



SOP for populating NCBI submission templates for SARS-CoV-2 (BioSample, SRA, and GenBank)

Forked from [Populating the NCBI pathogen metadata template](#)

Ruth Timme¹, Emma Griffiths², Duncan MacCannell³, Stacia Wyman⁴

¹US Food and Drug Administration; ²University of British Columbia; ³Centers for Disease Control and Prevention;

⁴Innovative Genomics Institute, UC Berkeley

1 Works for me dx.doi.org/10.17504/protocols.io.bf89jr26

GenomeTrakr Coronavirus Method Development Community 1 more workspace

Ruth Timme
US Food and Drug Administration

ABSTRACT

PURPOSE: Guidance on how to populate the three NCBI metadata submission templates for SARS-CoV-2 submissions, maximizing interoperability for COVID-19 surveillance.

DOI

dx.doi.org/10.17504/protocols.io.bf89jr26

PROTOCOL CITATION

Ruth Timme, Emma Griffiths, Duncan MacCannell, Stacia Wyman 2020. SOP for populating NCBI submission templates for SARS-CoV-2 (BioSample, SRA, and GenBank). **protocols.io**
dx.doi.org/10.17504/protocols.io.bf89jr26

FORK FROM

Forked from [Populating the NCBI pathogen metadata template](#), Ruth Timme

KEYWORDS

GenomeTrakr, metadata, Pathogen package, NCBI Pathogen Detection, INSDC

LICENSE

This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

CREATED

May 11, 2020

LAST MODIFIED

Jul 08, 2020

PROTOCOL INTEGER ID

36865

Three templates needed for NCBI SARS-CoV-2 submission

1 Guidance for populating the three NCBI metadata templates for SARS-CoV-2 submission.

START HERE FIRST: Read the [PHA4GE contextual data specification](#) BEFORE populating your submission templates!

Steps 3-4 help to map the NCBI submission template fields to the PHA4GE metadata specification, however, the primary PHA4GE guidance should be followed first to ensure the correct controlled vocabularies and ontology terms are used to populate these fields.

Guidance included in this protocol:

- **Step 2)** BioSample metadata template (modified for PHA4GE)
- **Step 3)** SRA metadata template
- **Step 4)** GenBank source modifier template (modified for PHA4GE)

BioSample metadata

2 PHA4GE pathogen template for BioSample submission

Download template here:

☐ [BioSample-Pathogen-PHA4GE_200708.xlsx](#)

2.1

attributes	guidance	example
sample_name	<p>Sample Name is a unique identifier for the sample (it cannot be left blank).</p> <p>Populate with the same ID as "Isolate", or include just the lab ID:.</p> <p>Populate this field using the values in the PHA4GE specification for "specimen collector sample ID".</p>	<p>Example 1: SARS-CoV-2/human/USA/MI-MDHHS-SC20654/2020</p> <p>Example 2: MI-MDHHS-SC20654</p>
bioproject_accession	<p>The accession number of the BioProject(s) to which the BioSample belongs (PRJNAxxxxx). A valid BioProject accession has prefix PRJN, PRJE or PRJD. This cannot be left blank.</p> <p>**Double-check that you are submitting to the correct data BioProject for your laboratory, and not the umbrella BioProject for the entire effort.</p> <p>Populate this field using the values in the PHA4GE specification for "bioproject accession".</p>	PRJNA12345
attribute_package	<p>Specify the pathogen type.</p> <p>"Pathogen.cl" (for clinical or host-associated pathogen) or "Pathogen.env" (for environmental, food or other pathogen).</p> <p>The value provided in this field drives validation of other fields.</p>	Pathogen.cl
organism	<p>Populate this field using the values in the PHA4GE specification for "organism".</p>	Severe acute respiratory syndrome coronavirus 2

