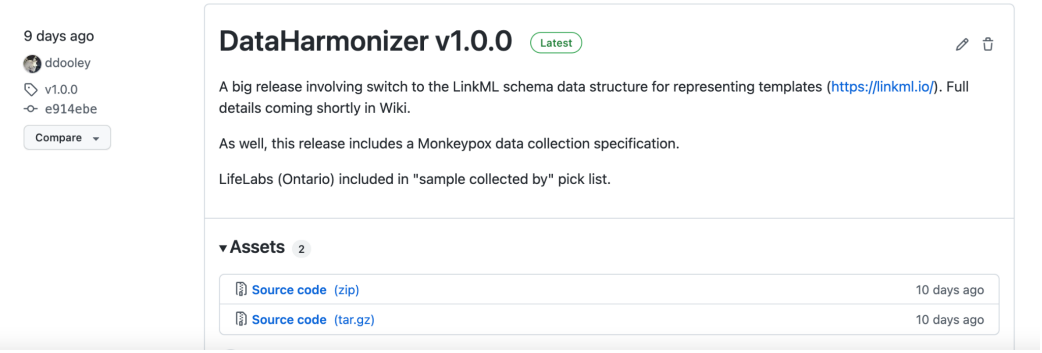
Contextual Data (Metadata) Curation

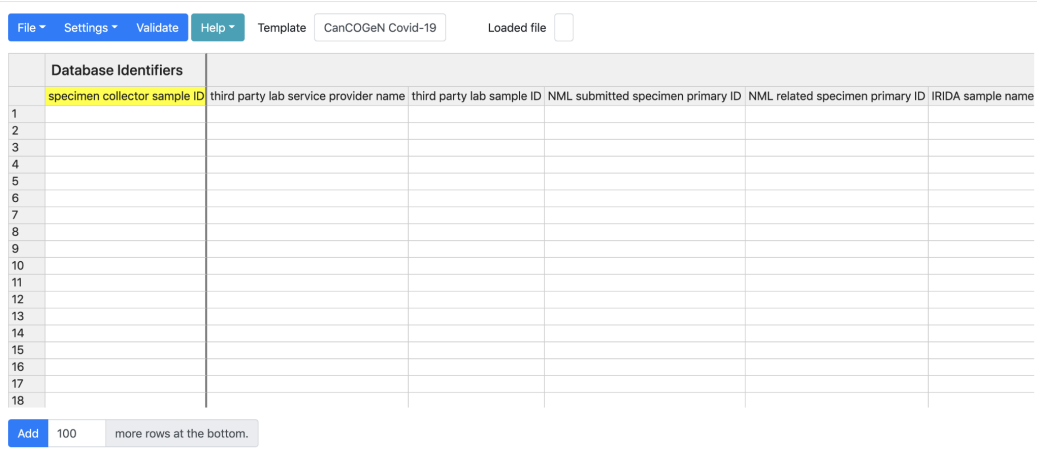
I. **Purpose:** To harmonize Monkeypox contextual data across data providers worldwide.
   a. Data providers will extract and curate lab-specific contextual data according to the steps outlined in the procedure below.
   b. Laboratories will populate the harmonized template with information from their datasets using the *DataHarmonizer* application.
   c. Data providers will share harmonized data according to their organization's data sharing policies and with the approval of the authorized data steward.

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

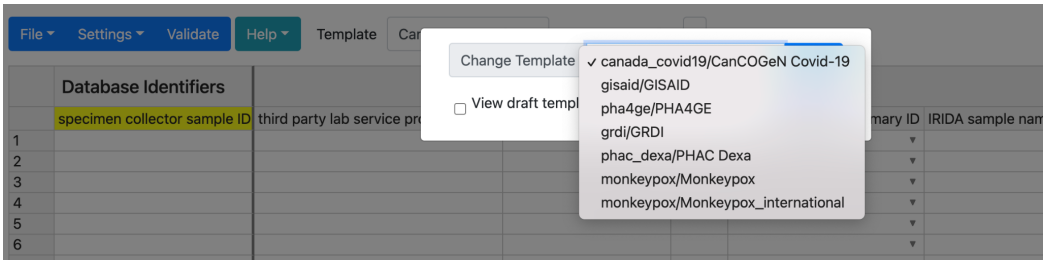
III. **Procedure:**

	Action	Related docs
1	Download the zip file ("Source code (zip))" containing The DataHarmonizer application from the following link: https://github.com/Public-Health-Bioinformatics/covid19ValidationGrid/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



The DataHarmonizer enables contextual data harmonization for different pathogens and projects. Select the Monkeypox template by selecting the “monkeypox/Monkeypox_international template” from the **Template** menu beside the **Help** button.

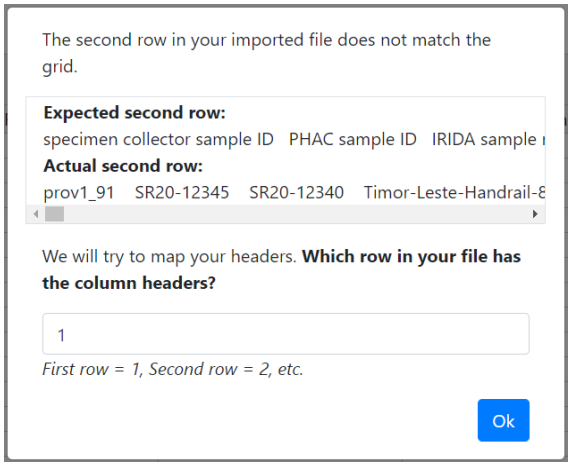


Data can be entered into the validator application manually, by typing values into the application’s spreadsheet, or data can be imported from local `xlsx`, `xls`, `tsv` and `csv` files.

