

Feb 27, 2024

Version 11

Guidance for populating and validating GenomeTrakr metadata templates (BioSample and SRA) V.11

 In 1 collection

DOI

<https://dx.doi.org/10.17504/protocols.io.eq2ly3x1pgx9/v11>

Maria Balkey¹, Ruth Timme¹, Candace Bias¹, Errol Strain¹, Tina Lusk Pfefer¹

¹US Food and Drug Administration

GenomeTrakr

Tech. support email: genomeTrakr@fda.hhs.gov



Ruth Timme

US Food and Drug Administration



Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.eq2ly3x1pgx9/v11>

Protocol Citation: Maria Balkey, Ruth Timme, Candace Bias, Errol Strain, Tina Lusk Pfefer 2024. Guidance for populating and validating GenomeTrakr metadata templates (BioSample and SRA). **protocols.io**

<https://dx.doi.org/10.17504/protocols.io.eq2ly3x1pgx9/v11>Version created by **Ruth Timme**

Manuscript citation:

Timme, R.E., Wolfgang, W.J., Balkey, M. et al. Optimizing open data to support one health: best practices to ensure interoperability of genomic data from bacterial pathogens. One Health Outlook 2, 20 (2020). <https://doi.org/10.1186/s42522-020-00026-3>

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

Protocol status: Working

We use this protocol and it's working

Created: January 12, 2024

Last Modified: February 27, 2024

Protocol Integer ID: 93402

Keywords: GenomeTrakr, metadata, Pathogen package, NCBI Pathogen Detection, INSDC, validating genometrakr metadata template, genometrakr metadata template, health enteric biosample package for all submission, pathogen biosample template, health enteric biosample package, extended pathogen biosample template, biosample metadata template, biosample metadata, extended pathogen package, pathogen package, draft one health enteric package, sequencing data submission, genometrakr, picklists in the genometrakr, health enteric package, health enteric package custom template, available one health enteric package custom template, enteric bacterial pathogen, bacterial pathogen, metadata template, sra metadata, metadata templates necessary for direct submission, national center for biotechnology information, pathogen, biotechnology information, biosample, sequencing, genome, level metadata template, guidelines for obtaining, guideline

Disclaimer

Please note that this protocol is public domain, which supersedes the CC-BY license default used by protocols.io.



Abstract

PURPOSE: This protocol provides instructions for preparing and filling out the metadata templates necessary for direct submission to the National Center for Biotechnology Information (NCBI). These instructions are relevant for the majority of whole genome sequencing data submissions derived from enteric bacterial pathogens collected for surveillance purposes.

SCOPE: This protocol provides detailed instructions for the following two metadata templates:

1. **BioSample metadata:** guidelines for obtaining, populating, and validating the BioSample metadata template.
2. **SRA metadata:** Guidelines for populating sequence-level metadata template.

Version history:

v11: Change of guidance to use the One Health Enteric BioSample package for all submissions.

v10: updates to the GenomeTrakr-extended pathogen biosample template (**GT-pathogen package-OHE v0.3.xlsx**) and release of newly available One Health Enteric package custom templates.

v9: Bug fix

v8: Updated the picklists in the GenomeTrakr-extended pathogen package, "**GT-pathogen package-OHE v0.2.2.xlsx**". Also provided a direct link to the newly published One Health Enteric package.

v7: Updated the picklists in the GenomeTrakr-extended pathogen package, "**GT-pathogen package-OHE v0.2.2.xlsx**" and added an incremental update file for the **DRAFT One Health Enteric Package** that includes extensive edits compared to v6.

v6: Added the One Health Enteric package presented at IAFP 2021 meeting.

Materials

Gather the following contextual information for each pure culture isolate:

1. organism name
2. lab name that collected the sample
3. collection date
4. collection source
5. Geographic location of sample collection

Before start

Before collecting sequence data for your isolates, ensure that you can provide the minimum metadata recommended by your coordinating surveillance body.

