

Oct 16, 2024

Version 2

GenFS Metadata Cleanup Challenge protocol V.2

DOI

<https://dx.doi.org/10.17504/protocols.io.rm7vzj6prlx1/v2>



Ruth Timme¹, Martin Shumway², Candace Bias³, Maria Balkey³, Tina Pfefer⁴

¹US Food and Drug Administration; ²National Center for Biotechnology Information; ³US FDA-CFSAN; ⁴US FDA

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.rm7vzj6prlx1/v2>

Protocol Citation: Ruth Timme, Martin Shumway, Candace Bias, Maria Balkey, Tina Pfefer 2024. GenFS Metadata Cleanup Challenge protocol. protocols.io <https://dx.doi.org/10.17504/protocols.io.rm7vzj6prlx1/v2> Version created by **Ruth Timme**



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: October 16, 2024

Last Modified: October 16, 2024

Protocol Integer ID: 109934

Keywords: ncbi for the genfs metadata cleanup challenge exercise, genfs metadata cleanup challenge protocol this protocol, genfs metadata cleanup challenge protocol, genfs metadata cleanup challenge exercise, metadata cleanup challenge, submitting bulk biosample, bulk biosample, ncbi, finalized template

Abstract

This protocol provides guidance for submitting bulk BioSample updates to NCBI for the GenFS Metadata Cleanup Challenge exercise.

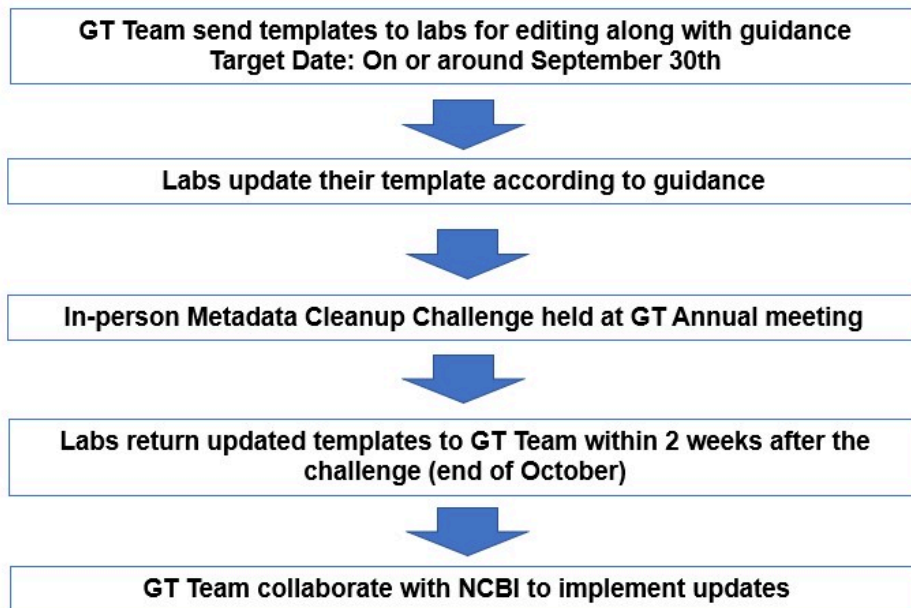
Table of Contents:

- Timeline/Workflow for the 2024 Metadata Cleanup Challenge
- Preparing the update file
- Making corrections and updates within the file
- Submitting the finalized template



Timeline/Workflow for the 2024 Metadata Cleanup Challenge

1



2 Overview and scope of exercise

Timeframe: focus on submissions between September 2023-August 2024. Your lab is welcome to include submission outside this timeframe.

Scope: review and curate entries for the following attributes, which are core requirements within the One Health Enteric package:

Focus on these attributes:

- source_type
- collected_by
- sequenced_by
- project_name
- host (for human and animal isolates)
- food_origin (for food isolates)
- isolation_source



Guidance for preparing the update file

3 Steps 4 and 5 describe the required format for the bulk update file.

4 **For labs in the GenomeTrakr network:**

We have generated a bulk metadata update template for every laboratory that has a BioProject linked to the GenomeTrakr umbrella BioProject at NCBI ([PRJNA593772](https://www.ncbi.nlm.nih.gov/BioProject/PRJNA593772)).

Pick up your metadata template here: [DOCS: metadata hackathon/2024 Hackathon Template Pickup](#).

Instructions for generating your own template (if your lab's BioProject is not linked to the GenomeTrakr Umbrella):

Navigate to SRA Run Selector: <https://www.ncbi.nlm.nih.gov/Traces/study/>

Enter your BioProject accession(s), then click the "Metadata" button in the download box.

