

**Wastewater
DataHarmonizer SOP 1.0**

Wastewater DataHarmonizer Download and Operation Instructions

- I. **Purpose:** To harmonize contextual data across data providers in the GRDI-AMR network.
- 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.
 - c. Data providers will upload harmonized data to IRIDA, and when permissible, the INSDC.
- II. **Data:** The contextual data describing identifiers and accession numbers, sample collection and processing, host information, sequencing, bioinformatics and QC metrics, AMR profiling data, risk assessment data, and public repository 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	

Wastewater DataHarmonizer SOP 1.0

	Action	Related docs
--	--------	--------------

2 weeks ago

ddooley

PGPv2.0.0

9e3909c

Compare

Pathogen Genomics Package 2.0.0

Includes NEW AMBR 1.0.0 template. The AMBR Project, led by the Harrison Lab at the University of Calgary, is an interdisciplinary study aimed at using 16S sequencing as part of a culturomics platform to identify antibiotic potentiators from the natural products of microbiota. The AMBR DataHarmonizer template was designed to standardize contextual data associated with the isolate repository from this work.

Includes DataHarmonizer v1.4.4

Template Name	Template Versionx.y.z	x changes (field)	y changes (values/IDs)	z changes (defs/formats/examples)
CanCOGeN (SC2)	1.0.1			
DEXA (One Health)	1.0.0			
GISAID (SC2)	1.0.0			
GRDI	5.2.1	new fields, AMR_measurement field removed	new IDs	des/formats/examples for new fields, and new examples for several existing fields
Monkeypox	3.3.2			
Monkeypox-international	3.3.2			
PHA4GE (SC2)	1.0.1			
AMBR	1.0.0			

▼ Assets

2

Source code (zip)

5 days ago

Source code (tar.gz)

5 days ago

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:

File

Settings

Validate

Help

Template

CanCOGeN Covid-19

Loaded file

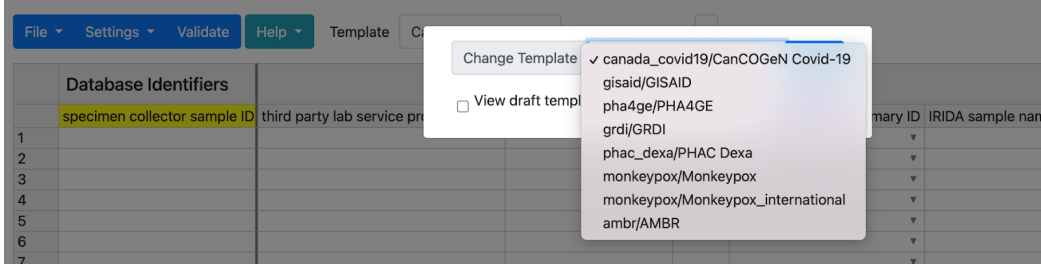
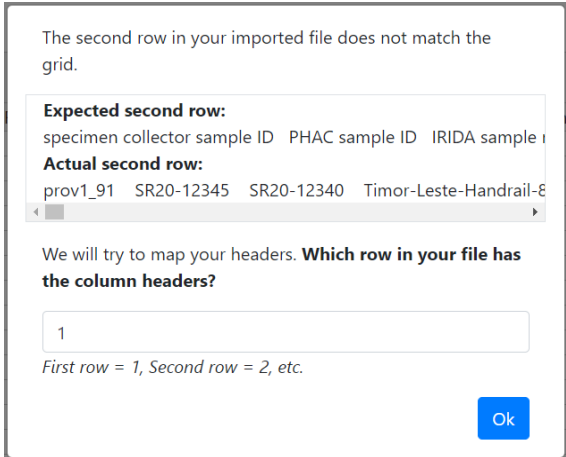
Database Identifiers						
	specimen collector sample ID	third party lab service provider name	third party lab sample ID	NML submitted specimen primary ID	NML related specimen primary ID	IRIDA sample name
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						

Add



100

more rows at the bottom.

