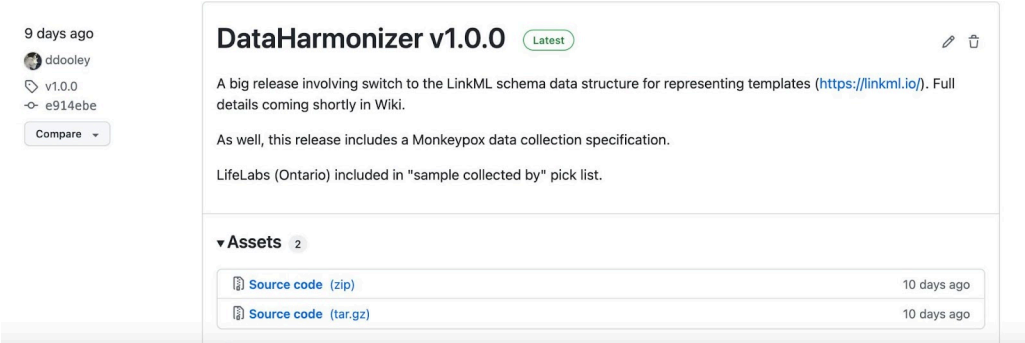


Contextual Data (Metadata) Curation

- I. **Purpose:** To harmonize MPox contextual data across Canadian data providers.
- Data providers will extract and curate lab-specific contextual data according to the steps outlined in the procedure below.
 - Laboratories will populate the harmonized template with information from their datasets using the *DataHarmonizer* application.
 - 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	<p>Download the zip file ("Source code (zip)") containing The DataHarmonizer application from the following link: https://github.com/cidgoh/pathogen-genomics-package/releases</p>  <p>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:</p>	

National Microbiology Laboratory - MPox
MPox Contextual Data Curation SOP 6.5

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

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

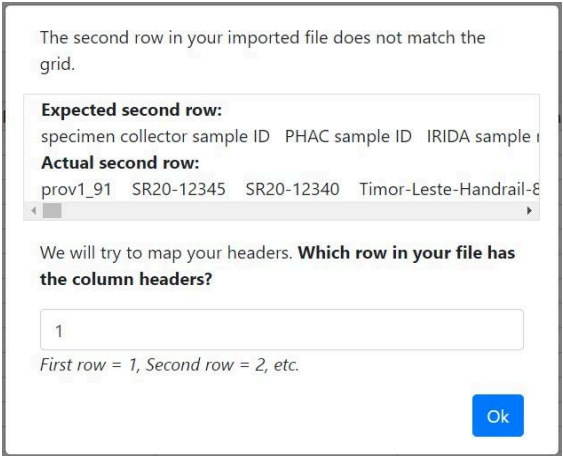
Data can be entered into the validator application manually, by typing values into the application's spreadsheet, or data can be imported from local **xlsx**, **xls**, **tsv** and **csv** files.

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:

National Microbiology Laboratory - MPox
MPox Contextual Data Curation SOP 6.5

	Action	Related docs
	 <p style="text-align: center;"><i>Resolve by declaring “1” as the row in which your column headers reside.</i></p>	
2	<p>Before you begin to curate sample metadata:</p> <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	<p>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.</p>	
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>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.</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). 	

National Microbiology Laboratory - MPox
MPox Contextual Data Curation SOP 6.5

	Action	Related docs
	<ul style="list-style-type: none"> • 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 <i>required field</i> 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. <p>Requesting a new term:</p> <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 (see Appendix A for required field definitions and guidance, and Appendix B for examples of how to structure sample descriptions):</p>	

National Microbiology Laboratory - MPox
MPox Contextual Data Curation SOP 6.5



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

Subsection	Required Fields
Database Identifiers	specimen collector sample 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.</i> <i>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 collection date sample collection date precision geo_loc (country) geo_loc (province/territory) organism isolate purpose of sampling purpose of sampling details
Describing the material and/or site sampled. <i>Note: Seven 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. NML submitted specimen type is required for upload to CNPHI. Select the appropriate value from the available pick list (consult the reference guide for further support).</i>	NML submitted specimen type anatomical material anatomical part body product collection device collection method

