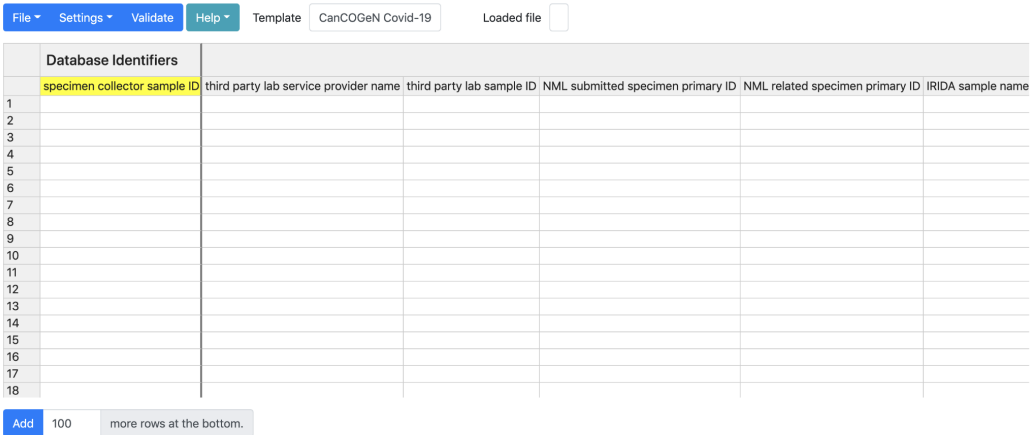


Contextual Data (Metadata) Curation

- I. **Purpose:** To harmonize SARS-CoV-2 contextual data across data providers in the CanCOGeN 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 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																																																																
1	<p>Download the zip file (“Source code (zip)”) containing The DataHarmonizer application from the following link:</p> <p>https://github.com/cidgoh/pathogen-genomics-package/releases/tag/PHPv2.0.3</p> <p>Releases / PHPv2.0.3</p> <div><div><h3>Pathogen Genomics Package 2.0.3</h3><div>Latest</div><div>ddooley released this Feb 17 · 1 commit to main since this release · PHPv2.0.3 · ae0874c</div><div>Includes DataHarmonizer v1.4.6</div><table><thead><tr><th>Pathogen Genomics Templates version</th><th>DH Version</th><th>Release Date</th><th>Template Name</th><th>Template Versionx.y.z</th><th>x changes (field)</th><th>y changes (values/IDs)</th><th>z changes (defs/formats/examples)</th></tr></thead><tbody><tr><td>CanCOGeN (SC2)</td><td>2.1.3</td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>DEXA (One Health)</td><td>1.0.0</td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>GISAID (SC2)</td><td>1.0.0</td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>GRDI</td><td>5.2.1</td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Monkeypox</td><td>3.3.3</td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Monkeypox-international</td><td>3.3.2</td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>PHA4GE (SC2)</td><td>1.0.1</td><td></td><td></td><td></td><td></td><td></td><td></td></tr></tbody></table></div></div> <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>	Pathogen Genomics Templates version	DH Version	Release Date	Template Name	Template Versionx.y.z	x changes (field)	y changes (values/IDs)	z changes (defs/formats/examples)	CanCOGeN (SC2)	2.1.3							DEXA (One Health)	1.0.0							GISAID (SC2)	1.0.0							GRDI	5.2.1							Monkeypox	3.3.3							Monkeypox-international	3.3.2							PHA4GE (SC2)	1.0.1						
Pathogen Genomics Templates version	DH Version	Release Date	Template Name	Template Versionx.y.z	x changes (field)	y changes (values/IDs)	z changes (defs/formats/examples)																																																										
CanCOGeN (SC2)	2.1.3																																																																
DEXA (One Health)	1.0.0																																																																
GISAID (SC2)	1.0.0																																																																
GRDI	5.2.1																																																																
Monkeypox	3.3.3																																																																
Monkeypox-international	3.3.2																																																																
PHA4GE (SC2)	1.0.1																																																																

CanCOGeN – SARS-CoV-2
CBW_1.0 Contextual Data Curation

	Action
	 <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 <code>xlsx</code>, <code>xls</code>, <code>tsv</code> and <code>csv</code> 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>
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>

CanCOGeN – SARS-CoV-2
CBW_1.0 Contextual Data Curation

	Action
--	--------

- 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, 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 <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</i>	sample collected by sequence submitted by sample collection date sample collection date precision geo_loc (country) geo_loc (province/territory) organism isolate purpose of sampling purpose of sampling details

