Jul 09, 2020

SOP for populating EBI submission templates (ENA)

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

Nabil-Fareed Alikhan[1], Emma Griffiths[2], Ruth Timme[3], Duncan MacCannell[4]

[1]Quadram Institute Bioscience; [2]University of British Columbia; [3]US Food and Drug Administration; [4]Centers for Disease Control and Prevention

1 Works for me dx.doi.org/10.17504/protocols.io.bh5dj826

Coronavirus Method Development Community PHA4GE

Nabil-Fareed Alikhan
Quadram Institute Bioscience

ABSTRACT

Guidance on how to populate the extended PHA4GE metadata package for SARS-CoV-2 submissions, maximizing interoperability for covid-19 surveillance.

DOI

dx.doi.org/10.17504/protocols.io.bh5dj826

PROTOCOL CITATION

Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell 2020. SOP for populating EBI submission templates (ENA). **protocols.io** dx.doi.org/10.17504/protocols.io.bh5dj826

FORK FROM

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

KEYWORDS

metadata, INSDC, ERC000033, ENA, EBI, SARS-Cov2, COVID-19

CREATED

Jul 01, 2020

LAST MODIFIED

Jul 09, 2020

PROTOCOL INTEGER ID

38789

Three templates needed for NCBI SARS-CoV-2 submission

1 Guidance for populating the three templates for SARS-CoV-2 submission to EBI.

This protocol helps to describe the fields needed to populate the ENA virus pathogen checklist, however, **the primary PHA4GE guidance should be followed** to ensure the controlled vocabularies and ontology terms are used to populate these fields.

Link to PHA4GE SARS-CoV-2 metadata specification: https://github.com/pha4ge/SARS-CoV-2-Contextual-

Citation: Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell (07/09/2020). SOP for populating EBI submission templates (ENA). https://dx.doi.org/10.17504/protocols.io.bh5dj826

Data-Specification

Link to PHA4GE SARS-CoV-2 EBI submission protocol: ENA, BioSample, and BioProject:
https://www.protocols.io/private/BD4C35AC52C942B2D927E662ABC3D195

Link to PHA4GE SARS-CoV2 EBI assembly submission protocol
https://www.protocols.io/private/39BBC8E9B6F911EAA1530A58A9FEAC2A

ENA virus pathogen reporting

2 PHA4GE ENA virus pathogen reporting standard checklist for sample metadata:

ENA submission spreadsheets are not tables with fields name in the first row, and specfics about each sample in subsequent rows. Instead the spreadsheet requires particular headers (denoted with "#"). e.g

#checklist_accession	*ERC000033*					
#unique_name_prefix						
sample_name	**tax_id**	**scientific_name**	**host age**	**organism**	**collection date**	**geographic location (country and/or sea)**
hCOV-sample-ENG-11	2697049	Severe acute respiratory syndrome coronavirus 2	50	Severe acute respiratory syndrome coronavirus 2	05/05/2020	United Kingdom
#template						
#units			*years*			

As described in the **PHA4GE SARS-CoV-2 EBI submission protocol**, there will be an oppotunity to download a template spreadsheet from ENA directly. We recommend that you add aditional fields as per the PHA4GE guidelines and keep a blank copy of this spreadsheet for subsequent submissions.

2.1 Guidlines to populate the sample metadata sheet

These fields are required for all ENA submission. They do not appear as part of the checklist. The description and guidance is tabulated below.

Description of required fields

ENA Required Fields	Definition
sample_name	The user-provided name of the sample.
tax_id	The NCBITaxon identifier for the organism being sequenced.
scientific_name	The taxonomic name of the organism.
common_name	The common name of the organism.
sample_description	Free text description of the sample.

Citation: Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell (07/09/2020). SOP for populating EBI submission templates (ENA). https://dx.doi.org/10.17504/protocols.io.bh5dj826

instrument_model	Name of the sequencing instrument.
library_source	Molecule type used to make the library.
library_selection	Library capture method.
library_strategy	Overall sequencing strategy or approach.
library_layout	Single or paired.
file_name	Include ALL of the files resulting from this library. **Add additional fields if there are more than two files (e.g. Filename3)

All sample submissions to ENA require this information. The description and mapping to the respective PHA4GE field should help you transfer your metadata from the PHA4GE table to the something acceptable for ENA submission.