SRA Run Selector

Accession: PRJNA692474 Search

Filters List

- 1 ☐ attribute_package
- 2 ☐ AvgSpotLen
- 3 ☐ Bases
- 4 ☐ BioProject
- 5 ☐ BioSampleModel
- 6 ☐ Bytes
- 7 ☐ Center Name
- 8 ☐ collected_by
- 9 ☐ Genus
- 10 ☐ geo_loc_name_country
- 11 ☐ geo_loc_name_country_co
- 12 ☐ Host
- 13 ☐ Host_disease
- 14 ☐ IFSAC_Category
- 15 ☐ IFSAC+_Category
- 16 ☐ Instrument

Common Fields

Consent	PUBLIC
Assay Type	WGS
DATASTORE filetype	FASTQ, RUN.ZQ, SRA
DATASTORE provider	GS, NCBI, S3
DATASTORE region	gs.us-east1, ncbi.public, s3.us-east-1
LibraryLayout	PAIRED
LibrarySource	GENOMIC
Platform	ILLUMINA
version	1

Select

	Runs	Bytes	Bases	Download
Total	241	66.33 Gb	136.36 Gb	Metadata or Accession List
Selected	0	0	0	Metadata or Accession List

The download metadata files will be a comma-delimited txt file with both the BioSample + SRA metadata. Open this in Excel. Remove all the metadata columns not relevant for this exercise, move the BioSample accessions to the first column, then proceed.

See the example template for guidance on which columns to keep:



Example Template.xlsx

5 Prepare the update file:

- Updates should be put into one bulk update file per lab, or per lab/organism, or some other aggregation.
- **Rows:** The first row, or header, must contain the column names. Each subsequent row must contain exactly one BioSample and its updatable fields, or attributes. Each row must contain exactly the same number of columns.
- **Columns:** The column names should be attribute names (**full name or harmonized name**) included in the One Health Enteric package. The first column should contain the BioSample accession (eg SAMN123456789)
- **File format:** tab-delimited text file using the file suffix .tsv or .txt. If you use Microsoft Excel for editing, export the final template into this format. Take care that dates are correctly exported from the spreadsheet.

FYI:

- Only those columns where at least one record has a replacement value need to be supplied.
- If a field is blank in the update file, and has an actual value in NCBI BioSample, that will be taken as "replace current value with NULL".
- If a field in a record already has a value in NCBI BioSample that will not be changed, then the field should be filled in with its current value.
- If a field is new to the BioSample, it will be added. A blank field whose attribute does not currently exist in the BioSample record will not be added.

5.1 What fields/attributes can be included in the bulk update file?

- collected_by
- geo_loc_name
- food_origin
- host
- project_name
- sequenced_by
- source_type
- isolation_source
- collection_date
- purpose_of_sampling
- env_local_scale
- env_medium
- animal_env

- env_broad_scale
- facility_type
- intended_consumer
- food_type_processed
- food_processing_method

5.2 What fields **CANNOT** be updated in the bulk update file?

DO NOT include changes to the following attributes in this bulk update file. Changes to the following fields can be requested separately by writing to **pd-help@ncbi.nlm.gov** directly. Changes to the following attributes affect multiple resources at NCBI and updates to them need to be coordinated.

- **biosample_acc/BioSample** – This is the primary key of the entire BioSample system - Keep this accession in the first column, but DO NOT EDIT OR UPDATE ENTRIES.
- **bioproject_acc/BioProject** - changes to linked bioProject, linked SRA, or linked Assembly.
- **strain** or **isolate**
- **sample_name**
- **attribute package** – This requires validation of all the fields together.
- **center name** – This is a property of SRA and cannot be changed using the bulk update channel.
- **organism** - Change to the BioSample species (the identification of the isolate).
- **Salmonella serovar/serotype** - changes to organism species or sub-species names (ex Salmonella enterica ⇒ Salmonella enterica subsp. enterica serovar Infantis)
- Any kind of SRA to BioSample, BioSample to BioProject, or SRA to BioProject mapping.
- Any of the **computed fields** in Pathogen Detection (epi_type, min_same, min_diff, computed_types, or amr/virulence analysis outputs). These attributes cannot be updated by the record owner.

For updating any of these attributes, send a TSV file of proposed changes to **pd-help@ncbi.nlm.gov** as prepared above.

Guidance for making corrections and updates within the file

- 6 **Steps 7-9:** These steps cover required fields. In your Excel template, ensure there's an entry in these columns for every row. If information is missing, then choose one of the null terms, "Not Applicable, Not Collected, Not Provided, Missing, or Restricted Access".

Guidance for making corrections and updates within the file

- 7 **source_type**



Review your entries, ensure every record has an `source_type` entry. Make corrections where needed.

human
animal
food
environmental
other
Not Applicable
Not Collected
Not Provided
Missing
Restricted Access

NCBI and US pathogen surveillance coordinators have a *strict* controlled vocabulary for this attribute, only the above terms are allowed.

8 **project_name**

Review the entries in this column, make corrections where needed and populate all missing fields. For US surveillance, please choose the coordinating body that best the isolate.

GenomeTrakr
GenomeTrakr; LFFM-FY1
GenomeTrakr; LFFM-FY2
GenomeTrakr; LFFM-FY3
GenomeTrakr; LFFM-FY4
GenomeTrakr; LFFM-FY5
NARMS
NARMS Cecal
NARMS Retail Meat
PulseNet
USDA-FSIS
Vet-LIRN
NAHLN
USDA-ARS

If you would like to create another term to communicate membership in another project or network, feel free to do that! Enter you new term directly into the update template. If you would like this term added to the picklist, send this term to **genomeTrakr@fda.hhs.gov**, and we'll add it to the next update.