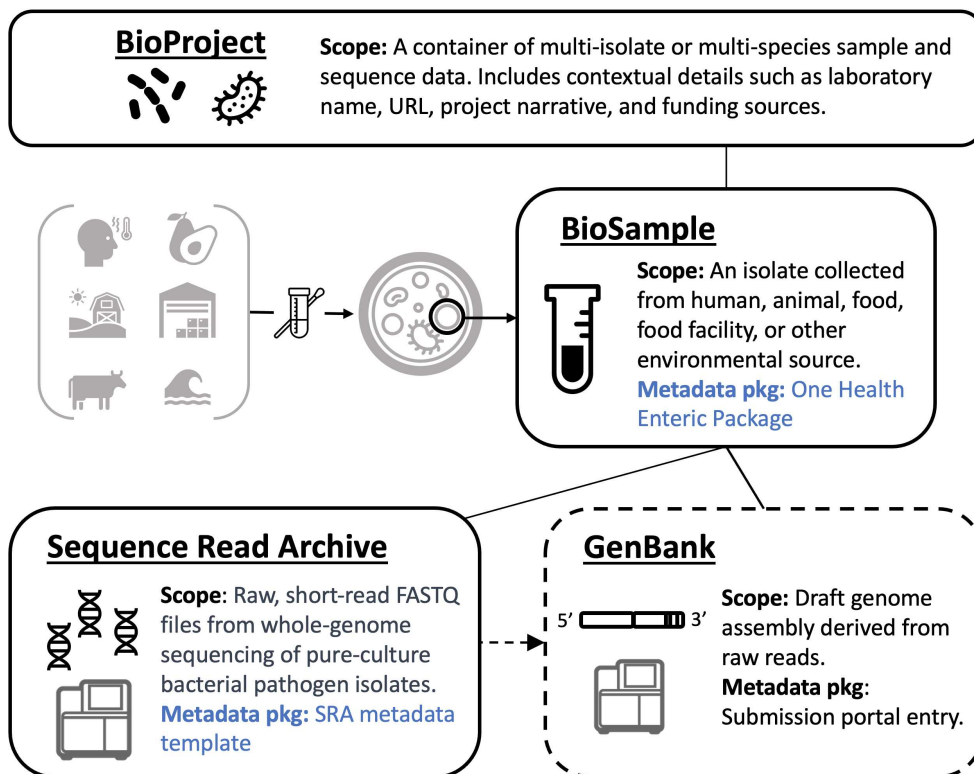
Overview

1

This protocol provides instructions on acquiring and completing two distinct metadata templates essential for the submission of enteric bacterial pathogen surveillance data to the National Center for Biotechnology Information (NCBI).

Two metadata templates are required for each NCBI submission:

1. **BioSample:** metadata describing the isolate, sample collected, and submitting lab information.
2. **SRA:** metadata describing the sequence data collection



BioSample metadata

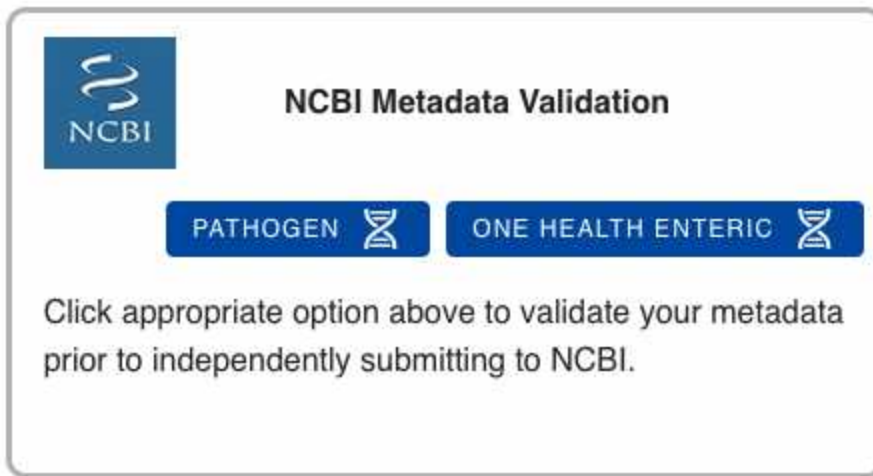
2 **Templates for BioSample submission:**

Visit GenomeTrakr Metadata Validation System (GMVS) at <https://gmvs.fda.gov/> to download custom, version-controlled, biosample metadata template(s). Current and

previous versions of these templates can also be at the [OHE GitHub page](#).

