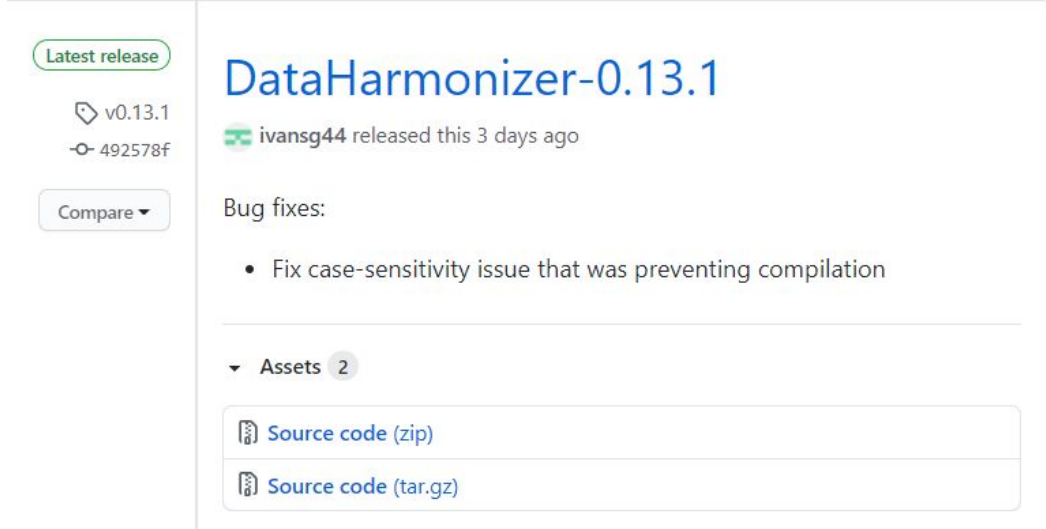


Contextual Data (Metadata) Curation

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

CanCOGeN – SARS-CoV-2

CanCOGeN_1.7 Contextual Data Curation

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

File ▾Settings ▾ValidateGetting Started

	Database Identifiers							
	specimen collector sample ID	PHAC sample ID	IRIDA sample name	umbrella bioproject accession	bioproject accession	biosample accession	SRA accession	Ge
1					▼			
2					▼			
3					▼			
4					▼			
5					▼			
6					▼			
7					▼			
8					▼			
9					▼			
10					▼			
11					▼			
12					▼			
13					▼			
14					▼			
15					▼			
16					▼			
17					▼			
18					▼			
19					▼			
20					▼			
21					▼			
22					▼			

Add100more rows at the bottom.

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.

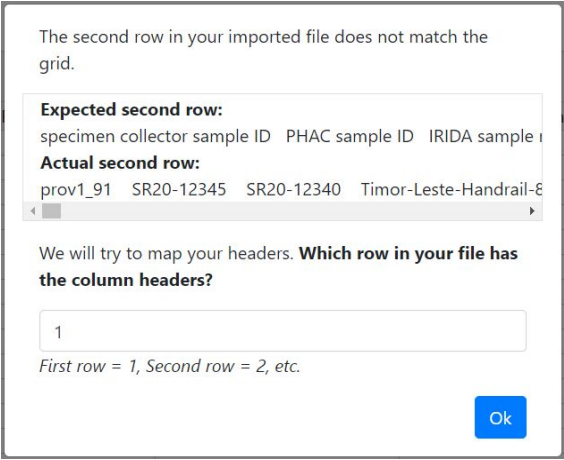
To import local data, click **File** on the top-left toolbar, and then click **Open**. To enter data in a new file, click **File** on the top-left toolbar, and then click **New**. Data entered into the spreadsheet can be copied and pasted.

Note: Only files containing the headers expected by the DataHarmonizer can be opened in the application. Example:

Database Identifiers					
specimen collector sample ID	NML submitter ID	NML related sample ID	IRIDA sample name	umbrella bioproject accession	bioproject accession
prov1_91					
prov1_92					
prov1_93					
prov1_94					

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

CanCOGeN – SARS-CoV-2
CanCOGeN_1.7 Contextual Data Curation

	Action	Related docs
	 <p><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). 	



CanCOGeN – SARS-CoV-2
CanCOGeN_1.7 Contextual Data Curation

	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 <u>required field</u> in each row. If data is missing or not collected, choose a null value from the picklist.<ul style="list-style-type: none">○ Not Applicable○ Missing○ Not Collected○ Not Provided○ Restricted Access• Free text can be provided when picklists are not available. <p>If a desired term is not present in a picklist, 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> <table><tr><th>Subsection</th><th>Required Fields</th></tr><tr><td>Sample Collection and Processing <i>Note: Evaluate with your supervisor whether the specimen collector sample ID is considered identifiable by your institutional policies. If not, copy the sample ID into the sample ID field in the validator spreadsheet. If yes, provide the alternative sample ID as specified by the lab. Be sure to keep a copy of the key.</i></td><td>specimen collector sample ID sample collected by sequence submitted by sample collection date geo_loc (country) geo_loc (province/territory) organism isolate purpose of sampling purpose of sampling details</td></tr><tr><td>Describing the material and/or site sampled.</td><td>anatomical material anatomical part</td></tr></table>	Subsection	Required Fields	Sample Collection and Processing <i>Note: Evaluate with your supervisor whether the specimen collector sample ID is considered identifiable by your institutional policies. If not, copy the sample ID into the sample ID field in the validator spreadsheet. If yes, provide the alternative sample ID as specified by the lab. Be sure to keep a copy of the key.</i>	specimen collector sample ID sample collected by sequence submitted by sample collection date geo_loc (country) geo_loc (province/territory) organism isolate purpose of sampling purpose of sampling details	Describing the material and/or site sampled.	anatomical material anatomical part	
Subsection	Required Fields							
Sample Collection and Processing <i>Note: Evaluate with your supervisor whether the specimen collector sample ID is considered identifiable by your institutional policies. If not, copy the sample ID into the sample ID field in the validator spreadsheet. If yes, provide the alternative sample ID as specified by the lab. Be sure to keep a copy of the key.</i>	specimen collector sample ID sample collected by sequence submitted by sample collection date geo_loc (country) geo_loc (province/territory) organism isolate purpose of sampling purpose of sampling details							
Describing the material and/or site sampled.	anatomical material anatomical part							