National Microbiology Laboratory - MPox
MPox Contextual Data Curation SOP 6.5

	<table><tr><td>Host Information</td><td>host (scientific name) host disease host age host age unit host age bin host gender</td></tr><tr><td>Sequencing</td><td>sequenced by sequence submitted by purpose of sequencing purpose of sequencing details sequencing instrument sequencing date</td></tr><tr><td>Bioinformatics and QC Metrics</td><td>raw sequencing data processing method hosting method consensus software name consensus software version sequence assembly software name sequence assembly software version bioinformatics protocol</td></tr></table>	Host Information	host (scientific name) host disease host age host age unit host age bin host gender	Sequencing	sequenced by sequence submitted by purpose of sequencing purpose of sequencing details sequencing instrument sequencing date	Bioinformatics and QC Metrics	raw sequencing data processing method hosting method consensus software name consensus software version sequence assembly software name sequence assembly software version bioinformatics protocol	
Host Information	host (scientific name) host disease host age host age unit host age bin host gender							
Sequencing	sequenced by sequence submitted by purpose of sequencing purpose of sequencing details sequencing instrument sequencing date							
Bioinformatics and QC Metrics	raw sequencing data processing method hosting method consensus software name consensus software version sequence assembly software name sequence assembly software version bioinformatics protocol							
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>							

National Microbiology Laboratory - MPox
MPox Contextual Data Curation SOP 6.5

	Action	Related docs
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. 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.</p> <ul style="list-style-type: none"> • Have the validated data reviewed by the data steward (i.e. your supervisor) 	
9	<p>Submit validated data to the national database.</p> <p>You can submit either by i) emailing the validated data to your NML contact, or ii) uploading the validated data directly through the NML's NextCloud Sharepoint</p> <ul style="list-style-type: none"> • Before uploading to the NextCloud Sharepoint, export your data in "NML-LIMS" format by clicking File on the top-left toolbar, then clicking Export To. Type in the file name, and select "NML-LIMS" from the Format picklist. Then click Export. • See the NML's NextCloud Sharepoint documentation for more information regarding Metadata Upload. 	
10	<p>Optional: Format validated data for GISAID submission.</p> <p>The DataHarmonizer will automate the preparation of a GISAID submission form from the entered data by exporting the data in GISAID format.</p>	

National Microbiology Laboratory - MPox
MPox Contextual Data Curation SOP 6.5

	Action	Related docs
	<ul style="list-style-type: none">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.	
11	<p>Additional Information:</p> <p>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.</p> <p>The latest version of the SOP is published online and accessible via a web browser at all times.</p> <p>Datasets that can be used for testing, training, and quality control purposes are also available.</p>	

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

National Microbiology Laboratory - MPox
MPox Contextual Data Curation SOP 6.5

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.

Use "MPox virus". This value is provided in the template.
e.g. MPox virus

isolate

Identifier of the specific isolate.

Provide the GISAID EpiPox virus name, which should be written in the format "hMpxV/CANADA/2 digit provincial ISO code-xxxxx/year". If the province code cannot be shared for privacy reasons, put "UN" for "Unknown".
e.g. hMpxV/Canada/UN-NML-12345/2022

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

purpose of sampling details

The description of why the sample was collected providing specific details.

National Microbiology Laboratory - MPox
MPox Contextual Data Curation SOP 6.5

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 MPox diagnosis.

Describing the material and/or site sampled.

NML submitted specimen type

The type of specimen submitted to the NML for testing.

This information is required for upload through the CNPHI LaSER system. Select the specimen type from the pick list provided. If sequence data is being submitted rather than a specimen for testing, select "Not Applicable".

e.g. Swab

anatomical material

A substance obtained from an anatomical part of an organism e.g. tissue, blood.

Provide a descriptor if an anatomical material was sampled. Use the picklist provided in the template. If a desired term is missing from the picklist, contact emma_griffiths@sfu.ca. If not applicable, do not leave blank. Choose a null value.

e.g. Lesion (Pustule)

anatomical part

An anatomical part/location of an organism e.g. oropharynx.