Include more than one term? Separate with "; ", for example, "GenomeTrakr; USDA-ARS".

9 **collected_by** and **sequenced_by**

Review the entries in both of these columns, make corrections where needed, and populate any missing fields.

Ensure terms are standardized across all your records. Check the **sequenced_by** picklist in the current One Health Enteric package file for your standardized laboratory name.

Send updates, corrections, or additions to your laboratory name to genomeTrakr@fda.hhs.gov and we'll make corresponding updates to the picklist terms.

Guidance for making corrections and updates within the file

- 10 **Steps 11-13:** For this exercise we're focused on two conditionally required attributes:
1. **host** should be populated for isolates derived from human or animal samples
 2. **food_origin** should be populated for isolates derived from food products, or other commercial products sampled for pathogens (medical products, cosmetics, tattoo ink, etc).

Use the **source_type** column to filter for these sample types, first for animal/human samples (step 11 + 12), then for food samples (step 13).

Guidance for making corrections and updates within the file

11 **host**

FOR HUMAN and ANIMAL ISOLATES ONLY

Host is a required for host associated human and animal isolates.

Use the **source_type** column to identify human and animal isolates (filter in Excel for human and animal).

Use the entries in the **isolation_source** column to help determine what the **host** entry should be.

Where possible, enter a scientific or binomial name, for example *Homo sapiens* or *Bos taurus*. If scientific name is unknown, use a common name recognized by the [NCBI taxonomy database](#) (porcine, bovine, etc).

When **host** is completely unknown, provide one of NCBI's null values:

Not Applicable
Not Collected
Not Provided
Missing
Restricted Access

Guidance for making corrections and updates within the file

12 isolation_source



****For human and animal isolates: remove all taxonomic references included in isolation_source.****

This information should solely reside in **host**.

Example edit to both **isolation_source** and **host**:

Pathogen: environmental/food/other sample from Salmonella enterica

Identifiers	BioSample: SAMN07774651; SRA: SRS3721856; CFSAN: CFSAN070305
Organism	Salmonella enterica cellular organisms; Bacteria; Pseudomonadota; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Salmonella
Package	Pathogen: environmental/food/other; version 1.0
Attributes	Interagency Food Safety Analytics Collaboration (IFSAC) category veterinary clinical/research cow strain CFSAN070305 source type animal geographic location Nigeria collected by University of Ibadan collection date 2012 isolate name alias DE 52 latitude and longitude missing isolation source stool (<u>Bos taurus</u>) project name GenomeTrakr sequenced by FDA Center for Food Safety and Applied Nutrition ontological term bos taurus:NCBITAXON_9913, feces:UBERON_0001988 attribute_package environmental/food/other PublicAccession CFSAN070305 ProjectAccession PRJNA186035 Species enterica Genus Salmonella Host <u>Bos taurus</u>
BioProject	PRJNA186035 Salmonella Retrieve all samples from this project

13 food_origin

****FOR FOOD OR OTHER PRODUCTS ONLY****

Sort or filter on the **source_type** column to identify the **"food"** isolates.



Food and other products have two attributes describing geographic location information:

geo_loc_name: geographic location where the sample was physically collected

food_origin: geographic *origin* of food product or other product sampled for pathogens

Check location data in **geo_loc_name**:

- If the location information in **geo_loc_name** reflects the *state or country of origin* for the food product (e.g. "India", reflecting an imported food product sampled in the US), move this geographic information to **food_origin**.
- If your isolates contain the location where the sample was collected (**e.g. port of entry, US state of grocery store, etc**), leave **geo_loc_name** as is. Populate **food_origin** with the country or state of origin.

If **food_origin** is unknown, provide one of NCBI's null values:

Not Applicable

Not Collected

Not Provided

Missing

Restricted Access

OPTIONAL: bring your records up to OHE standards

- 14 Review and populate the other conditionally required fields for OHE sub-packages. Download [latest version of the One Health Enteric package](#) for picklist terms.

Filter on **source_type** to locate isolates belonging to the four OHE sub-packages. The terms listed under the sub-packages are the conditionally mandatory attributes for those subpackages.

animal samples, source_type = animal
animal_env

food product samples, source_type = food
intended_consumer
food_processing_method

facility inspection samples, source_type = environmental
facility_type
food_type_processed

farm/environment samples, source_type = environmental
env_broad_scale
env_medium



env_local_scale

Finalize the update file and submit the bulk template!

15 Check for duplications - ensure there are no double BioSample entries.

15.1 Bring your template to the October 16th, 2024 GenomeTrakr metadata cleanup challenge - finalize edits at the hackathon.

15.2 Remove columns that you're not updating. Ensure that the BioSample accession is retained in Column 1 and strain is kept for tracking the update.

Keep

Biosample
strain

Remove

package name
Title
Center_Name
Release_Date
Bioproject
Run
sample_name

15.3 Save as a tab-delimited .tsv file and transfer the .tsv file to protocols.io:
DOCS: metadata hackathon/2024 Hackathon Template Submission