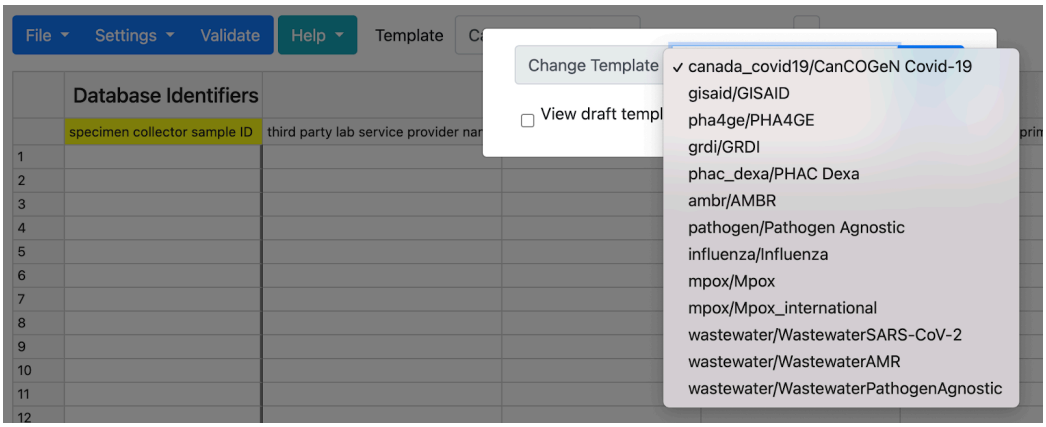
Template Name	Template Versionx.y.z	x changes (field)	y changes (values/IDs)	z changes (defs/formats/examples)
CanCOGeN (SC2)	2.3.4		new picklist IDs	
DEXA (One Health)	1.0.0			
GISAID (SC2)	1.0.0			
GRDI	9.0.0	new unit fields	added IDs under food_products and food_product_properties	minor edits (typos etc) to defs and guidance
Mpox	4.3.3			
Mpox-international	4.3.3			
PHA4GE (SC2)	1.0.1			
AMBR	2.3.0			
Pathogen_Agnostic	1.0.0			



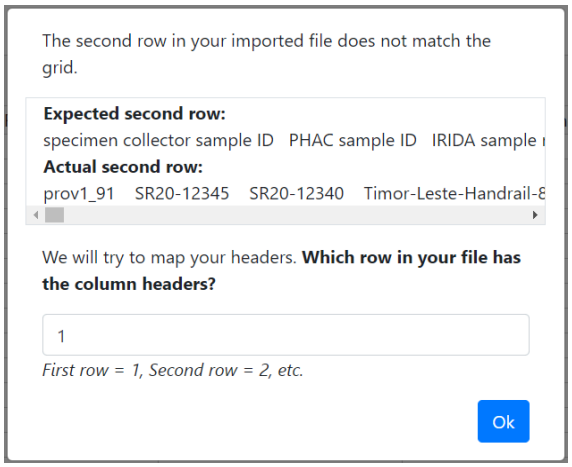
Extract the zip file's contents, and navigate into the extracted folder. Open **main.html**. The validator application will open in your default browser. It should look like this:



The DataHarmonizer enables contextual data harmonization for different pathogens and projects. Select one of the wastewater templates by clicking on the desired package (wastewater/WastewaterSars-CoV-2, wastewater/WastewaterAMR, wastewater/WastewaterPathogenAgnostic) from the **Template** menu beside the **Help** button.



Data can be entered into the validator application manually, by typing values into the application's spreadsheet, or data can be imported from local `xlsx`, `xls`, `tsv` and `csv` files.

	To import local data, click **File** on the top-left toolbar, and then click **Open**. To enter data in a new file, click **File** on the top-left toolbar, and then click **New**. Data entered into the spreadsheet can be copied and pasted. *Note: Only files containing the headers expected by the DataHarmonizer can be opened in the application.* *If you are missing the first row, you will get the following warning:*  *Resolve by declaring "1" as the row in which your column headers reside.*	
2	Before you begin to curate sample metadata: ● Review your dataset ● Review the fields in the template of the DataHarmonizer ● Review the field descriptions in the Reference Guide and curation SOP	
3	Familiarize yourself with DataHarmonizer functionality by reviewing the "**Getting Started**". To access "Getting Started", click on the green **Help** button on the top-left toolbar, then click **Getting Started**. Definitions, examples and further guidance are available by double clicking on the field headers, or by using the "**Reference Guide**". To access the "Reference Guide" click on the **Help** button, then click **Reference Guide**.	
4	Confirm mapping of your data fields to those in the harmonized template with the data steward (e.g. your supervisor). *If data is to be shared with trusted partners or public repositories, confirm the level of granularity of information that can be shared with the data steward. The most detailed information allowable should be included here.*	

5

Enter data into the validator spreadsheet.

