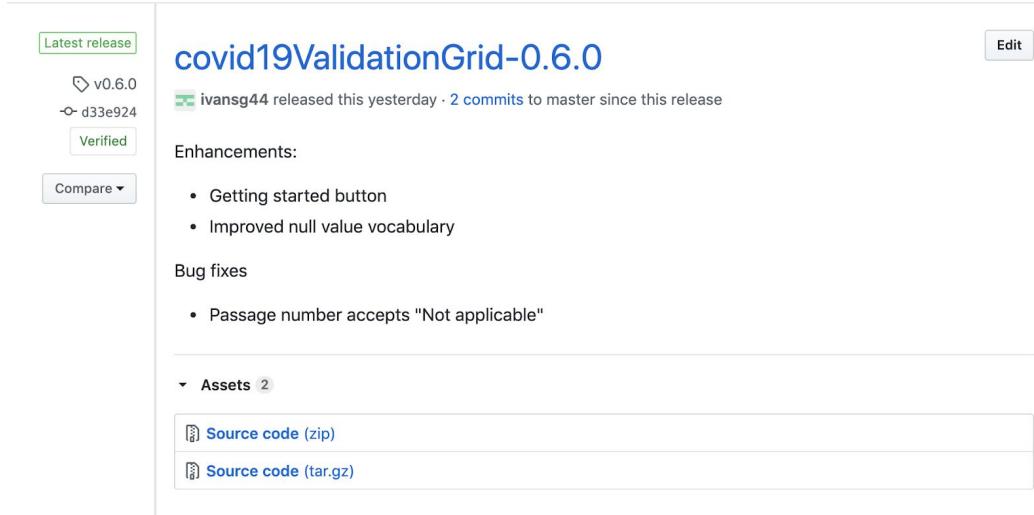
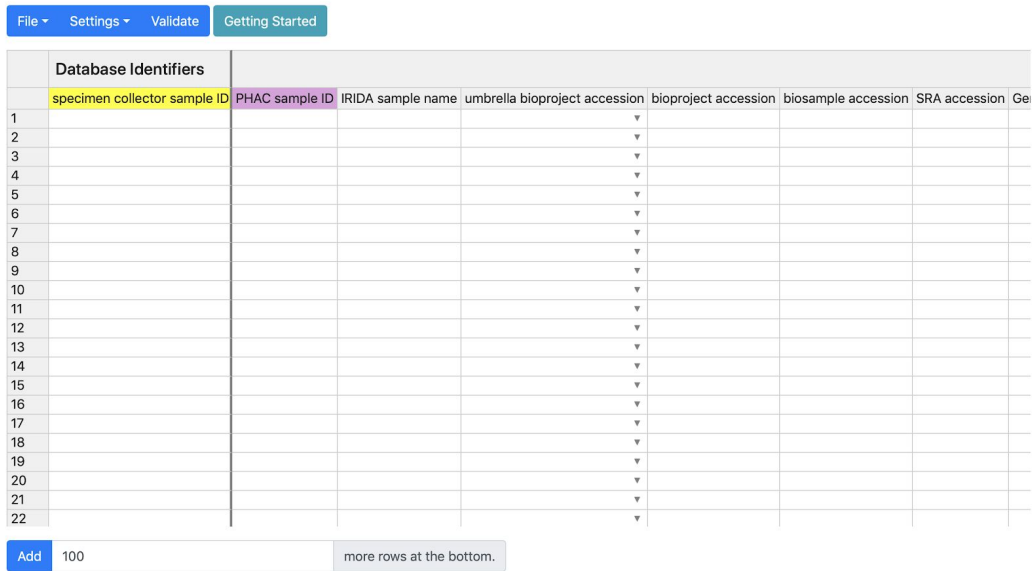


## Contextual Data (Metadata) Curation

- I. **Purpose:** To harmonize SARS-CoV-2 contextual data across data providers in the CanCOGeN network.
- 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: <a href="https://github.com/Public-Health-Bioinformatics/covid19ValidationGrid/releases">https://github.com/Public-Health-Bioinformatics/covid19ValidationGrid/releases</a></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\_1.1 Contextual Data Curation