Guidance of required fields

ENA Required Fields	PHA4GE Field	PHA4GE Guidance
sample_name	specimen collector sample ID	This field can be populated by the PHA4GE field "specimen collector sample ID".
tax_id	N/A	Use "2697049" as the tad_id for SARS-CoV-2.
scientific_name	organism	This field can be populated by the PHA4GE field "organism". Provide the full name "Severe acute respiratory syndrome coronavirus 2".
common_name	N/A	This field can be populated by the PHA4GE field "host (common name)". Provide "Sars-CoV-2".
sample_description	N/A	
instrument_model	N/A	See ENA SRA pick list. (e.g. Illumina MiSeq, iSeq 100, GridION, MinION, PacBio Sequel II)
library_source	N/A	See ENA SRA pick list. (e.g. viral RNA, metagenomic)

Citation: Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell (07/09/2020). SOP for populating EBI submission templates (ENA). https://dx.doi.org/10.17504/protocols.io.bh5dj826

library_selection	N/A	See ENA SRA pick list. (e.g. random, PCR)
library_strategy	N/A	See ENA SRA pick list. (e.g. WGS, RNA-Seq, Amplicon)
library_layout	N/A	See ENA SRA pick list. (single, paired)
file_name	r1 fastq filename	This field can be populated by the PHA4GE field "r1 fastq filename".

All sample submissions to ENA require this information. The description and mapping to the respective PHA4GE field should help you transfer your metadata from the PHA4GE table to the something acceptable for ENA submission.

For SARSCov2 submission, ENA ask that the metadata comply with Checklist ERC000033: https://www.ebi.ac.uk/ena/browser/view/ERC000033. These fields below are drawn from the ENA Checklist, with description and mapping to the respective PHA4GE field should help you transfer your metadata from the PHA4GE table to the something acceptable for ENA submission.

ENA Virus Checklist Field	ENA Definition	ENA Requirement Status
subject exposure	Exposure of the subject to infected human or animals, such as poultry, wild bird or swine. If multiple exposures are applicable, please state them separated by semicolon. Example: poultry; wild bird	optional
subject exposure duration	Duration of the exposure of the subject to an infected human or animal. If multiple exposures are applicable, please state their duration in the same order in which you reported the exposure in the field 'subject exposure'. Example: 1 day; 0.33 days	optional
type exposure	Setting within which the subject is exposed to animals, such as farm, slaughterhouse, food preparation. If multiple exposures are applicable, please state their type in the same order in which you reported the exposure in the field 'subject exposure'. Example: backyard flock; confined animal feeding operation	optional
personal protective equipment	Use of personal protective equipment, such as gloves, gowns, during any type of exposure. Example: mask	optional
hospitalisation	Was the subject confined to a hospital as a result of virus infection or problems occurring secondary to virus infection?	optional
illness duration	The number of days the illness lasted. Example: 4	optional
illness symptoms	The symptoms that have been reported in relation to the illness, such as cough, diarrhea, fever, headache, malaise, myalgia, nausea, runny_nose, shortness_of_breath, sore_throat. If multiple exposures are applicabl	optional
collection date	The date of sampling, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated i.e. all of these are valid ISO8601 compliant times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008.	recommended

Citation: Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell (07/09/2020). SOP for populating EBI submission templates (ENA). https://dx.doi.org/10.17504/protocols.io.bh5dj826