- Our custom templates include extensive guidance and controlled vocabularies for most attributes in the package.
- Sub-packages are available for download covering the major One Health samples types (human/animal hosts, food, food facilities, and farm/environment). Users can choose to populate the full package, or one more more of the sub-packages.

When visiting GMVS, click on the ONE HEALTH ENTERIC icon within the NCBI Metadata Validation box.



Follow GMVS instructions to download BioSample metadata template (click on the cloud download icon). Chose the most appropriate template for your sample types (the full package or one of the sub-packages).



One Health Enteric Metadata Sheet Upload

First select the One Health Enteric package that you will be validating. GMVS supports validating either the full One Health Enteric package or a sub-package by selecting the appropriate radio button below. Then click the paperclip icon to browse and attach the completed One Health Enteric metadata sample sheet. Click the **Validate** button after attaching. If needed, click the  button next to the radio button label to download the latest template for that package. The latest version number is included in the radio button label as well. If you currently have a template, you can find the version number at the top of the Instructions tab.

- ☒ One Health Enteric Full Package (1.3  
- ☐ One Health Enteric Food Group Package (1.3  
- ☐ One Health Enteric Food Facility Package (1.3  
- ☐ One Health Enteric Farm Environment Package (1.3  
- ☐ One Health Enteric Human Animal Host Package (1.3  



Attach the One Health Enteric metadata sample spreadsheet

Search/Select Lab Name

Search and Select Lab Name Conducting Validation.

 VALIDATE

 CANCEL

One Health Enteric Metadata Sheet Upload

2.1 Review the excel -Instructions- sheet within the OHE excel file.

This sheet is the instruction sheet to help outline purpose of each metadata field.				
Version Date: 12/16/2023 10:53:14	Version Number: 1.3	Header Color Category	Required	Description
Column	Field Name (Mandatory)	Format		
A	sample_name	Free Text	Mandatory	Sample Name is another unique identifier for the pure culture isolate and required by NCBI for BioSample submission (it cannot be left blank). It can have any format, but we suggest that it be the same as the strain name or contain another identifier important to the isolate or submitting laboratory. NCBI validates this attribute for uniqueness, so you cannot use "missing" or "not collected". This identifier is NOT available in NCBI-PD. Example: MDQ-185-12-KY-M5619-200619.
B	bioproject_accession	Free Text	Mandatory	The accession number of the BioProject(s) to which the BioSample belongs (P/N/A/xxxxxx). This cannot be left blank. Double check that you are submitting to the correct BioProject (the organism name must match the one designated for your BioProject). For species that fall outside of NCBI pathogen detection, we recommend establishing a separate multi-species "research" bioproject for publishing data outside of the structured Pathogen Detection surveillance effort. Example: PRJNA578395.
C	isolate_name_alian	Free Text	Optional	Other IDs associated with this isolate or strain. Separate with ';' if more than one. Example: ABC123;StanUd567.
D	strain	Free Text	Mandatory	This is the authoritative ID used for foodborne pathogen genomic epidemiology and within NCBI Pathogen Detection. Although the strain ID can have any format, we suggest that it be unique, concise, and consistent within your laboratory (e.g. CFSAN123456).
E	culture_collection	Free Text	Optional	Name of source institute and unique culture identifier. See the description for the proper format and list of allowed institutes, http://www.inrdc.org/en/controlled-vocabulary/culturecollection-qualifier . Example: ATCC-BAA-464.
F	reference_material	Dropdown	Optional	Describe a laboratory reference or control strain. Leave blank if not applicable. Example: proficiency testing isolate.
The organism name should include the most descriptive information you have at time of submission, adhering to proper nomenclature in NCBI taxonomy database: https://www.ncbi.nlm.nih.gov/taxonomy/Browser/iewtax.cgi . Check spelling carefully! Levels of valid organism names are as follows: Genus species: Salmonella enterica Listeria monocytogenes Genus species and subspecies: Salmonella enterica subsp. enterica Determined serotype or serovar (trad or WGS-based): Escherichia coli O104:H7 Salmonella enterica subsp. enterica serovar Agona Salmonella enterica subsp. diarizonae serovar 38-103b/A,N,X,15 Listeria monocytogenes serotype 1/2a If NCBI doesn't have the desired organism name, enter the name determined by your laboratory. After submission, a "taxonomy consult" will take place to evaluate the new name. Sometimes the organism name is changed to a canonical server name and the submission proceeds. It is also possible that the server is a novel one not currently in the NCBI database and the Taxonomy team will work with the submitter to get the new name added to the database.				
G	organism	Free Text	Mandatory	