Provide a descriptor if an anatomical part was sampled. Use the picklist provided in the template. If a desired term is missing from the picklist, contact emma_griffiths@sfu.ca. If not applicable, do not leave blank. Choose a null value.

e.g. Genital area

body product

A substance excreted/secreted from an organism e.g. feces, urine, sweat.

Provide a descriptor if a body product was sampled. Use the picklist provided in the template. If a desired term is missing from the picklist, contact emma_griffiths@sfu.ca. If not applicable, do not leave blank. Choose a null value.

e.g. Feces

collection device

The instrument or container used to collect the sample e.g. swab.

Provide a descriptor if a device was used for sampling. Use the picklist provided in the template. If a desired term is missing from the picklist, contact emma.griffiths@bccdc.ca. If not applicable, do not leave blank. Choose a null value.

e.g. Swab

National Microbiology Laboratory - MPox
MPox Contextual Data Curation SOP 6.5

collection method

The process used to collect the sample e.g. phlebotomy, necropsy.

Provide a descriptor if a collection method was used for sampling. Use the picklist provided in the template. If a desired term is missing from the picklist, contact emma.griffiths@bccdc.ca. If not applicable, do not leave blank. Choose a null value.

e.g. Biopsy

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 "MPox" from the pick list provided in the template.

e.g. MPox

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

host age bin

The age category of the host at the time of sampling.

The age bins, in 10 year intervals have been provided. If a host's age cannot be specified due to privacy concerns, an age bin can be used as an alternative.

e.g. 50 - 59

host gender

The gender of the host at the time of sample collection.

If known, select a value from the pick list.

e.g. male

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 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)

sequencing instrument

The model of the sequencing instrument used.

Select a sequencing instrument from the picklist provided in the template.

e.g. Minlon

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.

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

raw sequencing data processing method

The names of the software and version number used for raw data processing such as removing barcodes, adapter trimming, filtering etc.

Provide the software name followed by the version.

e.g. Trimmomatic v. 0.38, Porechop v. 0.2.3

dehosting method

The method used to remove host reads from the pathogen sequence.

Provide the name and version number of the software used to remove host reads.

e.g. BWA 0.7.17

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

sequence assembly software name

The name of the software used to assemble a sequence.

Provide the name of the software used to assemble the sequence.

e.g. SPAdes Genome Assembler, Canu, wtdbg2, velvet

sequence assembly software version

The version of the software used to assemble a sequence.

Provide the version of the software used to assemble the sequence.

e.g. 3.15.5

bioinformatics protocol

A description of the overall bioinformatics strategy used.

Further details regarding the methods used to process raw data, and/or generate assemblies, and/or generate consensus sequences can. This information can be provided in an SOP or protocol or pipeline/workflow. Provide the name and version number of the protocol, or a GitHub link to a pipeline or workflow.

National Microbiology Laboratory - MPox
MPox Contextual Data Curation SOP 6.5

e.g. <https://github.com/phac-nml/ncov2019-artic-nf>

National Microbiology Laboratory - MPox
MPox Contextual Data Curation SOP 6.5

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	MPox	Lesion (Pustule)	Genital area	Swab

e.g. anal dry swab should be recorded:

host (scientific name)	host disease	anatomical part	collection device
Homo sapiens	MPox	Anus	Dry swab

e.g. saliva should be recorded:

host (scientific name)	host disease	anatomical material
Homo sapiens	MPox	Saliva

e.g. gargarism should be recorded:

host (scientific name)	host disease	collection method
Homo sapiens	MPox	Saline gargle (mouth rinse and gargle)

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

host (scientific name)	host disease	anatomical material	anatomical part
Homo sapiens	MPox	Lesion (Pustule); Fluid	Chest

e.g. throat swab should be recorded:

host (scientific name)	host disease	anatomical part	collection device
Homo sapiens	MPox	Oropharynx (OP)	Swab

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>

National Microbiology Laboratory - MPox
MPox Contextual Data Curation SOP 6.5

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.
6.5	October 2024	Charlie Barclay	Version compatibility update