geographic location (country and/or sea)	The geographical origin of the sample as defined by the country or sea. Country or sea names should be chosen from the INSDC country list (http://insdc.org/country.html).	mandatory
geographic location (latitude)	The geographical origin of the sample as defined by latitude and longitude. The values should be reported in decimal degrees and in WGS84 system	recommended
geographic location (longitude)	The geographical origin of the sample as defined by latitude and longitude. The values should be reported in decimal degrees and in WGS84 system	recommended
geographic location (region and locality)	The geographical origin of the sample as defined by the specific region name followed by the locality name.	recommended
sample capture status	Reason for the sample collection.	recommended
host disease outcome	Disease outcome in the host.	recommended
host common name	common name of the host, e.g. human	mandatory
host subject id	a unique identifier by which each subject can be referred to, de-identified, e.g. #131	mandatory
host age	age of host at the time of sampling; relevant scale depends on species and study, e.g. could be seconds for amoebae or centuries for trees	recommended
host health state	Health status of the host at the time of sample collection.	mandatory
host sex	Gender or sex of the host.	mandatory
host scientific name	Scientific name of the natural (as opposed to laboratory) host to the organism from which sample was obtained.	mandatory
virus identifier	Unique laboratory identifier assigned to the virus by the investigator. Strain name is not sufficient since it might not be unique due to various passsages of the same virus. Format: up to 50 alphanumeric characters	recommended
collector name	Name of the person who collected the specimen. Example: John Smith	mandatory
collecting institution	Name of the institution to which the person collecting the specimen belongs. Format: Institute Name, Institute Address	mandatory
receipt date	Date on which the sample was received. Format:YYYY-MM-DD. Please provide the highest precision possible. If the sample was received by the institution and not collected, the 'receipt date' must be provided instead. Either the 'collection date' or 'receipt date' must be provided. If available, provide both dates.	recommended
sample storage conditions	Conditions at which sample was stored, usually storage temperature, duration and location	optional
definition for seropositive sample	The cut off value used by an investigatior in determining that a sample was seropositive.	recommended
serotype (required for a seropositive sample)	Serological variety of a species characterised by its antigenic properties. For Influenza, HA subtype should be the letter H followed by a number between 1-16 unless novel subtype is identified and the NA subtype should be the letter N followed by a number between 1-9 unless novel subtype is identified. If only one of the subtypes have been tested then use the format H5Nx or HxN1. Example: H1N1	recommended
isolate	individual isolate from which the sample was obtained	mandatory
strain	Name of the strain from which the sample was obtained.	optional
host habitat	Natural habitat of the avian or mammalian host.	recommended
isolation source host-associated	Name of host tissue or organ sampled for analysis. Example: tracheal tissue	recommended

Citation: Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell (07/09/2020). SOP for populating EBI submission templates (ENA). https://dx.doi.org/10.17504/protocols.io.bh5dj826

host description	Other descriptive information relating to the host.	optional
gravidity	Whether or not the subject is gravid. If so, report date due or date post-conception and specify which of these two dates is being reported.	optional
host behaviour	Natural behaviour of the host.	recommended
isolation source non-host-associated	Describes the physical, environmental and/or local geographical source of the biological sample from which the sample was derived. Example: soil	recommended

Fields from ENA Checklist ERC000033. The description and mapping to the respective PHA4GE field should help you transfer your metadata from the PHA4GE table to the something acceptable for ENA submission.

ENA Virus Checklist Field	PHA4GE Field	PHA4GE Guidance
subject exposure	exposure event	This field can be populated by the PHA4GE field "exposure event". Caution: this may be sensistive information. Consult the data steward before sharing. If the information is unknown, not applicable, or can not be shared, leave blank or provide a null value.
subject exposure duration	N/A	If the information is unknown, not applicable, or not available, leave blank or provide a null value.
type exposure	exposure event	The PHA4GE field "exposure event" describes subject exposures and settings. Exposure information captured in the PHA4GE specification can be provided either in the ENA "subject exposure" or the "type exposure" field.
personal protective equipment	N/A	If the information is unknown, not applicable, or not available, leave blank or provide a null value.

Citation: Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell (07/09/2020). SOP for populating EBI submission templates (ENA). https://dx.doi.org/10.17504/protocols.io.bh5dj826

hospitalisation	specified as value under host health status	This information can be found as values in the "host health status" field in the PHA4GE specification e.g. Hospitalized, Hospitalized (ICU), Hospitalized (Non-ICU). The ENA "hospitalisation" field requires yes/no values. Provide yes/no values for this field if submitting to ENA. If the information is unknown, or can not be shared, leave blank or provide a null value.
illness duration	N/A	If the information is unknown, not applicable, or not available, leave blank or provide a null value.
illness symptoms	signs and symptoms	This field can be populated by the PHA4GE field "signs and symptoms". If the information is unknown, not applicable, or not available, leave blank or provide a null value.
collection date	sample collection date	This field can be populated by the PHA4GE field "sample collection date". Caution: the sample collection date may be considered public health identifiable information. Consult the data steward before sharing. Acceptable formats are YYYY-MM-DD, YYYY-MM, or YYYY.
geographic location (country and/or sea)	geo_loc (country)	This field can be populated by the PHA4GE field "geo_loc name (country)". The values in the PHA4GE pick list are derived from the INSDC country list.