Instructions Sheet within OHE excel file

Proceed to fill out the BioSample metadata template in the -UserEntry- excel sheet. Where possible, use terms from dropdown menus for each metadata attribute.

reference_material	organism *	collected_by *	collection_date *	cult_isol_date	geo_loc_name *	isolation_source *	source_type *
					USA:NC		
					USA:ND		
					USA:OH		
					USA:OK		
					USA:OR		
					USA:PA		
					USA:RI		
					USA:SC		
					USA:SD		
					USA:TN		
					USA:TX		
					USA:UT		

User Entry Sheet within the OHE excel file.

Validate BioSample metadata template

2.2 Upload the completed OHE metadata template to GMVS and click on -VALIDATE- icon.

One Health Enteric Metadata Sheet Upload

First select the One Health Enteric package that you will be validating. GMVS supports validating either the full One Health Enteric package or a sub-package by selecting the appropriate radio button below. Then click the paperclip icon to browse and attach the completed One Health Enteric metadata sample sheet. Click the **Validate** button after attaching. If needed, click the  button next to the radio button label to download the latest template for that package. The latest version number is included in the radio button label as well. If you currently have a template, you can find the version number at the top of the Instructions tab.

- ☒ One Health Enteric Full Package (1.3  
- ☐ One Health Enteric Food Group Package (1.3  
- ☐ One Health Enteric Food Facility Package (1.3  
- ☐ One Health Enteric Farm Environment Package (1.3  
- ☐ One Health Enteric Human Animal Host Package (1.3  



Attach the One Health Enteric metadata sample spreadsheet



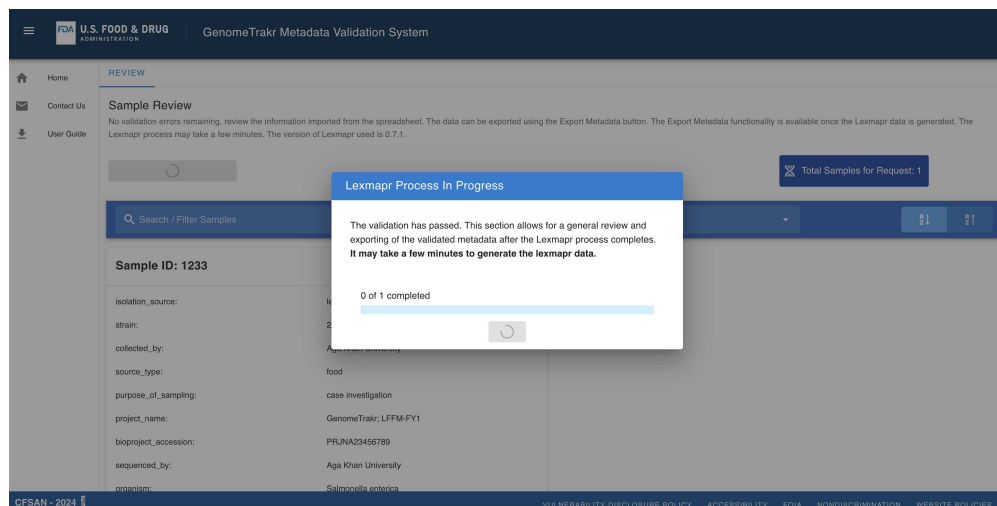

 OneHealthEntericMetadata(1).xlsx (214.0 kB)