isolate	<p>Full name of the virus: SARS-CoV-2/source/location/isolateID/year</p> <p>Populate this field using the values in the PHA4GE specification for "isolate".</p>	SARS-CoV-2/human/USA/CA-CDPH-001/2020
collected_by	<p>Full name of laboratory or institute that collected the sample.</p> <p>Populate this field using the values in the PHA4GE specification for "sample collected by".</p>	Utah Public Health Laboratory
collection_date	<p>the date on which the sample was collected; "YYYY-mm-dd", "YYYY-mm", or "YYYY". Including the month or month/day of collection is extremely valuable for accessing seasonality in the database.</p> <p>Populate this field using the values in the PHA4GE specification for "sample collection date".</p> <p>If unknown, put "not collected", or other null value like "missing".</p>	2020-05-15
geo_loc_name	<p>Geographical origin of the sample; use the appropriate name from this list http://www.insdc.org/documents/country-qualifier-vocabulary. Use a colon to separate the country from more detailed information about the location,</p> <p>Populate this field combining the PHA4GE specification fields for "geo_loc name (country)" and "geo_loc name (state/province/region)". Put more country field first, followed by state/province/region, separated by a colon.</p> <p>If unknown, put "not collected", or other null value like "missing".</p>	USA: Utah
isolation_source	<p>Describe the sample as "Clinical", "Animal", or "Environmental".</p> <p>Add the statement "See additional sample source fields for further information".</p>	Clinical, see additional sample source fields for further information
lat_lon	<p>The coordinates of the geographical location of sample or host collection.</p> <p>If known, provide the geographical coordinates of the location where the sample was collected. Specify as degrees latitude and longitude in the format "d[d.ddd] N S d[dd.ddd] W E", eg, 38.98 N 77.11 W.</p> <p>Populate this field by combining the PHA4GE specification fields "geo_loc latitude" and "geo_loc longitude"</p> <p>**DO NOT PROVIDE LAT/LON OF THE INSTITUTION, NOR THE CENTER OF A CITY/REGION WHERE THE SAMPLE WAS COLLECTED.</p> <p>If unknown, put "not collected", or other null value like "missing".</p>	<p>38.98 N 77.11 W</p> <p>not collected</p>

host	<p>If the combined attribute package is being used, this field can be left empty for Pathogen.ev isolates.</p> <p>Populate this field using the values in the PHA4GE specification for "host (scientific name)".</p> <p>This field is only required for host-associated samples (Pathogen.cl specified in attribute_package).</p>	Homo sapiens
host_disease	<p>This field is only required for host-associated samples (Pathogen.cl specified in attribute_package). If the combined attribute package is being used, this field can be left empty for Pathogen.ev isolates.</p> <p>Populate this field using the values in the PHA4GE specification for "host_disease".</p> <p>If the host is healthy or the information is unknown, put "missing"</p>	COVID-19
host_health_state	<p>Information regarding health state of the individual sampled at the time of sampling.</p> <p>Populate this field using the values in the PHA4GE specification for "host health state".</p> <p>If the information is unknown, or can not be shared, leave blank or provide a null value.</p>	asymptomatic
host_disease_outcome	<p>Final outcome of disease, e.g., death, chronic disease, recovery</p> <p>Populate this field using the values in the PHA4GE specification for "host disease outcome".</p> <p>If the information is unknown, or can not be shared, leave blank or provide a null value.</p>	e.g., death, chronic disease, recovery
host_sex	<p>The gender of the host at the time of sample collection.</p> <p>Populate this field using the values in the PHA4GE specification for "host gender".</p> <p>If the information is unknown, or can not be shared, leave blank or provide a null value.</p>	male
host_age	<p>Age of host at the time of sampling</p> <p>Populate this field using the values in the PHA4GE specification for "host age". Provide age in years. Age-binning is also acceptable.</p> <p>If the information is unknown or can not be shared, leave blank or provide a null value.</p>	31
host_subject_id	<p>a unique identifier by which each subject can be referred to, de-identified.</p> <p>Populate this field using the values in the PHA4GE specification for "host subject ID".</p> <p>If the information is unknown or can not be shared, leave blank or provide a null value.</p>	clinical123456