Citation: Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell (07/09/2020). SOP for populating EBI submission templates (ENA). https://dx.doi.org/10.17504/protocols.io.bh5dj826

geographic location (latitude)	geo_loc latitude	This field can be populated by the PHA4GE field "geo_loc latitude". Caution: this is likely sensitive information. Consult the data steward before sharing. Do not provide latitude of the institution, nor the centre of the city/region where the sample was collected as this falsely implicates an existing location. If the information is unknown or can not be shared, leave blank or provide a null value.
geographic location (longitude)	geo_loc longitude	This field can be populated by the PHA4GE field "geo_loc longitude". Caution: this is likely sensitive information. Consult the data steward before sharing. Do not provide longitude of the institution, nor the centre of the city/region where the sample was collected as this falsely implicates an existing location. If the information is unknown or can not be shared, leave blank or provide a null value.
geographic location (region and locality)	geo_loc (state/province/region)	This field can be populated by the PHA4GE field "geo_loc (state/province/region". Caution: this maye be sensitive information depending on the number of cases in this geographic area. Consult the data steward before sharing. If the information is unknown or can not be shared, leave blank or provide a null value.

Citation: Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell (07/09/2020). SOP for populating EBI submission templates (ENA). https://dx.doi.org/10.17504/protocols.io.bh5dj826

sample capture status	purpose of sampling	While the meanings of ENA's "sample capture status" and PHA4GE's "purpose of sampling" fields overlap in meaning, ENA provides a specific pick list of terms for populating the field. Use the ENA pick list if submitting to ENA. If the information is unknown or can not be shared, leave blank or provide a null value.
host disease outcome	host disease outcome	While the meanings of ENA's "host disease outcome" and PHA4GE's "host disease outcome" fields overlap in meaning, ENA provides a specific pick list of terms for populating the field. Use the ENA pick list if submitting to ENA. If the information is unknown or can not be shared, leave blank or provide a null value.
host common name	host (common name)	This field can be populated by the PHA4GE field "host (common name)".
host subject id	host subject ID	This field can be populated by the PHA4GE field "host subject ID". Caution: the host subject ID may be considered public health identifiable information. Consult the data steward before sharing. If unknown or considered identifiable, provide an alternative ID or a null value.

Citation: Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell (07/09/2020). SOP for populating EBI submission templates (ENA). https://dx.doi.org/10.17504/protocols.io.bh5dj826

host age	host age	This field can be populated by the PHA4GE field "host age". Caution: the host age may be considered public health identifiable information. Consult the data steward before sharing. If the information is unknown or can not be shared, leave blank or provide a null value.
host health state	host health state	While the meanings of the ENA and PHA4GE "host health state" fields overlap, ENA requires certain values in this field. If known, provide "diseased" or "healthy". If the information is unknown or can not be shared, provide a null value.
host sex	host gender	While the meanings of ENA's "host sex" and PHA4GE's "host gender" fields overlap in meaning, ENA provides a specific pick list of terms for populating the field. Use the ENA pick list if submitting to ENA. Caution: the host gender may be considered public health identifiable information. Consult your data steward before sharing. If the information is unknown or can not be shared, provide a null value.
host scientific name	host (scientific name)	This field can be populated by the PHA4GE field "host (scientific name)".

Citation: Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell (07/09/2020). SOP for populating EBI submission templates (ENA). https://dx.doi.org/10.17504/protocols.io.bh5dj826

virus identifier	specimen collector sample ID	This field can be populated by the PHA4GE field "specimen collector sample ID". Caution: the sample ID may be considered sensitive information. Consult the data steward. You may need to provide an alternative ID.
collector name	N/A	Provide the name of the person who collected the sample. If the information is unknown or can not be shared, provide a null value.
collecting institution	sample collected by	This field can be populated by the PHA4GE field "sample collected by". Caution: if the name of the lab reveals geographic information, this may be considered public health identifiable information. Consult the data steward before sharing. If information ca not be shared, provide a null value.
receipt date	received date	This field can be populated by the PHA4GE field "received date". If the information is unknown or can not be shared, leave blank or provide a null value.
sample storage conditions	N/A	If the information is unknown, not applicable, or not available, leave blank or provide a null value.
definition for seropositive sample	N/A	If the information is unknown, not applicable, or not available, leave blank or provide a null value.
serotype (required for a seropositive sample)	N/A	If the information is unknown, not applicable, or not available, leave blank or provide a null value.