United States Food and Drug Administration | Center for Food Safety and Applied Nutrition

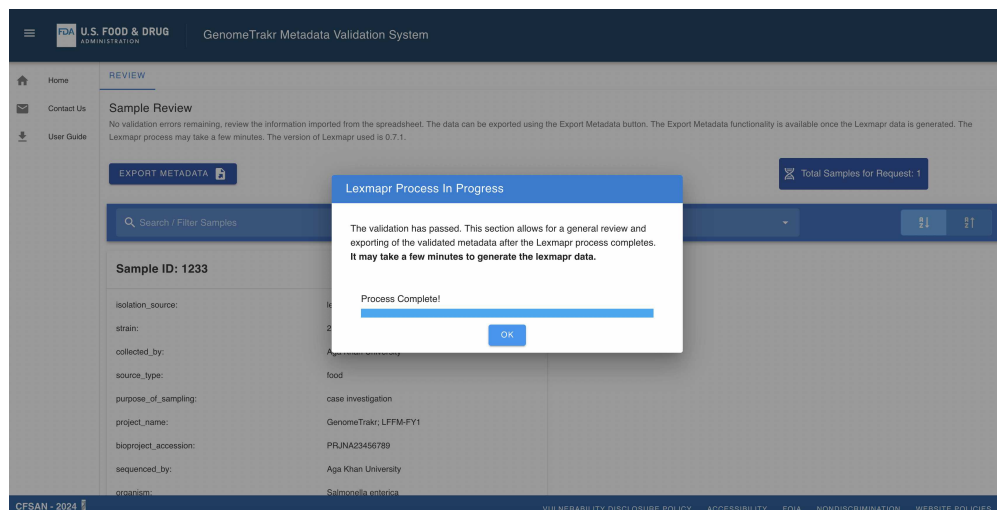
Search and Select Lab Name Conducting Validation.

 **VALIDATE**  **CANCEL**

The GMVS validation system will check each entry and also run **LexMapr** for auto-assignment of the IFSAC category.



After completion GMVS will report out results of the validation.



Click -OK-.

2.3 No validation errors:

If metadata passes GMVS validation, each record will be displayed with all the metadata and you will have an option to export metadata.

REVIEW

Sample Review

No validation errors remaining, review the information imported from the spreadsheet. The data can be exported using the Export Metadata button. The Export Metadata functionality is available once the Lexmapr data is generated. The Lexmapr process may take a few minutes. The version of Lexmapr used is 0.7.1.

EXPORT METADATA

Total Samples for Request: 1

Search / Filter Samples

Sample ID: 1233

isolation_source:	lettuce
strain:	2331231
collected_by:	Aga Khan University
source_type:	food
purpose_of_sampling:	case investigation
project_name:	GenomeTrakr; LFFM-FY1
bioproject_accession:	PRJNA23456789
sequenced_by:	Aga Khan University
organism:	Salmonella enterica

CFSAN - 2024 [VULNERABILITY DISCLOSURE POLICY](#) [ACCESSIBILITY](#) [FOIA](#) [NONDISCRIMINATION](#) [WEBSITE POLICIES](#)

Click on the -EXPORT METADATA- icon.

Review validated BioSample metadata and lexmapr output (cleaned up isolation_source entries and proposed IFSAC_category). ➡ go to step #3 for reviewing lexmapr output.

2.4 Address validation errors:

If there are validation errors, GMVS will generate a log report. If few errors are reported, edit values by clicking the EDIT icon, otherwise, export reviewed template by clicking -EXPORT ERRORS-.