CanCOGeN – SARS-CoV-2
CBW_1.0 Contextual Data Curation

	Action	
	<p><i>ID as specified by the lab. Be sure to keep a copy of the key.</i></p>	
	<p>Describing the material and/or site sampled.</p> <p><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></p>	<p>anatomical material anatomical part body product environmental material environmental site collection device collection_method NML submitted specimen type</p>
	<p>Host Information</p>	<p>host (scientific name) host disease host age host age unit host age bin host gender</p>
	<p>Sequencing</p>	<p>sequencing instrument sequencing date purpose of sequencing purpose of sequencing details</p>
	<p>Bioinformatics and QC Metrics</p>	<p>raw sequencing data processing method hosting method consensus sequence software name consensus sequence software version bioinformatics protocol</p>
6	<p>Validate the entered data by clicking on the Validate button on the top-left toolbar.</p>	

CanCOGeN – SARS-CoV-2
CBW_1.0 Contextual Data Curation

	Action
	<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. Export to IRIDA, GISAID, 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 CNPHI Collaboration Centre interface.</p> <ul style="list-style-type: none"> • Before uploading to CNPHI, 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 CNPHI 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> <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	Additional Information:

CanCOGeN – SARS-CoV-2
CBW_1.0 Contextual Data Curation

	Action
	<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

sequence submitted 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)

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

CanCOGeN – SARS-CoV-2
CBW_1.0 Contextual Data Curation

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 Severe acute respiratory syndrome coronavirus 2. This value is provided in the template.
e.g. Severe acute respiratory coronavirus 2

isolate

Identifier of the specific isolate.

Provide the isolate name. The isolate name should be identical to the GISAID virus name in the format "hCov-19/CANADA/xxxxx/2020", where xxxxx represents the sample ID.
e.g. hCov-19/CANADA/prov_rona_99/2020

purpose of sampling

The reason that the sample was collected.

The reason a sample was collected may provide information about potential biases in sampling strategy. Provide the purpose of sampling from the picklist in the template. Most likely, the sample was collected for Diagnostic testing. 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.

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. The sample was collected to investigate the prevalence of variants associated with mink-to-human transmission in Canada.

Describing the material and/or site sampled.

anatomical material

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

CanCOGeN – SARS-CoV-2
CBW_1.0 Contextual Data Curation

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@bccdc.ca. If not applicable, do not leave blank. Choose a null value.
e.g. Blood

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@bccdc.ca. If not applicable, do not leave blank. Choose a null value.
e.g. Nasopharynx (NP)

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@bccdc.ca. If not applicable, do not leave blank. Choose a null value.
e.g. Feces

environmental material

A substance or object obtained from the natural or man-made environment e.g. soil, water, sewage.

Provide a descriptor if an environmental material was sampled. 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. Face Mask

environmental site

An environmental location may describe a site in the natural or built environment e.g. contact surface, metal can, hospital, wet market, bat cave.

Provide a descriptor if an environmental site was sampled. 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. Building floor

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

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. Bronchoalveolar Lavage (BAL)

CanCOGeN – SARS-CoV-2
CBW_1.0 Contextual Data Curation

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

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.

This field is only required if there was a host. If the host was a human select COVID-19 from the pick list. If the host was asymptomatic, this can be recorded under "host health state details". If the host is not human, and the disease state is not known or the host appears healthy, put "Not Applicable".

e.g. COVID-19

host age

Age of host at the time of sampling.

Enter the age of the host in years. If not available, provide a null value. If there is not host, put "Not Applicable".

e.g. 79

host age bin

Age of host at the time of sampling, expressed as an age group.

Select the corresponding host age bin from the pick list provided in the template. If not available, provide a null value. The "host age bin" field will automatically propagate with the bin that corresponds to the input in "host age". If not available or you are not permitted to share, put a null value.

Age Bins:

0 - 9

10 -19

20 - 29

30 - 39

40 - 49

50 - 59

60 - 69

70 - 79

80 - 89

90 - 99

100+

CanCOGeN – SARS-CoV-2
CBW_1.0 Contextual Data Curation

host age unit

The unit used to measure the host age, in either months or years.