Citation: Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell (07/09/2020). SOP for populating EBI submission templates (ENA). https://dx.doi.org/10.17504/protocols.io.bh5dj826

isolate	isolate	This field can be populated by the PHA4GE field "isolate".
strain	N/A	If the information is unknown, not applicable, or not available, leave blank or provide a null value.
host habitat	N/A	If the information is unknown, not applicable, or not available, leave blank or provide a null value.
isolation source host-associated	anatomical material; anatomical part; body product	This field can be populated by the PHA4GE fields "anatomical material", "anatomical part" and "body product". If the information is unknown, not applicable, or can not be shared, leave blank or provide a null value.
host description	N/A	Can be left blank.
gravidity	N/A	If the information is unknown, not applicable, or not available, leave blank or provide a null value.
host behaviour	N/A	If the information is unknown, not applicable, or not available, leave blank or provide a null value.
isolation source non-host-associated	environmental site; environmental material	This field can be populated by the PHA4GE fields "environmental site" and "environmental material". If the information is unknown, not applicable, or can not be shared, leave blank or provide a null value.

Fields from ENA Checklist ERC000033. The description and mapping to the respective PHA4GE field should help you transfer your metadata from the PHA4GE table to the something acceptable for ENA submission.

There is extended guideance available at the **PHA4GE SARS-CoV-2 metadata specification:** https://github.com/pha4ge/SARS-CoV-2-Contextual-Data-Specification

ENA run metadata

Citation: Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell (07/09/2020). SOP for populating EBI submission templates (ENA). https://dx.doi.org/10.17504/protocols.io.bh5dj826

3 Populate ENA run metadata table:

PRO TIPS:

1. If you have sequences to submit that have drastically different metadata, create a separate submission + metadata table for each case.
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 As described in the **PHA4GE SARS-CoV-2 EBI submission protocol**, there will be an oppotunity to download a template spreadsheet from ENA directly. There is a description of these fields in the table below.

Field	Description	Example
Sample reference	Include the same ID here as you entered for "sample_name" in the BioSample submission template. This field can be populated by the PHA4GE field "specimen collector sample ID".	UT-12345
Library name	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. This field can be populated by the PHA4GE field "library_id".	UT-12345.6
Title	Short, free text description that identifies the data on public pages. For Example: {methodology} of {organism}: {sample_name}	Amplicon-based sequencing of SARS-CoV-2: UT-12345
Library strategy	Overall sequencing strategy or approach. Choose from NCBI pick list	e.g. WGS, RNA-Seq, Amplicon
Library source	molecule type used to make the library	e.g. viral RNA, metagenomic
Library selection	Library capture method	e.g. random, PCR
instrument model	Name of the sequencing instrument	e.g. Illumina MiSeq, iSeq 100, GridION, MinION, PacBio Sequel II

Citation: Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell (07/09/2020). SOP for populating EBI submission templates (ENA). https://dx.doi.org/10.17504/protocols.io.bh5dj826

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
File name	Includes files resulting from this library. This maybe named "First file name" if multiple files need to be submitted This field can be populated by the PHA4GE field "r1 fastq filename".	genome_r1.fastq (*must be exact)
Second file name	genome_r2.fastq (*must be exact) This field can be populated by the PHA4GE field "r2 fastq filename". This field will only be shown for certain file types (i.e. paired FASTQ)	genome_r2.fastq (*must be exact)
Filename3-8	list other fastq file names (e.g. for NextSeq data)	
(First/Second) MD5 Checksum	MD5 checksum of the file being submitted	07182d8b0...

Citation: Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell (07/09/2020). SOP for populating EBI submission templates (ENA). https://dx.doi.org/10.17504/protocols.io.bh5dj826