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

In 5 collections

Maria Balkey[1], Ruth Timme[1], Candace Hope Bias[1], Errol Strain[1], Tina Lusk Pfefer[1]

[1]US Food and Drug Administration

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

Ruth Timme
US Food and Drug Administration

VERSION 11

FEB 26, 2024

OPEN ACCESS

DISCLAIMER

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

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

Protocol Citation: Maria Balkey, Ruth Timme, Candace Hope 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/v11Version 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

Protocol status: Working
We use this protocol and it's working

Created: Jan 11, 2024

Last Modified: Feb 26, 2024

PROTOCOL integer ID: 93402

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

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 INSTRUCTIONS

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

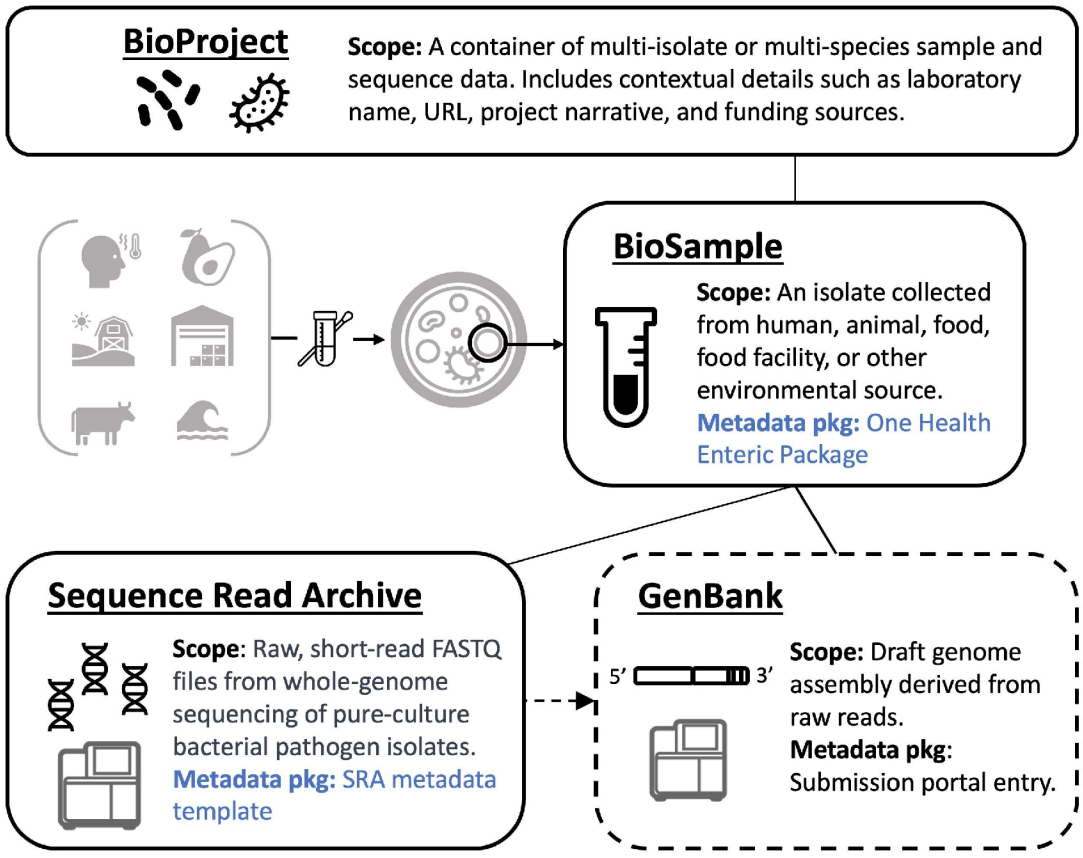
Overview

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



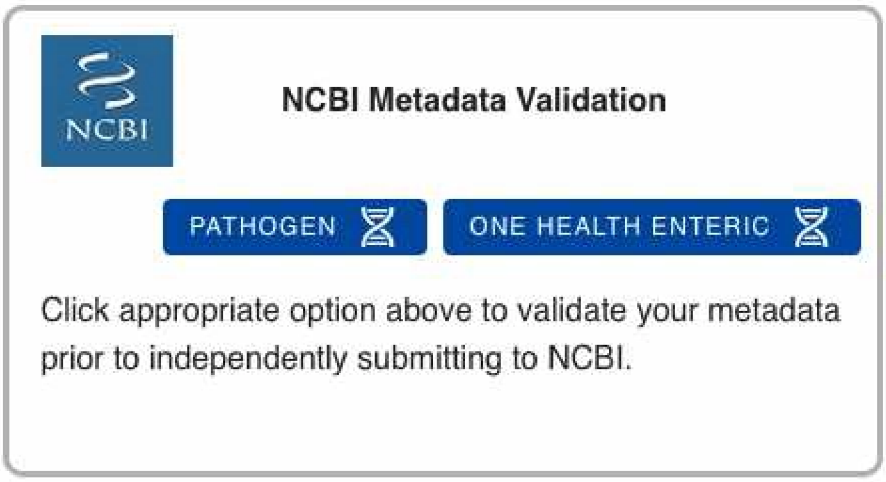
BioSample metadata

2 **Templates for BioSample submission:**

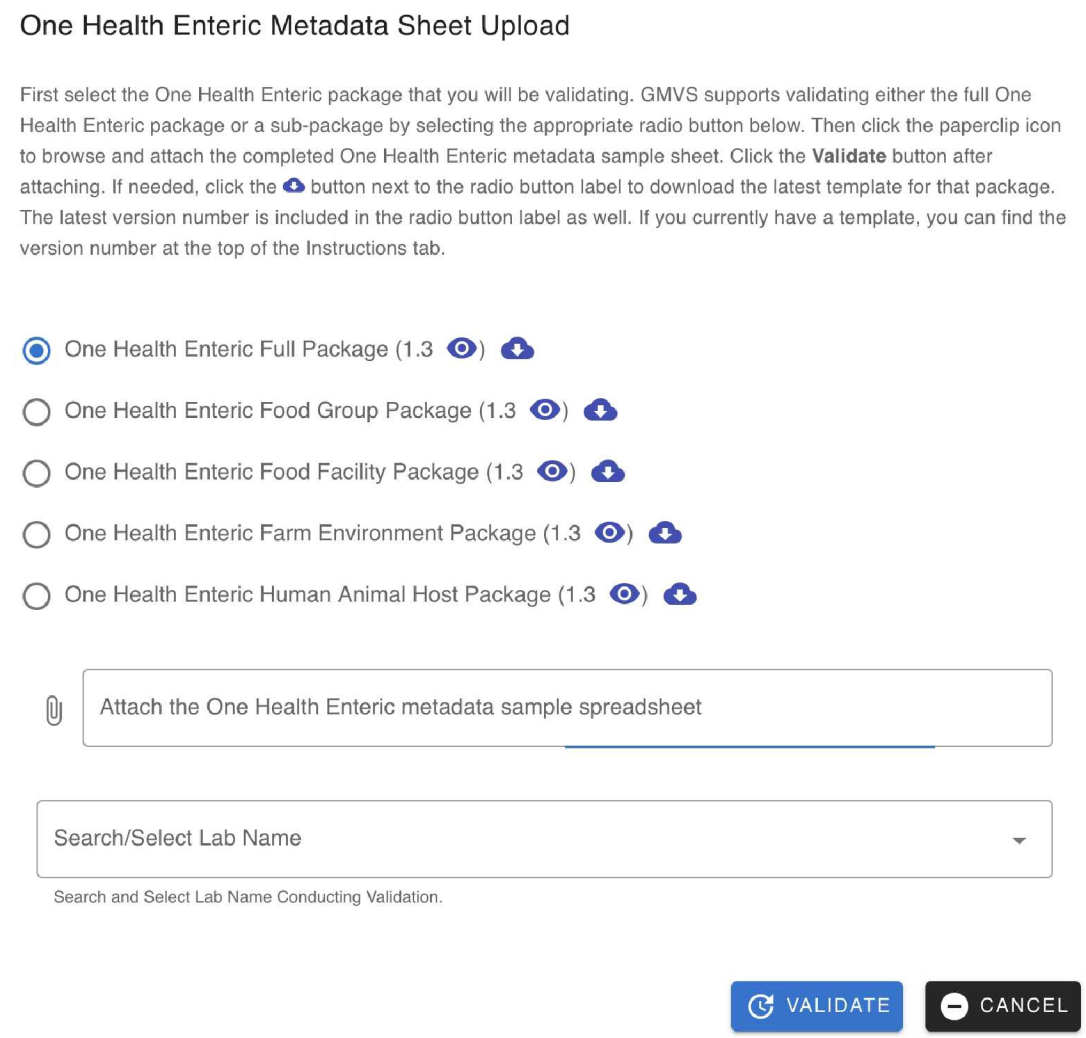
Visit GenomeTrakr Metadata Validation System (GMVS) at https://gmvs.fda.gov/ to download custom, version-controlled, biosample metadata template(s). Current and previous versions of these templates can also be at the **OHE GitHub page**.