Indicate whether the host age is in months or years. Age indicated in months will be binned to the 0 - 9 year age bin.

host gender

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

Select the corresponding host gender from the pick list provided in the template. If not available, choose a null value. If there is no host, put "Not Applicable".

e.g. Male

Sequencing

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.

The reason an isolate was sequenced may provide information about potential biases in sequencing strategy. Provide the purpose of sequencing from the picklist in the template. Most likely, the sample was collected for Surveillance or Research. The reason why a sample was originally collected may differ from the reason why it was selected for sequencing. The reason for sample collection should be indicated in the "purpose of sampling" field.

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. The sample was sequenced to investigate the differences in lineages circulating in Canada during the spring and fall waves of the pandemic.

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

CanCOGeN – SARS-CoV-2
CBW_1.0 Contextual Data Curation

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 sequence 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 sequence 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

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.

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

CanCOGeN – SARS-CoV-2
CBW_1.0 Contextual Data Curation

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

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

e.g. nasal swab should be recorded:

host (scientific name)	host (common name)	host disease	anatomical part	collection device
Homo sapiens	Human	COVID-19	Nasopharynx (NP)	Swab

e.g. throat swab should be recorded:

host (scientific name)	host (common name)	host disease	anatomical part	collection device
Homo sapiens	Human	COVID-19	Oropharynx (OP)	Swab

e.g. saliva should be recorded:

host (scientific name)	host (common name)	host disease	anatomical material
Homo sapiens	Human	COVID-19	Saliva

e.g. salt water gargle should be recorded:

host (scientific name)	host (common name)	host disease	collection method
Homo sapiens	Human	COVID-19	Saline gargle (mouth rinse and gargle)

e.g. human feces should be recorded:

host (scientific name)	host (common name)	host disease	body product
Homo sapiens	Human	COVID-19	Feces

e.g. swab of a hospital bed rail should be recorded:

environmental site	environmental material	collection device
Hospital	Bed Rail	Swab

e.g. tissue from a bat (*Platyrrhinus lineatus*) in a cave should be recorded:

CanCOGeN – SARS-CoV-2
CBW_1.0 Contextual Data Curation

Host (common name)	Host (scientific name)	host disease	anatomical_part	environmental_site
Bat	Platyrrhinus lineatus	Not applicable	Tissue	Cave

e.g. *particulates from air filter* should be recorded:

environmental material	collection method
Particulate Matter	Air Filtration

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>

VII. Appendix D: Ethical, Practical, and Privacy Considerations

An effective and equitable response to the COVID-19 pandemic requires rapid and sustained international collaboration and data sharing. Many of the contextual data elements described in the CanCOGeN contextual data specification are critical for effective public health surveillance and response. However, many of these same elements have ethical, practical, and privacy issues which must be considered before data can be shared. Data governance policies may vary between data types and jurisdictions, thus users of the specification should consult data stewards and privacy officers regarding organization-specific and jurisdiction-specific policies. Below, we highlight a series of common issues and provide suggestions for ways forward.

CanCOGeN – SARS-CoV-2 CBW_1.0 Contextual Data Curation

Note: This guidance is not intended to apply to all situations and use cases. Decisions regarding implementation of the specification must ultimately be made by the user in consultation with data providers and data stewards. If the intended use of the information collected is for research purposes, there will likely be many additional administrative and ethical requirements (e.g. Research Ethics Board (REB) review).

Identifiers and Repository Accession Numbers

Sharing consensus sequence and raw data, as well as contextual data, with public repositories enables tracking of global spread of the SARS-CoV-2 virus, phylodynamics analyses, development and improvement of diagnostics, and much more. Laboratories world-wide are sharing SARS-CoV-2 sequence and minimal contextual data with public repositories such as GISAID and the INSDC. When you share information with a public database, you will receive an accession number (a unique identifier in a database enabling the tracking of multiple versions of the data). If you have shared data with a public database, make sure to capture the accession numbers. GISAID will provide you with a single accession number. Make sure to record it. INSDC members (NCBI, EMBL-EBI, DDBJ) may provide you with different accession numbers depending on what you share, and how. You can share assemblies and consensus sequences with GenBank, raw data with Sequence Read Archive (SRA), and contextual data as a BioSample (see reference guide for further information). Information may be organized in BioProjects, and at a higher organizational level, Umbrella BioProjects. Make sure to record all of the applicable accession numbers.