- Hide non-required fields (colour-coded purple and white/grey) by clicking **Settings** on the top-left toolbar, followed by clicking on **Show Required Columns** (colour-coded in yellow).
- Double click in the field headers to see definitions and detailed guidance as needed (or consult Appendix A).
- Jump to a specific field header by clicking **Settings** on the top-left toolbar, followed by clicking on **Jump to**, then select the field header of the column you would like to view from the drop down list.
- Populate the validator template with the information from your dataset.
- Use picklists when provided.
- A value must be entered for every *required field* in each row. If data is missing or not collected, **choose a null value from the picklist**.
    - Not Applicable
    - Missing
    - Not Collected
    - Not Provided
    - Restricted Access
- Free text can be provided when picklists are not available.
- For filling an entire column with the same data, use the **Fill Column** function. Click **Settings**, followed by **Fill Column**. Type in the name of the desired field, followed by the value that should be used to fill every row in that column. Then click **OK**.

If a desired term is not present in a picklist, use the New Term Request System to request new vocabulary. Alternatively, contact Emma Griffiths at ega12@sfu.ca.

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.

Required fields are organized into subsections.

A data curation SOP is available with specific instructions for how to fill in fields (see the Wastewater Metadata Curation SOP). Below summarises the required (in bold) and/or recommended fields for each subsection (*Note: may vary by template*).

Subsection	Required/Recommended Fields
Database identifiers	**specimen collector sample ID** BioSample accession sampling site ID

Wastewater Metadata Curation SOP

	sampling event ID
Sample Collection and Processing *Note: Evaluate with your supervisor whether the specimen collector sample ID is considered identifiable by your institutional policies. If not, copy the sample ID into the sample ID field in the validator spreadsheet. If yes, provide the alternative sample ID as specified by the lab. Be sure to keep a copy of the key.*	**sample collected by** sample collector contact email **geo_loc (country)** **geo_loc (state/province/territory)** **organism** **purpose of sampling** scale of sampling **sample collection date** sample collection time sample collection time duration value sample collection time duration unit environmental site environmental material environmental material properties wastewater system type collection method endogenous control details
Environmental conditions and processing *Note: Populate only the fields that pertain to your sample. Provide the most granular information allowable according to your organization's data sharing policies. Select the appropriate value from the available pick list (consult the reference guide and curation SOP for further support).*	water catchment area human population measurement value precipitation measurement value precipitation measurement unit turbidity measurement value turbidity measurement unit fecal contamination indicator fecal contamination value fecal contamination unit
Strain and Isolate Information	microbiological method **isolate ID** alternative isolate ID serovar serotyping method
Sequence Information	**purpose of sequencing** sequenced by sequenced by contact name sequenced by contact email **sequence submitted by** sequence submitter contact email

	sequencing instrument sequencing assay type sequencing protocol amplicon pcr primer scheme genomic target enrichment method	
Bioinformatics and QC metrics	raw sequence data processing method dehosting method **consensus sequence software name** **consensus sequence software version** **sequence assembly software name** **sequence assembly software version**	
Taxonomic identification information	read mapping software name read mapping software version **read mapping software name** **read mapping software version** taxonomic reference database name taxonomic reference database version	
AMR detection information	AMR analysis software name AMR analysis software version AMR reference database name AMR reference database version AMR analysis report filename	

6 Validate the entered data by clicking on the **Validate** button on the top-left toolbar.

Missing information and invalid entries in required fields will be highlighted in red.

- Observe invalid rows by clicking **Settings** in the top-left toolbar, and then clicking on **Show invalid rows.**
- Address errors systematically by clicking the **Next Error** button. When all errors have been corrected, the **Next Error button** will disappear.
- Observe valid rows by clicking **Settings** in the top-left toolbar, and then clicking on **Show valid rows.**

	• Return view to all rows by clicking **Settings** in the top-left toolbar, and then clicking on **Show all rows.** *Note: Row viewing options only appear after a validation attempt has been made.*	
7	Address any invalid data that was flagged in red in the template. • Pale Red = Incorrect data format • Dark Red = Required data missing *Note: It is possible to export incomplete or invalid data. Make sure to review any errors prior to exporting.*	
8	Export validated data by clicking **File** on the top-left toolbar, and then clicking on **Save as**. Enter the file name and press **Save**.	

Appendix A: Document Revision History

Version	Date	Writer	Description of Change
1.0	Feb 09 2024	Emma Griffiths	Initial release
1.0	Feb 09 2024	Charlie Barclay	Updated links and required fields subsection
3.1	Oct 08 2024	Charlie Barclay	Version compatibility update