- 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.

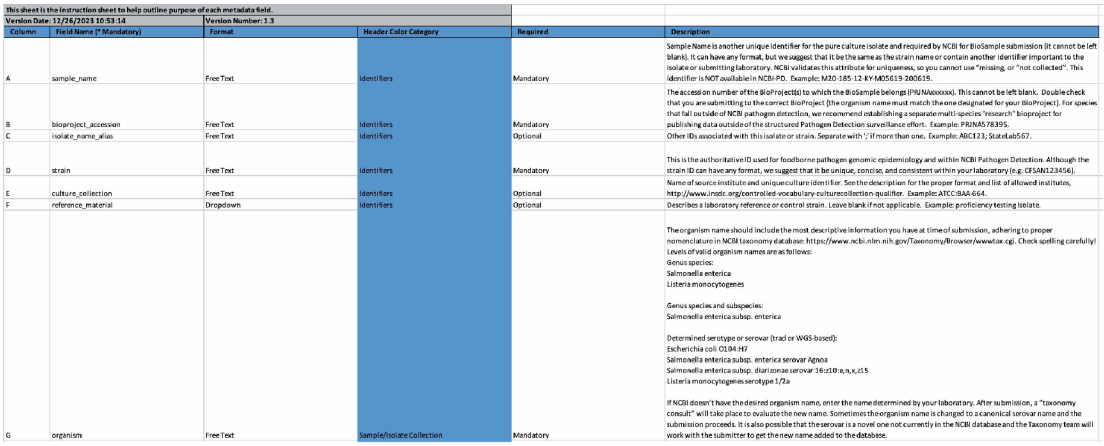


Follow GMVS instructions to download BioSample metadata template (click on the cloud download icon). Chose the most appropriate template for your sample types (the full package or one of the sub-packages).



One Health Enteric Metadata Sheet Upload

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: 12/26/2023 10:53:14 | Version Number: 1.3

Column	Field Name (* Mandatory)	Format	Header Color Category	Required	Description
A	sample_name	Free Text	Identifiers	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: M20-185-12-KY-M05619-280619.
B	bioproject_accession	Free Text	Identifiers	Mandatory	The accession number of the BioProject(s) to which the BioSample belongs (PRJNAxxxxxx). 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_alias	Free Text	Identifiers	Optional	Other IDs associated with this isolate or strain. Separate with ';' if more than one. Example: ABC123; StateLab567.
D	strain	Free Text	Identifiers	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	Identifiers	Optional	Name of source institute and unique culture identifier. See the description for the proper format and list of allowed institutes, http://www.insdc.org/controlled-vocabulary-culturecollection-qualifier. Example: ATCC:BAA-664.
F	reference_material	Dropdown	Identifiers	Optional	Describes a laboratory reference or control strain. Leave blank if not applicable. Example: proficiency testing isolate.
G	organism	Free Text	Sample/Isolate Collection	Mandatory	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/wwwtax.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 16:z10:e,n,x,z15 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 serovar name and the submission proceeds. It is also possible that the serovar 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.

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.