anatomical_material	<p>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, use this look-up service to identify a standardized term: https://www.ebi.ac.uk/ols/ontologies/uberon.</p> <p>Populate this field using the values in the PHA4GE specification for "anatomical material".</p> <p>If not applicable, leave blank.</p>	blood
anatomical_part	<p>An anatomical part of an organism e.g. oropharynx.</p> <p>Populate this field using the values in the PHA4GE specification for "anatomical part".</p> <p>If not applicable, leave blank.</p>	nasopharynx
body_product	<p>A substance excreted/secreted from an organism e.g. feces, urine, sweat.</p> <p>Populate this field using the values in the PHA4GE specification for "body product".</p> <p>If not applicable, leave blank.</p>	feces
passage_history	<p>Number of passages</p> <p>Populate this field using the values in the PHA4GE specification for "passage number".</p> <p>If not applicable, leave blank.</p>	3
lab_host	<p>Name and description of the laboratory host used to propagate the organism or material from which the sample was obtained.</p> <p>This field can be populated by the PHA4GE field "lab_host".</p> <p>If not applicable, leave blank.</p>	Vero E6 cell line

passage_method	<p>Passage method.</p> <p>Populate this field using the values in the PHA4GE specification for "passage method".</p> <p>If not applicable, leave blank.</p>	<p>AVL buffer+30 %EtOH lysate received from Respirator y Lab. P3 passage in Vero-1 via bioreactor large- scale batch passage. P3 batch derived from the SP- 2/referenc e lab strain.</p>
purpose_of_sampling	<p>The reason that the sample was collected.</p> <p>Populate this field using the values in the PHA4GE specification for "purpose of sampling".</p> <p>If not applicable, leave blank.</p>	<p>Select a value from the pick list in the template. (e.g. diagnostic testing)</p>
environmental_material	<p>A substance obtained from the natural or man-made environment e.g. soil, water, sewage, door handle, bed handrail, face mask.</p> <p>Populate this field using the values in the PHA4GE specification for "environmental material".</p> <p>If not applicable, leave blank.</p>	<p>face mask</p>
environmental_site	<p>An environmental location may describe a site in the natural or built environment e.g. hospital, wet market, bat cave.</p> <p>Populate this field using the values in the PHA4GE specification for "environmental site".</p> <p>If not applicable, leave blank.</p>	<p>hospital room</p>
collection_device	<p>The instrument or container used to collect the sample e.g. swab.</p> <p>Populate this field using the values in the PHA4GE specification for "collection device".</p> <p>If not applicable, leave blank.</p>	<p>swab</p>

collection_method	<p>The process used to collect the sample e.g. phlebotomy, necropsy.</p> <p>Populate this field using the values in the PHA4GE specification for "collection method".</p> <p>If not applicable, leave blank.</p>	Bronchoalveolar lavage (BAL)
culture_collection	<p>Name of source institute and unique culture identifier for the organism (SARS-CoV-2).</p> <p>See the description for the proper format and list of allowed institutes, http://www.insdc.org/controlled-vocabulary-culturecollection-qualifier</p> <p>If not applicable, leave blank.</p>	
host_specimen_voucher	<p>Identifier for the physical host specimen.</p> <p>Populate this field using the values in the PHA4GE specification for "host specimen voucher".</p> <p>if not applicable, leave blank.</p>	<p>URI example: http://portal.vertnet.org/o/fmnh/mammals?id=33e55cfe-330b-40d9-aaae-8d042cba7542,</p> <p>INSDC triplet example: UAM:Mamm:52179</p>
description	Optional field for additional description of the sample.	

SRA metadata

3 Populate SRA's batch metadata table:

Download template here:

ftp://ftp-trace.ncbi.nlm.nih.gov/sra/metadata_table/SRA_metadata.xlsx

PRO TIPS:

1. If you have sequences to submit that belong to more than one BioProject, create a separate submission + metadata table for each of your BioProjects.
2. *Entering fastq filenames in the spreadsheet.* On a Mac, you can directly copy the file names from the folder into a spreadsheet. This is not possible on a PC using copy and paste but can be done with some command-line operation.
3. Finally, it is important to develop a QA/QC step to make sure the files are associated with the correct sample name. For example, use a left function in excel to strip of the appended text in the file name and then use the exact match to make sure the name matches the sample name.