Wastewater DataHarmonizer SOP 1.0

	Action	Related docs
	<p>The DataHarmonizer enables contextual data harmonization for different pathogens and projects. Select the AMBR template by selecting “grdi/GRDI” from the Template menu beside the Help button.</p>  <p>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.</p> <p>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.</p> <p><i>Note: Only files containing the headers expected by the DataHarmonizer can be opened in the application.</i></p> <p><i>If you are missing the first row, you will get the following warning:</i></p>  <p><i>Resolve by declaring “1” as the row in which your column headers reside.</i></p>	
2	Before you begin to curate sample metadata:	

Wastewater DataHarmonizer SOP 1.0

	Action	Related docs
	<ul style="list-style-type: none"> Review your dataset Review the fields in the template of the Validator application Review the field descriptions in the SOP Appendix 	
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	<p>Confirm mapping of your data fields to those in the harmonized template with the data steward (e.g. your supervisor).</p> <p><i>Confirm the level of granularity of information that can be shared in IRIDA with the data steward. The most detailed information allowable should be included here.</i></p>	
5	<p>Enter data into the validator spreadsheet.</p> <ul style="list-style-type: none"> 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).AMR-GRDI Metadata Curation SOP 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 <u>required field</u> in each row. If data is missing or not collected, choose a null value from the picklist. <ul style="list-style-type: none"> 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. 	Wastewater Metadata Curation SOP



**Wastewater
DataHarmonizer SOP 1.0**

	Action	Related docs						
	<p>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.</p> <p><i>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.</i></p> <p>Required fields are organized into subsections.</p> <p>Data should be entered into the DataHarmonizer in the same manner as other GRDI-AMR curation templates (i.e. the Excel version of the template). 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.</p> <table><tr><th>Subsection</th><th>Required/Recommended Fields</th></tr><tr><td>Database identifiers</td><td>specimen collector sample ID BioSample accession sampling site ID sampling event ID</td></tr><tr><td>Sample Collection and Processing <i>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.</i></td><td>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</td></tr></table>	Subsection	Required/Recommended Fields	Database identifiers	specimen collector sample ID BioSample accession sampling site ID sampling event ID	Sample Collection and Processing <i>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.</i>	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	
Subsection	Required/Recommended Fields							
Database identifiers	specimen collector sample ID BioSample accession sampling site ID sampling event ID							
Sample Collection and Processing <i>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.</i>	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							

**Wastewater
DataHarmonizer SOP 1.0**

	Action	Related docs
	Environmental conditions and processing <i>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).</i>	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	isolate_ID IRIDA_isolate_ID IRIDA_project_ID organism
	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

**Wastewater
DataHarmonizer SOP 1.0**

	Action	Related docs
6	<p>Validate the entered data by clicking on the Validate button on the top-left toolbar.</p> <p>Missing information and invalid entries in required fields will be highlighted in red.</p> <ul style="list-style-type: none"> • 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. <p><i>Note: Row viewing options only appear after a validation attempt has been made.</i></p>	
7	<p>Address any invalid data that was flagged in red in the template.</p> <ul style="list-style-type: none"> •  Pale Red = Incorrect data format •  Dark Red = Required data missing <p><i>Note: It is possible to export incomplete or invalid data. Make sure to review any errors prior to exporting.</i></p>	
8	<p>Export validated data by clicking File on the top-left toolbar, and then clicking on Save as. Enter the file name and press Save.</p>	
11	<p>Optional: Format validated data for IRIDA submission.</p> <p>You or your team members should have already created a project specific for the GRDI in IRIDA (irida.ca).</p> <p>Under File, select Save As. Make sure that the "Save as" type is "Excel Workbook" or .XLSX. Save the file with the name and location of your choice.</p> <p>Open the file and remove the top row containing the section headers. Re-save the file.</p> <p>Use the IRIDA Metadata Uploader to import your contextual data into the GRDI Project using the instructions provided here: https://phac-nml.github.io/irida-documentation/user/user/sample-metadata/</p>	<p>Upload to IRIDA SOP:</p> <p>https://irida.cerfacility.ca/documentation/user/samples/#adding-a-new-sample</p>

**Wastewater
DataHarmonizer SOP 1.0**

	Action	Related docs
	<i>Note: The IRIDA uploader only accepts Excel files, not csv files. If the top row containing the broad headings (Sample collection and processing, Host information, Sequencing, etc) is not removed, the IRIDA metadata upload will fail.</i>	

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

