Contextual Data (Metadata) Curation
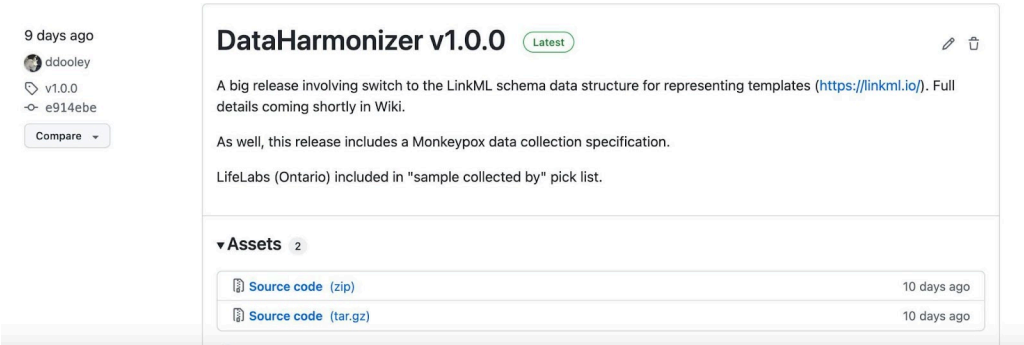
I.   **Purpose:** To harmonize highly pathogenic avian influenza (HPAI) contextual data across Canadian data providers.
   a. Data providers will extract and curate lab-specific contextual data according to the steps outlined in the procedure below. There are four scenario specific templates that can be utilised to capture contextual data for either food, wastewater, host or environmental samples. Please ensure you are using the correct template.
   b. Laboratories will populate the harmonized template with information from their datasets using the *DataHarmonizer* application.
   c. Data providers will share the harmonized data with the national database according to the agreed upon mechanism.

II.   **Data:** The contextual data describing sample collection and processing, host information, sequencing, and bioinformatics and QC metrics as supplied by the data provider.

III.   **Procedure:**

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

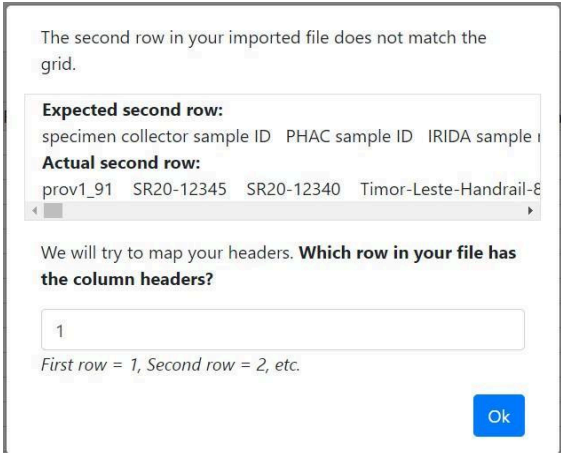|	**Action**	**Related docs**
1 cont.	The DataHarmonizer enables contextual data harmonization for different pathogens and projects. By default, the CanCOGeN template will load. To navigate to the HPAI template, select the **Template** menu beside the **Help** button. A dropdown selection will appear. Select the appropriate template for your data from the Template menu. All HPAI templates will begin with the prefix **HPAI/** 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.*	

	Action	Related docs
	- 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. *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 - Ensure you have chosen the appropriate template for your HPAI contextual data. There are four templates to choose from, which contain tailored fields for each data context. They are: **Environment** **Food** **Host** **Wastewater** - Review the fields in the template of the DataHarmonizer application or review the fields in the reference guide.	

Within the table cell, the warning dialog reads:

> The second row in your imported file does not match the grid.
>
> **Expected second row:**
> specimen collector sample ID   PHAC sample ID   IRIDA sample
> **Actual second row:**
> prov1_91   SR20-12345   SR20-12340   Timor-Leste-Handrail-8
>
> We will try to map your headers. **Which row in your file has the column headers?**
>
> 1
>
> *First row = 1, Second row = 2, etc.*
>
> Ok

	Action	Related docs
	○ To review the required and recommended fields in the application go to **Settings** and under the dropdown select **Show required and recommended columns**. ○ Alternatively review the required (field names are highlighted yellow ▉) and recommended fields (field names are highlighted purple ▉) in the reference guide. ● Review the field descriptions in the SOP Appendix of this document, or review the Definitions column in the reference guide.	
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). *Note: Confirm the level of granularity of information that can be shared publicly vs in the national database, with the data steward and/or the privacy officer. The most detailed information allowable should be included here.*	
5	**Data Entry:** 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). ● 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**.	