3.1

Field	Description	Example
-------	-------------	---------

sample_name	<p>Include the same ID here as you entered for "sample_name" in the BioSample submission template.</p> <p>Populate this field using the values in the PHA4GE specification for "specimen collector sample ID".</p>	UT-12345
library_ID	<p>The library name should be a unique ID relevant to your workflow. It can be an autogenerated ID from your LIMS system or a modification of your sample_name.</p> <p>Populate this field using the values in the PHA4GE specification for "library_id".</p>	UT-12345.6
Title	<p>Short, free text description that identifies the data on public pages.</p> <p>For Example: {methodology} of {organism}: {sample_name}</p>	Amplicon-based sequencing of SARS-CoV-2: UT-12345
library_strategy	<p>Overall sequencing strategy or approach. Choose from NCBI pick list</p>	See NCBI SRA pick list. (e.g. WGS, RNA-Seq, Amplicon)
library_source	molecule type used to make the library	See NCBI SRA pick list. (e.g. viral RNA, metagenomic)
library_selection	Library capture method	See NCBI SRA pick list. (e.g. random, PCR)
Library_layout	Choose from NCBI pick list	See NCBI SRA pick list. (single, paired)
platform	Sequencing platform	See NCBI SRA pick list. (e.g., Illumina, Oxford_nanopore, PacBio_SMR).

instrument_model	Name of the sequencing instrument. Populate this field using the values in the PHA4GE specification for "sequencing instrument"	See NCBI SRA pick list. (e.g. Illumina MiSeq, iSeq 100, GridION, MinION, PacBio Sequel II)
Design_description	optional field for free text description of methods	ARTIC PCR-tiling of viral cDNA (V3), sequenced on Illumina MiSeq with DNA Flex library prep-kit. Only reads aligned to SARS-CoV-2 reference (NC_045512.2) retained
Filetype	File format name for the raw sequence data Choose from NCBI pick list	See NCBI SRA pick list. (e.g. Fastq, OxfordNano pore_native, PacBio_HDF5)
Filename	include ALL of the files resulting from this library. **Add additional fields if there are more than two files (e.g. Filename3). Populate this field using the values in the PHA4GE specification for "r1 fastq filename".	genome_r1.fastq (*must be exact)
Filename2	genome_r2.fastq (*must be exact) Populate this field using the values in the PHA4GE specification for "r2 fastq filename".	genome_r2.fastq (*must be exact)
Filename3-8	list other fastq file names (e.g. for NextSeq data)	

Save the second sheet (SRA_data) as a TSV (tab-delimited file) for upload in the "SRA metadata" tab within the submission portal.

*NCBI should also accept the original excel formatted file.

4 Populate GenBank source modifier template:

 **GenBank-source_modifiers-PHA4GE_200708.xlsx**

4.1

Source modifier	Guidance	Example
Sequence_ID	<p>ID to link the fasta sequence to these source modifiers for a batch submission. This ID must match the ID listed as the fasta file header.</p> <p>Suggested: use the sample_name submitted to BioSample, or the short lab ID contained within the full isolate name.</p> <p>Populate this field using the values in the PHA4GE specification for "specimen collector sample ID".</p>	CA-CDPH-001
country	<p>This field can be populated by combining the PHA4GE fields "geo_loc name (country)" and "geo_loc name (state/province/region)". Put more country field first, followed by state/province/region, separated by a colon.</p>	USA: Virginia
host	<p>This field is only required for host-associated samples (Pathogen.cl specified in attribute_package). Leave blank for environmental isolates.</p> <p>Populate this field using the values in the PHA4GE specification for "host (scientific name)".</p>	Homo sapiens
isolate	<p>Full name of the virus:</p> <p>SARS-CoV-2/source/location/isolateID/year</p> <p>Example: SARS-CoV-2/human/USA/CA-CDPH-001/2020</p> <p>Populate this field using the values in the PHA4GE specification for "isolate".</p>	SARS-CoV-2/human/USA/CA-CDPH-001/2020
collection-date	<p>the date on which the sample was collected; "YYYY-mm-dd", "YYYY-mm", or "YYYY"</p> <p>Populate this field using the values in the PHA4GE specification for "sample collection date".</p>	2020-06-04
isolation-source	<p>"Clinical", "Animal", or "Environmental".</p> <p>Add the statement "See additional sample source fields for further information."</p>	clinical; See additional sample source fields for further information
BioSample	<p>The accession number of the BioSample registered for this sample.</p> <p>Populate this field using the values in the PHA4GE specification for "biosample accession".</p>	SAMN15187145
BioProject	<p>The accession number of the BioProject(s) to which the BioSample belongs.</p> <p>Populate this field using the values in the PHA4GE specification for "bioproject accession".</p>	PRJNA625551