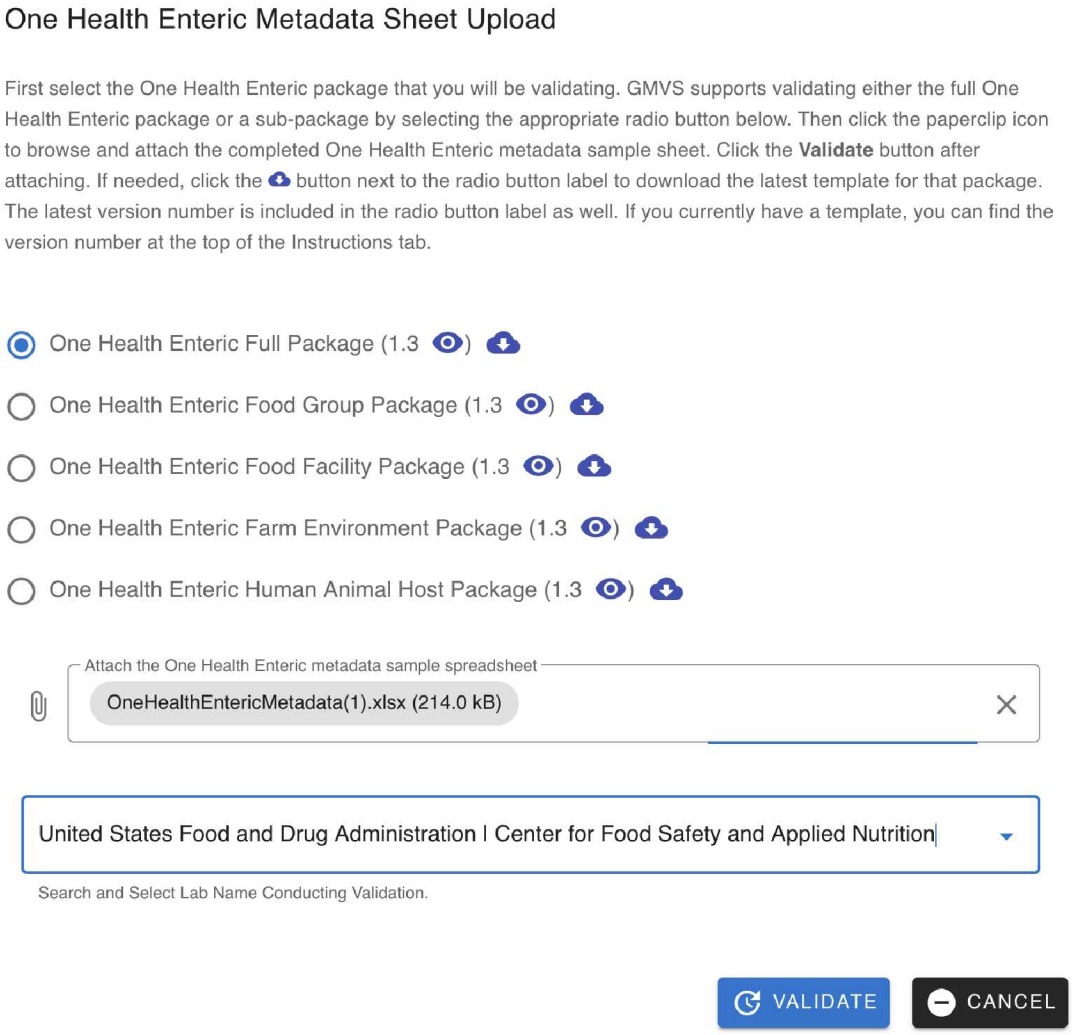
eference_material	organism *	collected_by *	collection_date *	cult_isol_date	geo_loc_name *	isolation_source *	source_type *