	Action	Related docs

Subsection	Required Fields
Database Identifiers	specimen collector sample 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_collection_date sample_collection_date_precision geo_loc (country) geo_loc (province/territory) organism influenza_subtype purpose_of_sampling purpose_of_sampling_details
Describing the material and/or site sampled. *Note: Fields have been introduced to capture different kinds of anatomical and environmental samples, as well as collection methods. Populate only the fields that pertain to your sample - provide null values for the fields that are not applicable. Provide the most granular information allowable according to your organization's data sharing policies.*	food_product* ***for food template only**
Host Information	host (scientific name)** host disease** host_age** host_age_unit** ****for host template only**
Sequencing	purpose_of_sequencing sequenced_by sequenced_by_contact_name sequenced_by_contact_email

	Action	Related docs
6	**Requesting a new term:** **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 (see **Appendix A** for required field definitions and guidance, and **Appendix B** for examples of how to structure sample descriptions):	
7	**Validate your data:** Check 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.*	
8	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.*	

	Action	Related docs
9	**Export validated data:** Clicking **File** on the top-left toolbar, and then clicking on **Save as**. Enter the file name and press **Save**. Export to GISAID, Biosample (NCBI), VirusSeq Data Portal, or NML-LIMS formats by clicking **File** on the top-left toolbar, and then clicking **Export to**. Have the validated data reviewed by the data steward (i.e. your supervisor).	
10	**Additional Information:** A local copy of the **Standard Operating Procedure (SOP)** is included in every download of the DataHarmonizer. To access it, click on the green **Help** button on the top-left toolbar, then click **SOP**. The latest version of the SOP is published online and accessible via a web browser at all times. Datasets that can be used for testing, training, and quality control purposes are also available.	

IV.   **Appendix A: Required Field Definitions and Guidance**

Field definitions for required fields, as well as guidance and examples, are provided below. This information has been sourced from the DataHarmonizer reference guide. Guidance for strongly recommended and optional fields can be found in the reference guide. For access to information on non-required fields, refer to "Procedure - Action 3"**.**

Database Identifiers

specimen collector sample ID

The user-defined name for the sample.
Store the collector sample ID. If this number is considered identifiable information, provide an alternative ID. Make sure to store the key between this alternative ID and the original ID for traceability. Every collector sample ID from a single submitter must be unique. It can have any format, but we suggest that you make it concise, unique and consistent within your lab.
e.g. prov_rona_99

Sample Collection and Processing

sample collected by

The name of the agency that collected the original sample.
The name of the sample collector should be written out in full, (with minor exceptions) and be consistent across multiple submissions e.g. Public Health Agency of Canada, Public Health Ontario, BC Centre for Disease Control. The sample collector specified is at the discretion of the data provider (i.e. may be hospital, provincial public health lab, or other).
e.g. BC Centre for Disease Control

sample collection date

The date on which the sample was collected.
Sample collection date is critical for surveillance and many types of analyses. Required granularity includes year, month and day. Record the collection date accurately in the template. Before sharing this data, ensure you have consulted the data steward and/or your privacy officer regarding whether they consider this date to be identifiable information. If this date is considered

identifiable, it is acceptable to add "jitter" to the collection date you share by adding or subtracting a calendar day (acceptable by GISAID). Do not change the collection date in your original records. Alternatively, "received date" may be used as a substitute in the data you share. The date should be provided in ISO 8601 standard format "YYYY-MM-DD".
e.g. 2020-03-16

sample collection date precision

The precision to which the "sample collection date" was provided.
Provide the precision of granularity to the "day", "month", or "year" for the date provided in the "sample collection date" field. The "sample collection date" will be truncated to the precision specified upon export; "day" for "YYYY-MM-DD", "month" for "YYYY-MM", or "year" for "YYYY".
e.g. year

geo_loc_name (country)

The country where the sample was collected.
Provide the country name from the controlled vocabulary provided.
e.g. Canada

geo_loc_name (province/territory)

The province/territory where the sample was collected.
Provide the province/territory name from the controlled vocabulary provided.
e.g. Saskatchewan

organism

Taxonomic name of the organism.
Provide the official nomenclature for the organism(s) present in the sample. Multiple organisms can be entered, separated by semicolons. Avoid abbreviations. Search for taxonomic names here: ncbi.nlm.nih.gov/taxonomy.
e.g. Influenza A virus

influenza subtype

The taxonomic name for the specific subtype.
Provide the appropriate influenza subtype.
e.g. H5N1 subtype (Influenza A virus) [NCBITaxon:102793]

purpose of sampling

The reason that the sample was collected.
As all samples are taken for diagnostic purposes, "Diagnostic Testing" should be chosen from the picklist at this time. The reason why a sample was originally collected may differ from the reason why it was selected for sequencing, which should be indicated in the "purpose of sequencing" field.
e.g. Diagnostic testing

Describing the material and/or site sampled.

food product

A material consumed and digested for nutritional value or enjoyment.
This field includes animal feed. If applicable, select the standardized term and ontology ID for the anatomical material from the picklist provided. Multiple values can be provided, separated by a semi-colon.
e.g. Bone meal [ENVO:02000054]

