# NCBI BioSample Submission Requirements - Table of Contents

# 1. [*Getting Started -* Long Description and Instructions for the Excel Metadata File](#_1.__) 1

This section provides instructions for the Data Entry worksheet, which populates the “master” NCBI Biosample worksheet; Required BioSample fields from the latter (NCBI Biosample) master worksheet will be uploaded to NCBI

[Entering Data on the “DATA ENTRY” worksheet](#_Entering_Data_on)

[Data Entry Worksheet - Field Definitions](#_Data_Entry_Worksheet)

# 2. [Curating Data and Viewing in the “NCBI BIOSAMPLE” Worksheet](#_2.__) 9

This section describes how to make your entries as accurate and descriptive as possible

# 3. [META DATA Short Cheat Sheet](#_3.__) 10

Skip to this section for a summary of the essential requirements

# *Getting Started -* Long Description and Instructions for the Excel Metadata File:

* There are two worksheets of interest to all parties.
* The first is the “**Data Entry**” worksheet. It is used to populate the “**NCBI Biosample**” worksheet
* “Freeze panes” is currently enabled in both the "Data Entry" and the “NCBI Biosample” worksheet to assist you. To change the Freeze panes option, go to the Excel "View" tab and reselect "Freeze panes" to turn off / on as required
* No row filters are currently on in the NCBI Biosample worksheet - should you wish to filter rows in the NCBI Biosample worksheet to identify which strains are yours for example, just place the mouse on the top row, choose "Data" from the Excel tab menu and select the "Filter" function

The expected way to enter data for each isolate is to employ the ““Data Entry” worksheet to manually enter data.

## **Entering Data on the “DATA ENTRY” worksheet:**

*The advantage of this approach is that it pre-populates the NCBI Biosample with the* ***controlled vocabulary terms*** *used to convey the metadata categories*

*Instructions:*

* Start with the **Record** area and select the **Strain identifier** from the drop down menu at the right of the Record row cell
* With the desired isolate selected, proceed to scroll down the spreadsheet with your mouse or used and arrow down or page down function on your computer
* Fill in appropriate terms to your project. Leave blank items that are not relevant
* For windows users, right-click on any cell to select a drop down menu that provides a listing of controlled vocabulary choices
* Upon completion and curation of the data entered, you need to save the entry. To achieve, scroll back up to the top of the DATA ENTRY sheet and depress the big "**SAVE**" button with your mouse - located up beside the Pathogen Isolate Information in row 6.  
  Once the sample is saved, it will populate the information into the NCBI Biosample spreadsheet for that sample.
* Proceed to the next isolate number on your list
* Fields with an asterisk (\*) are mandatory. Your submission will fail if any mandatory fields are not completed. If information is unavailable for any mandatory field (or to clarify the data collection status of any field) enter the appropriate keyword to the right. The 8 mandatory fields are as follows:
* There are 8 mandatory fields on the NCBI Biosample data sheet that are required for genome submission to NCBI:
  + **Sample name** (this is your FWS (or LiDS-NG) ID and is already in the sheet)
  + **Organism** (e.g., E. coli or SE or Lmo and is already filled in for you)
  + **Attribute package** (this has been filled in for you already)
  + **Collected by** (this is your organization or an appropriate other)
  + **Collection Date** (this is the year you or the original submittor collected the sample - field sampling date, whatever is the closest to the isolation data)
  + **Geo\_loc** (we need the country at least, if not the province or more specific. This could be sensitive information, so consider the amount of detail you wish to include)
  + **Lat-lon** (if you don't have this or don't want to disclose, put "Missing". This is a required field so SOMETHING must be entered.) OR it can be calculated from the city location - which can be the provincial capital for the province. Better if more specific, but we appreciate there may be sensitivities
  + **Isolation source** (WE MUST PUT SOMETHING IN THIS FIELD. Not only is it required, but this is what makes the sample make sense to others. If one chooses not to employ the dropdown menus, please specify the material (the WHAT) (e.g. water, food type, feces, environmental swab) and WHERE the sample was taken from (e.g. cow rectum, cow manure, river, abattoir, retail store). We need a "what" and we need a "from where". Something that works is "Human feces" or "water - irrigation ditch".
* Other fields to enter, as appropriate.
* ***NCBI requirements for environmental samples are more extensive than for clinical or food samples.*** For example, **Isolation Source** needs to be well specified for all isolates but especially for environmental samples. Even if "Terrain" etc is specified. Just putting a river name in the free text column is not sufficient. First, there's nothing populating the required field, secondly, it's not informative enough. We need to know What did they sample from the river? Biofilm? Sediment? Water? At minimum, we need what material was sampled and where it's from.
* Many fields have controlled vocabulary terms to employ with an enhanced as-you-type lookup function. *Right-click with your mouse in the field cell to access the drop down menus*. Without use of these drop-down menus, mandatory fields will be left blank in the NCBI BioSample worksheet and the NCBI submission will fail*.*
* Most light yellow-coloured input field cells on the Data Entry worksheet are included in the NCBI Biosample submission table and will be uploaded to NCBI. *For fields that you wish to collect for the purpose of the project, but don’t wish to enter, you can toggle whether they get uploaded by italicising the column header in the NCBI Biosample spreadsheet*.
* Similarly, all other input field cells (coloured beige) are merely stored in this worksheet for your reference or for internal use - they are NOT included in the NCBI Biosample submission process to NCBI. They may be used for internal project needs or they might be used in the text of project reports, manuscripts, etc. for better data interpretation.