USA:NC
USA:ND
USA:OH
USA:OK
USA:OR
USA:PA
USA:RI
USA:SC
USA:SD
USA:TN
USA:TX
USA:UT

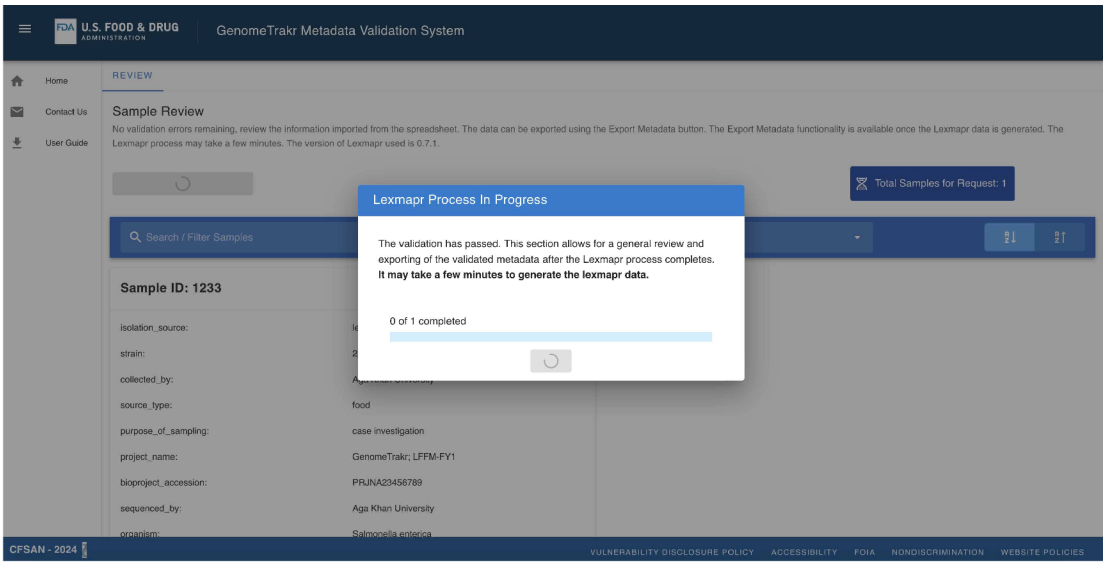
User Entry Sheet within the OHE excel file.

Validate BioSample metadata template

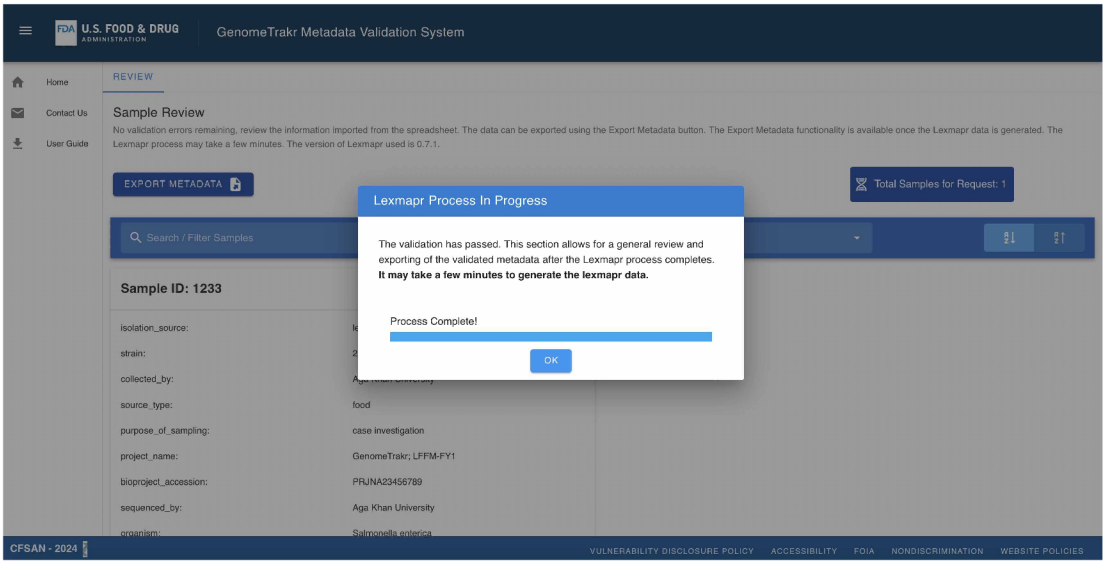
2.2 Upload the completed OHE metadata template to GMVS and click on -VALIDATE- icon.