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. Example:

If you are missing the first row, you will get the following warning:

	Action	Related docs
	 *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 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	Confirm mapping of your data fields to those in the harmonized template with the data steward (e.g. your supervisor). *Note: A version of this information will be made public in GISAID and NCBI, however, another version of this data will be captured in the access controlled national database. 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	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).	

	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.
- 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, 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):

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.*	sample collected by sequenced by sequence submitted by sample collection date geo_loc (country) geo_loc (province/territory) organism isolate

	Action	Related docs

If yes, provide the alternative sample ID as specified by the lab. Be sure to keep a copy of the key.	purpose of sampling purpose of sampling details
Host Information	host (scientific name) host disease
Sequencing	sequencing instrument purpose of sequencing purpose of sequencing details
Bioinformatics and QC Metrics	consensus software name consensus software version

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**. Export to GISAID and Biosample (NCBI) by clicking **File** on the top-left toolbar, and then clicking **Export to**.	

	Action	Related docs
	● Have the validated data reviewed by the data steward (i.e. your supervisor)	
9	Optional: Format validated data for GISAID submission. The DataHarmonizer will automate the preparation of a GISAID submission form from the entered data by exporting the data in GISAID format. ● Export your data in “GISAID” format by clicking **File** on the top-left toolbar, then clicking **Export To**. Type in the file name, and select “**GISAID**” from the Format picklist. Then click **Export**.	
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.

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_mpx_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

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 submitted by

The name of the agency that submitted the sequence to a database.

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)

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

geo_loc_name (country)
The country where the sample was collected.
Provide the country name from the controlled vocabulary provided.
e.g. United States of America [GAZ:00002459]

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.
Use "Monkeypox virus". This value is provided in the template.
e.g. Monkeypox virus [NCBITaxon:10244]

isolate
Identifier of the specific isolate.
This identifier should be an unique, indexed, alpha-numeric ID within your laboratory. If submitted to the INSDC, the "isolate" name is propagated throughout different databases. As such, structure the "isolate" name to be INSDC compliant in the following format: "MpxV/host/country/sampleID/date".
e.g. MpxV/human/USA/CA-CDPH-001/2020

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 [GENEPIO:0100002]

purpose of sampling details
The description of why the sample was collected providing specific details.
Provide an expanded description of why the sample was collected using free text. The description may include the importance of the sample for a particular public health investigation/surveillance activity/research question. If details are not available, provide a null value.
e.g. Symptomology and history suggested Monkeypox diagnosis.

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 [NCBITaxon:9606]

host disease
The name of the disease experienced by the host.
Select "Monkeypox" from the pick list provided in the template.
e.g. Monkeypox [MONDO:0002594]

Sequencing

sequencing instrument
The model of the sequencing instrument used.
Select a sequencing instrument from the picklist provided in the template.
e.g. Oxford Nanopore MinION [GENEPIO:0100142]

purpose of sequencing
The reason that the isolate was sequenced.
The reason why a sample was originally collected may differ from the reason why it was selected for sequencing. The reason a sample was sequenced may provide information about potential biases in sequencing strategy. Provide the purpose of sequencing from the picklist in the template. The reason for sample collection should be indicated in the "purpose of sampling" field.
e.g. Baseline surveillance (random sampling) [GENEPIO:0100005]

purpose of sequencing details
The description of why the sample was sequenced providing specific details.
Provide an expanded description of why the sample was sequenced using free text. The description may include the importance of the sequences for a particular public health investigation/surveillance activity/research question. If details are not available, provide a null value.
e.g. Outbreak in MSM community

Bioinformatics and QC Metrics

consensus software name
The name of software used to generate the consensus sequence.
Provide the name of the software used to generate the consensus sequence.
e.g. iVar

consensus software version
The version of the software used to generate the consensus sequence.
Provide the version of the software used to generate the consensus sequence.
e.g. 1.3

V. Appendix B: Structuring Sample Descriptions (Examples)

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

e.g. Swab, pustule lesion from the groin should be recorded:

host (scientific name)	host disease	anatomical material	anatomical part	collection device
Homo sapiens [NCBITaxon:9606]	Monkeypox [MONDO:0002594]	Lesion (Pustule) [NCIT:C78582]	Genital area [BTO:0003358]	Swab [GENEPIO:0100027]

e.g. tissue biopsy should be recorded:

host (scientific name)	host disease	anatomical material	collection method
Homo sapiens [NCBITaxon:9606]	Monkeypox [MONDO:0002594]	Tissue [UBERON:0000479]	Biopsy [OBI:0002650]

e.g. saliva should be recorded:

host (scientific name)	host disease	anatomical material
Homo sapiens [NCBITaxon:9606]	Monkeypox [MONDO:0002594]	Saliva [UBERON:0001836]

e.g. gargarism should be recorded:

host (scientific name)	host disease	collection method
Homo sapiens [NCBITaxon:9606]	Monkeypox [MONDO:0002594]	Saline gargle (mouth rinse and gargle) [GENEPIO:0100034]

e.g. Pustule Fluid R upper chest should be recorded:

host (scientific name)	host disease	anatomical material	anatomical part
Homo sapiens [NCBITaxon:9606]	Monkeypox [MONDO:0002594]	Lesion (Pustule) [NCIT:C78582]; Fluid [UBERON:0006314]	Chest [UBERON:0001443]

e.g. throat swab should be recorded:

host (scientific name)	host disease	anatomical part	collection device
Homo sapiens [NCBITaxon:9606]	Monkeypox [MONDO:0002594]	Oropharynx (OP) [UBERON:0001729]	Swab [GENEPIO:0100027]

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 24 2022	Emma Griffiths	Created protocol