## Data Entry Worksheet - Field Definitions

### Under the category of *Pathogen Isolate:*

**Pathogen type**. This denotes the attributes package for the NCBI submission. While mandatory it is determined by the entries below. Allowed values are "Pathogen.cl" (for clinical or host-associated pathogen) or "Pathogen.env" (for environmental, food or other pathogen). The value provided in this field drives validation of other fields. For animal-associated food items, use "Clinical"

**Sample name** (yellow field) - Sample Name is a project isolate name or name chosen for the sample. It can have any format, but we suggest that you make it concise, unique and consistent within your lab, or your project and as informative as possible. Every Sample Name from a single Submitter or project consortium within the same Bioproject must be unique.

**Sample title** (yellow field) - Notes: Alternative Sample Name

**Description** (yellow field) - Notes: Free text description of sample. Details can be included for distinguishing isolates based on important attributes or experimental conditions. Entries should use semicolons to separate distinct descriptor phrases.

**BioProject ID** (yellow field) – NCBI Note: Associated BioProject(s), assigned by NCBI

**Organism (species)** (yellow field) - NCBI note: The most descriptive organism name for this sample (to the species, if relevant).

**Strain** (yellow field) - Taxonomic level below species determined by molecular or phenotypic methods. These designations will likely be project specific and are often pre-determined. For example, GRDI-FWS E. coli & Salmonella samples, "Strain" is considered to be the same as “Sample Name” without underscores. e.g. FWSEC001 or LiDS0001.

**Collection date** (yellow field) - NCBI Note: Date of sampling: ISO 8601 standard YYYY-MM-DD

### Under the category of *Isolate Identifiers*: **-** Identifying numbers associated with a sample. Assigned by collectors, testing facilities, lab or sequencing facilities, culture collection repository and Biosample IDs.

There are numerous Identifiers (multiple beige fields) – these are internal institute identifiers used to cross-reference, and will not be exported to NCBI. e.g., Identifying numbers associated with a sample. Assigned by collectors, testing facilities, lab or sequencing facilities, culture collection repository and Biosample IDs

* When there is need (i.e. *almost always*) to record additional organizational or facility identifiers provide them here. These are not included with the NCBI Biosample submission; rather used for cross-referenced within the project or individual submitter and testing labs. (See below)

For example:

* original\_facility\_id
* original\_submittor\_id

**Example Isolate Identifiers:** All are synonyms for the 'sample\_name' field that NCBI will use. The tracking of metadata is the primary reason these alternative Identifiers are tracked.

sample\_name = FWS isolate ID

original\_submitter\_id = sample label (Original Sample ID) – refers to the original isolate ID hopefully the first isolate ID given by the isolating lab

curator1\_isolate\_id = NML isolate ID

curator2\_isolate\_id = LFZ Sample ID

past1\_id = CIPARS lab Sample ID

past2\_id = SISO Sample ID

alt\_isolate\_id

etc.