Make required changes and click -RE-IMPORT SHEET- and proceed to re-validate the template.

U.S. FOOD & DRUG ADMINISTRATION **GenomeTrakr Metadata Validation System**

VALIDATION ERRORS

Sample Sheet Validation Errors

The validation has failed. Validation errors are listed below. Click "Edit" to update the data to meet the field's requirements or fix the errors in your spreadsheet and re-import.

EXPORT ERRORS **RE-IMPORT SHEET**

Validation Errors Remaining: 1

Row	Column Letter	Property Name	Value Entered	Error Message	Action
2	G	organism		organism is Mandatory	EDIT

isolation_source: lettuce strain: 2331231 collected_by: Aga Khan University source_type: food purpose_of_sampling: case investigation project_name: GenomeTrakr; LFFM-FY1 bioproject_accession: PRJNA23456789 sequenced_by: Aga Khan University organism: null collection_date: 2023 sample_name: 1233 geo_loc_name: Argentina.Chaco

Evaluation of LexMapr Output

- 3 LexMapr is a tool that processes free text from **isolation_source** and generates standard terminologies from controlled vocabulary/ontologies, including FoodOn, GenEpiO, UBERON, ENVO, NCBI Taxon, and specific food and environmental categories from Interagency Food Safety Analytics Collaboration (IFSAC) controlled vocabulary.

LexMapr: <https://github.com/Public-Health-Bioinformatics/LexMapr>.

Each GMVS record subject to validation is analyzed with LexMapr, the attribute isolation source gets an ontological descriptor and a category from IFSAC+ terminology for food safety. After records are processed with LexMapr, a report is generated

The LexMapr report generated at GMVS contains the following columns: strain, isolation_source, isolation_source (LexMapr generated), and IFSAC_category.

	strain	isolation_source	isolation_source (Lexmapr generated)	IFSAC_category
	FDA189213 897_s001	ENV swab sponge	environmental swab sponge	environmental- factory/production facility

LexMapr Output generated during validation.

Review the Lexmapr generated recommendations for isolation_source and IFSAC_category. If you agree with the recommendations, copy these the contents of these fields into the validated BioSample metadata template, under the isolation_source and IFSAC_category fields, respectively.

If the IFSAC category(s) recommended for the sample type are incorrect or *not* appropriate, leave that entry blank for the submission and submit a bug report to genometrakr@fda.hhs.gov.

Save the validated biosample metadata template and proceed with NCBI submissions.

SRA sequence metadata template

- 4 **Template for SRA metadata submission:**

Download the generic "**Metadata spreadsheet with sample names**" file from the NCBI Submission Templates page:

<https://submit.ncbi.nlm.nih.gov/templates/>

And follow the guidance in the following table:

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.

4.1

A	B	C
Field	Description	Example
sample_name	Include the same ID here as you entered for "sample_name" in the BioSample submission template.	UT-12345
library_ID	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.	UT-12345.6
Title	Short, free text description that identifies the data on public pages. For Example: {methodology} of {organism}: {sample_name}	WGS of Salmonella enterica: UT-12345
library_strategy	Overall sequencing strategy or approach. Choose from NCBI pick list	WGS
library_source	molecule type used to make the library	genomic
library_selection	Library capture method	random
Library_layout	Choose from NCBI pick list	paired
platform	Sequencing platform	Illumina
instrument_model	Name of the sequencing instrument.	MiSeq
Design_description	Free text description of methods	



A	B	C
Filetype	File format name for the raw sequence data Choose from NCBI pick list	Fastq
Filename	include ALL of the files resulting from this library. **Add additional fields if there are more than two files (e.g. Filename3).	genome_r1.fastq (*must be exact)
Filename2	genome_r2.fastq (*must be exact)	genome_r2.fastq (*must be exact)
Filename3-8	list other fastq file names (e.g. for NextSeq data)	

SRA metadata data template guidance and examples for WGS submission.

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.