CanCOGeN – SARS-CoV-2
CanCOGeN_1.7 Contextual Data Curation

	Action	Related docs								
	<table><tr><td><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.</i></td><td>body product environmental material environmental site collection device collection_method</td></tr><tr><td>Host Information</td><td>host (scientific name) host disease host age host age units host age bin host gender</td></tr><tr><td>Sequencing</td><td>sequencing instrument purpose of sequencing purpose of sequencing details</td></tr><tr><td>Bioinformatics and QC Metrics</td><td>consensus sequence method</td></tr></table>	<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.</i>	body product environmental material environmental site collection device collection_method	Host Information	host (scientific name) host disease host age host age units host age bin host gender	Sequencing	sequencing instrument purpose of sequencing purpose of sequencing details	Bioinformatics and QC Metrics	consensus sequence method	
<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.</i>	body product environmental material environmental site collection device collection_method									
Host Information	host (scientific name) host disease host age host age units host age bin host gender									
Sequencing	sequencing instrument purpose of sequencing purpose of sequencing details									
Bioinformatics and QC Metrics	consensus sequence method									
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.● 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>									

CanCOGeN – SARS-CoV-2
CanCOGeN_1.7 Contextual Data Curation

	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 	
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 CNPHI 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 LaSER interface.</p> <ul style="list-style-type: none"> • Before uploading to CNPHI, export your data in “CNPHI” format by clicking File on the top-left toolbar, then clicking Export To. Type in the file name, and select “CNPHI” from the Format picklist. Then click Export. • See CNPHI documentation for more information regarding Metadata Upload. 	<p>Submission to CNPHI LaSER SOP:</p> <p>https://docs.google.com/document/d/e/2PACX-1vRUAWGPP44_Dlbmpx8FimS-jweA2d3KIBGs_h2JgUiWzKIXT_Q4gz44fCbyqqH1JAnprWsRaFHkjZMKj/pub</p>
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	<p>Optional: Format validated data for IRIDA submission.</p> <p>The DataHarmonizer will automate the preparation of an IRIDA submission form from the entered data by exporting the data in IRIDA format.</p> <ul style="list-style-type: none"> • Export your data in “IRIDA” format by clicking File on the top-left toolbar, then clicking Export To. Type in the file name, and select “IRIDA” from the Format picklist. Then click Export. <p><i>Note: If the top row containing the broad headings (Database identifiers, Sample collection and processing, Host information, Sequencing, Bioinformatics and QC, Authors) is not removed, the IRIDA metadata upload will fail. By</i></p>	<p>Upload to IRIDA SOP:</p> <p>https://irida.org/efacility.ca/documentation/user/samples/#adding-a-new-sample</p>

CanCOGeN – SARS-CoV-2
CanCOGeN_1.7 Contextual Data Curation

	Action	Related docs
	<i>exporting the sheet in IRIDA format, the DataHarmonizer completes this formatting for you.</i>	
12	<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>	

CanCOGeN – SARS-CoV-2
CanCOGeN_1.7 Contextual Data Curation

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”.

Sample Collection and Processing

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 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
CanCOGeN_1.7 Contextual Data Curation

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.

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

CanCOGeN – SARS-CoV-2
CanCOGeN_1.7 Contextual Data Curation

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)

Host Information

host (scientific name)

CanCOGeN – SARS-CoV-2
CanCOGeN_1.7 Contextual Data Curation

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+

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.

CanCOGeN – SARS-CoV-2
CanCOGeN_1.7 Contextual Data Curation

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

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

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

consensus sequence method

The name and version number of the protocol used to produce the consensus sequence.

Provide the software name followed by the version.
e.g. iVar 1.2

CanCOGeN – SARS-CoV-2
CanCOGeN_1.7 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. 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 (*Chiroptera*) in a cave should be recorded:

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

CanCOGeN – SARS-CoV-2
CanCOGeN_1.7 Contextual Data Curation

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>

CanCOGeN – SARS-CoV-2
CanCOGeN_1.7 Contextual Data Curation

Revision History

Version	Date	Writer	Description of Change
0.0	May 25, 2020	Lauren Tindale, Emma Griffiths	Created protocol
1.0	June 8, 2020	Emma Griffiths	Protocol edited
1.1	June 16, 2020	Emma Griffiths	Protocol edited
1.2	October 05, 2020	Rhiannon Cameron	Protocol edited
1.3	October 06, 2020	Rhiannon Cameron	Protocol edited
1.4	October 23, 2020	Rhiannon Cameron	Protocol & Appendix A edited
1.5	November 04, 2020	Rhiannon Cameron, Emma Griffiths	Protocol & Appendix A edited
1.6	November 12, 2020	Emma Griffiths, Rhiannon Cameron	Protocol & Appendix A edited
1.7	November 13, 2020	Rhiannon Cameron	Protocol edited