	 <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 xlsx, tsv and csv files.</p> <p><i>Note: Only files containing the headers expected by the DataHarmonizer can be opened in the application.</i></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"> <li>• Review your dataset</li> <li>• Review the fields in the template of the Validator application</li> <li>• Review the field descriptions in the Appendix</li> </ul>	
3	<p>Familiarize yourself with DataHarmonizer functionality by reviewing the “Getting Started” section of the application (green button). Definitions, examples and further guidance are available by double clicking on the field headers.</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>	

CanCOGeN – SARS-CoV-2  
CanCOGeN\_1.1 Contextual Data Curation

- 5 Enter data into the validator spreadsheet.
- Hide non-required fields (colour-coded purple and white) 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).
  - 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.
  - Free text can be provided when picklists are not available.

**If a desired term is not present in a picklist, contact [emma.griffiths@bccdc.ca](mailto:emma.griffiths@bccdc.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
<b>Sample Collection and Processing</b>  <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
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</i>	anatomical material anatomical part body product environmental material environmental site collection device collection_method



CanCOGeN – SARS-CoV-2  
**CanCOGeN\_1.1 Contextual Data Curation**

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"> <li>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.</li> </ul>	
----	--	--

#### 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.

##### **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 organization or 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 e.g. Public Health Agency of Canada, Public Health Ontario, BC Centre for Disease Control.  
e.g. Public Health Ontario

###### **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

CanCOGeN – SARS-CoV-2  
**CanCOGeN\_1.1 Contextual Data Curation**

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 (country)**

*The country where the sample was collected.*

Use the controlled vocabulary provided in the template pick list.

e.g. Canada

**geo\_loc (province/territory)**

*The province/territory where the sample was collected.*

Use the controlled vocabulary provided in the template pick list.

e.g. Saskatchewan

**organism**

*The taxonomic name of the organism.*

Use Severe Acute Respiratory Coronavirus-2. This value is provided in the template.

**isolate**

*Identifier of the specific isolate.*

Provide the isolate name. This identifier should be an unique, indexed, alpha-numeric ID within your laboratory. The isolate name is often the same as the specimen collector sample ID.

Suggested: Isolate name should be identical to the GISAID virus name, which should be written in the format "hCov-19/CANADA/xxxxx/2020".

**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](mailto:emma.griffiths@bccdc.ca). If not applicable, do not leave blank. Choose a null value.

**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](mailto:emma.griffiths@bccdc.ca). If not applicable, do not leave blank. Choose a null value.

e.g. Nasopharynx

**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](mailto:emma.griffiths@bccdc.ca). If not applicable, do not leave blank. Choose a null value.

e.g. Feces

CanCOGeN – SARS-CoV-2  
**CanCOGeN\_1.1 Contextual Data Curation**

**environmental material**

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

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](mailto: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. 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](mailto:emma.griffiths@bccdc.ca). If not applicable, do not leave blank. Choose a null value.

e.g. Hospital

**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](mailto: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](mailto: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)**

*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".

**host age**

*Age of host at the time of sampling.*

CanCOGeN – SARS-CoV-2  
**CanCOGeN\_1.1 Contextual Data Curation**

Enter the age of the host in years at the time of same collection e.g. 47. If not available or you are not permitted to share, put a null value. Age bins are also acceptable and a picklist can be made available.

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

**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.

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. Illumina MiSeq

**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. ARTIC protocol v3

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	Swab

**e.g. throat swab** should be recorded:



CanCOGeN – SARS-CoV-2  
CanCOGeN\_1.1 Contextual Data Curation

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

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

environmental material	collection method
Particulate Matter	Air Filtration

## 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

CanCOGeN – SARS-CoV-2  
**CanCOGeN\_1.1 Contextual Data Curation**