Samples, libraries, patients, sequences (raw, processed, consensus etc) and so on can have many identifiers, especially if there is a division of labour or sharing of information across agencies and organizations. The specification has provided fields to capture many of those that are common, but may not capture all of the IDs you require. **It is essential to track IDs of original materials and information** to establish chain-of-custody and for follow-up, if necessary. It is better to track too many IDs than too few. If you require more fields to capture the IDs you need, add them. Some IDs are considered public health identifiable information (PHII). Make sure to check with the appropriate authorities whether the IDs you plan to share are considered identifiable information. If considered identifiable, you may need to create an alternative set of IDs. If you do, make sure to store the key in a safe and secure place.

Geographical Information

Geographical information (country, province/state/region, city, postal code, latitude/longitude etc) is very informative for tracking spread of the virus at different scales. Detailed geographical information for human clinical samples is often considered PHII depending on the number of cases in that locality, or may be specially regulated, and so must be abstracted before it can be shared. If the specification is being used for a sequencing project and detailed geographic information can be recorded, additional standardized fields such as geo_loc name (city), geo_loc name (county), host contact information (postal code) can be added to your collection template as needed. It is important to note that most geographic location fields in the specification **describe the sample**. Other fields have been provided to capture geo_loc information about the origin of the host and the likely country of exposure. Curators should ensure that the information they are entering correctly refers to the sample or the host. Before sharing data, especially with public repositories, it is important to ensure the data being submitted complies with the permitted level of granularity. Discuss this with the data steward. If sharing latitude and longitude coordinates, do not use the centre of the city/region/province/state/country or the location of your agency as a proxy, as this implicates a real location and is misleading.

CanCOGeN – SARS-CoV-2
CBW_1.0 Contextual Data Curation

The “host origin geo_loc (country)” and suspected “location of exposure geo_loc name (country)” can be highly sensitive. If the information is shared and patients re-identified, it can have extreme consequences for the patient, the data collector, and the data provider, and political relations. However, this information is important for characterizing risk, understanding transmission, and how the disease impacts some groups more than others (i.e. due to systemic health care inequity, poverty, racism etc). There may also be issues of equitable access and benefit sharing that should be considered for genomics data, particularly regarding Indigenous communities. Institutional, national and international resources regarding these issues should be consulted for best practices.

Date Information

Geographical and temporal information are key elements of infectious disease surveillance programs. Temporal information consists of dates e.g. sample collection date, sample received date, sample sequenced date, symptom onset date etc. Dates can be considered PHII on their own for human clinical samples, or in combination with other types of contextual data (e.g. geographical information), or in context of how many cases have been reported in a locality. Elements such as “sample collection date” are usually held by the institution that collected the original specimen (e.g. performed the diagnostic test). As such, you may require permission to acquire this information, or it may be difficult to attain due to other burdens on the data provider (workload, system access, manual curation requirements). Alternatively, “received date” may be used as a substitute in the data you share.

Host Information

Outside of specifying the species scientific or common name, human host information is almost always considered PHII. Patient information is usually collected at the time of specimen collection (e.g. diagnostic test) using a case report form, and held by the institution that collected the original specimen. You will more than likely require permission to acquire this information, or it may be difficult to attain due to other burdens on the data provider (workload, system access, manual curation requirements).

“Host age” and “Host gender” are regularly collected for most surveillance programs and can be used to characterize case definitions, and for linkage between lab and epidemiological data. On their own, this information may not be considered PHII, however, they may be considered identifiable information when combined with other contextual data such as collection date and geographical location. Abstracting age information by using age binning is acceptable in the specification. Suggested age bins are as follows: 0-9 years, 10-19 years, 20-29 years, 30-39 years, 40-49 years, 50-59 years, 60-69 years, 70-79 years, 80+ years.

Methods Information

Methodological information, such as sampling and experimental design, laboratory procedures, bioinformatic processing, and quality control metrics, are crucial information to understand the context and limitations of analyses. Capturing as much well-structured information regarding your methods, and storing it in a centralized place (or single document) helps to future-proof the data as well as the work that went into collecting, processing, analyzing and interpreting the data. Capturing methodological information also enables better reproducibility, and increases quality control. The specification provides many fields for capturing experimental design, protocols, and scientific metrics. It is strongly recommended that as much of that information be captured and stored as possible.

Revision History

CanCOGeN – SARS-CoV-2
CBW_1.0 Contextual Data Curation

Version	Date	Writer	Description of Change
1.0	March 24 2023	Emma Griffiths	Created protocol