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



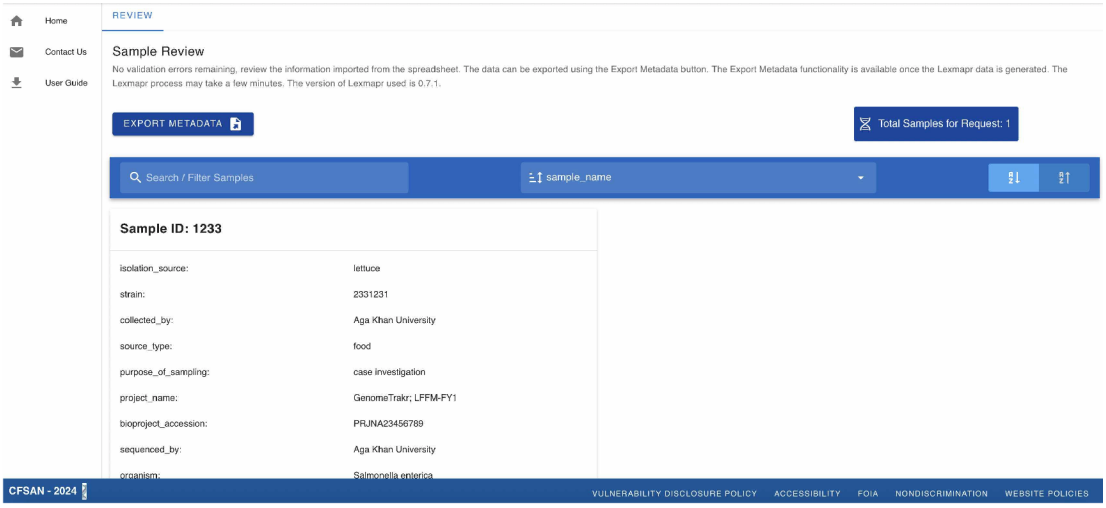
After completion GMVS will report out results of the validation.



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.



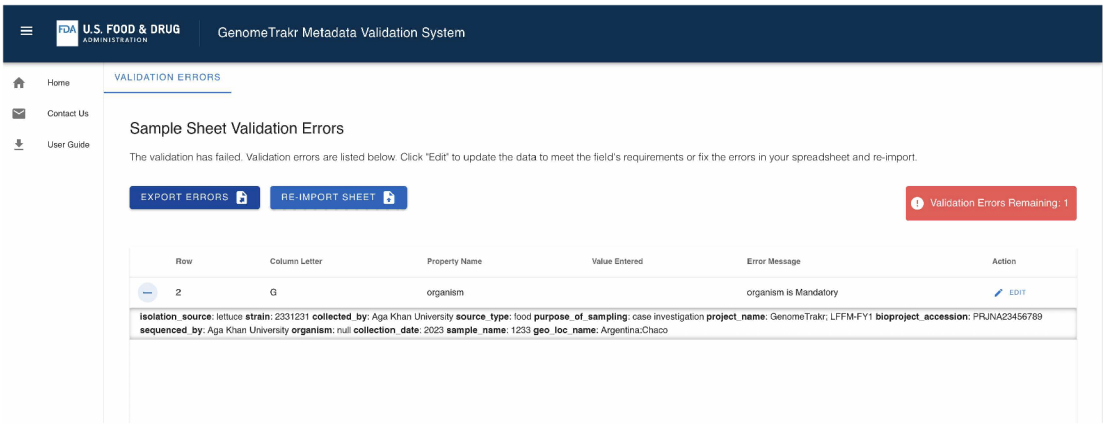
Click on the -EXPORT METADATA- icon.

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.



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
	FDA189213897_s001	ENV swab sponge	environmental swab sponge	environmental-factory/production facility

LexMapr Output generated during validation.

Review the Lexmpr generated recommendations for isolation_source and IFSAC_category. If you agree with the recomendations, 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	
	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.