Host Information

host (scientific name)

The taxonomic, or scientific name of the host.
Common name or scientific name are required if there was a host. Both can be provided, if known. Use terms from the pick lists in the template. Scientific name e.g. Homo sapiens, If the sample was environmental, put "Not Applicable".
e.g. Homo sapiens

host disease

The name of the disease experienced by the host.
Select "Influenza A" from the pick list provided in the template.
e.g. Influenza A

host age

The age of host at the time of sampling.
If known, provide the numerical age, otherwise age-binning complete the 'host age bin' field.
e.g. 32

host age unit

The units used to measure the host's age.
If known, provide the age units used to measure the host's age from the pick list.
e.g. year

Sequencing

sequenced by

The name of the agency that generated the sequence.
The name of the agency should be written out in full, (with minor exceptions) and be consistent across multiple submissions. If submitting specimens rather than sequencing data, please put the "National Microbiology Laboratory (NML)".
e.g. Public Health Ontario (PHO)

sequence by contact name

The name or title of the contact responsible for follow-up regarding the sequence.
Provide the name of an individual or their job title. As personnel turnover may render the contact's

name obsolete, it is more prefereable to provide a job title for ensuring accuracy of information and institutional memory. If the information is unknown or cannot be provided, leave blank or provide a null value.
 e.g. Enterics Lab Manager

sequencing date

The date the sample was sequenced.
Provide the date that the sample was sequenced in ISO 8601 standard "YYYY-MM-DD" format.
If the exact sequencing date is unknown, proxy dates may be used instead (e.g. library preparation date)
e.g. 2020-06-22

purpose of sequencing

The reason that the isolate was sequenced.
Select "Targeted surveillance (non-random sampling)" if the specimen fits any of the following criteria:
- Specimens attributed to individuals with no known intimate contacts to positive cases.
- Specimens attributed to youth/minors <18 yrs.
- Specimens attributed to vulnerable persons living in transient shelters or congregant settings.
- Specimens attributed to individuals self-identifying as "female".

For specimens with a recent international and/or domestic travel history, please select the most appropriate tag from the following three options:
- "Domestic travel surveillance"
- "International travel surveillance"
- "Travel-associated surveillance"

For specimens targeted for sequencing as part of an outbreak investigation, please select "Cluster/Outbreak investigation".

"Baseline surveillance (random sampling)" should be used in all other cases.

V.   **Appendix B: Structuring Sample Descriptions (Examples)**

Several examples are provided below which illustrate how to structure common sample descriptions.

e.g. Nasal swab from the snout of a cow as part of outbreak should be recorded:

host (scientific name)	host disease	anatomical part	body product	collection device	purpose of sampling
Bos taurus [NCBITaxon:9913]	Highly pathogenic avian influenza A (HPAI)	Snout [UBERON:0006333]	Mucus [UBERON:0000912]	Swab [GENEPIO:0100027]	Cluster/Outbreak investigation [GENEPIO:0100019]

e.g. Wild bird feces should be recorded:

host (common name)	body product
Bird [NCBITaxon:8782]	Feces [UBERON:0001988]

e.g. Deer carcus should be recorded:

host (common name)	anatomical material	anatomical part
Deer [NCBITaxon:9850]	Tissue [UBERON:0000479]	Carcass [UBERON:0008979]

e.g. Milk testing (organic from farm) should be recorded:

food product	food product properties	environmental site
Milk [UBERON:0001913]	Organic food claim or use [FOODON:03510128]	Dairy farm [ENVO:03501416]

e.g. Wastewater from a treatment plant should be recorded:

environmental site	environmental material
Wastewater treatment plant [ENVO:00002272]	Wastewater [ENVO:00002001]

VI.   **Appendix C: Null Value Definitions**

Not Applicable

Information is inappropriate to report, can indicate that the standard itself fails to model or represent the information appropriately.

Missing

Information was known to be recorded in the past, but the observed value cannot be located or retrieved for some reason.

Not Collected

Information of an expected format was not given because it has not been collected.

Not Provided

Information of an expected format was not given, a value may be given at the later stage.

Restricted Access

Information exists but can not be released openly because of privacy concerns.

Source:
International Nucleotide Sequence Database Collaboration (INSDC) Missing Value Reporting Terms (2017-2018).
ENA Training Modules: https://ena-docs.readthedocs.io/en/latest/submit/samples/missing-values.html

Revision History

Version	Date	Writer	Description of Change
1.0	June 10 2022	Emma Griffiths	Created protocol
2.0	June 24 2022	Emma Griffiths	Updated field names, images in the instructions
3.0	June 30 2022	Emma Griffiths	Updated field names, instructions for sequence upload
5.4	September 2024	Charlie Barclay	Updated formatting, updated images for DH download and links. Added required fields in line with most up to date github version. Updated versioning to align with github and DH.