**Example facility identifiers**: These identify different facilities that have handled the isolate in some fashion.

original\_facility\_id -- describes the entity from which the isolate first derived. This may be a sampling site for those with field testing

curator1\_facility\_id

curator2\_facility\_id

past1\_facility\_id

past2\_facility\_id

alt\_facility\_id

* Sampling Program information

Should you wish to also capture the program for the original submitting facility, we optional columns are included in the spreadsheet

program\_reference\_id - the” program “ of the original\_Submittor\_lab. e.g. CIPARS, FoodNet, CANWARD, etc…

**NCBI contributing organization / project** (yellow field) – INSDC (International Nucleotide Sequence Database Collaboration) keeps an official list of contributors. Each has a special code; and sometimes individual codes for respective collection projects. Your agency and or project needs to be on this list for successful Biosample submission; if it isn't please apply to have it added both to this form and to the submission list below. <http://ftp.ncbi.nlm.nih.gov/pub/taxonomy/coll_dump.txt>

For consortium projects, there will be a single entity for the NCBI submission unless otherwise noted/pre-negotiated.

e.g. The Canadian Genomics Research and Development Initiative - Food and Water Safety (GRDI-FWS) Consortium or The Canadian Listeria Detection and Surveillance using Next Generation Genomics (LiDS-NG) Consortium

### Under the category of *Isolate Testing*:

**Serotype** (brown field) - Taxonomy below subspecies denoted by antigenic properties. Same as serovar and serogroup. Specify somatic (outer membrane and capsular) and flagellar antigens e.g. E. coli O157:H7. There are fields to choose from.

For *Escherichia coli*: optional but highly advisable to note - O, H, K (if known, 60 different K Ags have been recognized), other antigen (specify)

For *Listeria:*

**Serovar** (yellow field) – For Salmonella only, all Salmonella Serovars can be selected by name. All other types should be selected by formula. Be sure to RIGHT-CLICK for easier list lookup.

**Genotype**  (yellow field) - - DNA fingerprint nomenclature or designation such as a cluster code. e.g. 1001EC1837MP, a MLST or ST. where applicable. A cluster code does not necessarily imply it was a cluster but that the particular strain belonged to a group of isolates with a common pattern of interest to researchers.

**PFGE enzyme pattern** (brown field) - The PFGE designation and Enzyme used are species specific. Please note them.

**Clonal complex** (brown field) - if applicable. e.g., for Listeria

**Phage type** (yellow field) -An integer referring to a pattern of opaque or confluent lysis caused by a panel of bacteriophage infections e.g. 14a.

**Subgroup** (yellow field) - Optional additional subtype classification as per organism convention

**Subtype** (yellow field) - Optional additional subtype classification as per organism convention

*Institution code (brown field) – Assigned.*

**Collection code** (yellow field) - Optional additional identification code for a specific culture collection within a repository. Same as “Culture ID”.

**Specimen Id** (yellow field) - Optional additional identification code of a particular specimen within a given collection.

*Specimen (brown field) – Assigned.* Composite of “institution code:collection code:specimen id” e.g. UAM:Mamm:52179. These codes are controlled vocabulary identifiers for tracking specimens in sample collections. Structured vouchers include institution-codes (and optional collection-codes) taken from a controlled vocabulary maintained by the INSDC that denotes the museum or herbarium collection where the specimen resides, please visit: <http://www.insdc.org/controlled-vocabulary-specimenvoucher-qualifier>.

*Culture Id (brown field) – Assigned.* Identification code for a specific culture collection within a repository. Same as “Collection code”.

*Culture collection (brown field) – Assigned.* Notes: Composite identifier containing “institution code:collection code” e.g. ATCC:26370. Annotation with a culture\_collection qualifier implies that the sequence was obtained from a sample retrieved (by the submitter or a collaborator) from the indicated culture collection, or that the sequence was obtained from a sample that was deposited (by the submitter or a collaborator) in the indicated culture collection. See the description for the proper format and list of allowed institutes, http://www.insdc.org/controlled-vocabulary-culturecollection-qualifier.

### Facilitor vs Contact

### When submitting to NCBI this needs to be delineated to distinguish the submitting project consortium (GRDI FWS, LiDS, *etc*.) or individual lab vs. the facilitator (NML, IRIDA) that is submitting on behalf of a client project e.g., IRIDA SRA submissions on behalf of projects for streamlined batch processing. NML IRIDA gas an account that intereacts with NCBI, but IRIDA is just facilitating the submission and is not the contact person.

### For BioSample purposes, one can add in the following fields to make such a distinction:

### BioSample\_submission\_facilitator - IRIDA account @ NCBI

### BioSample\_contact\_information - project consortium lead for genome submissions (named indivudal or team contact that can deal with genome submissions questions

### Under the category of ***Isolate Source Location*:**

**Isolate source location -** Geographic details of the isolate source e.g. country. (we need the country at least, if not the province or more specific. This could be sensitive information, so consider the amount of detail you wish to include).

Country - This field is mandatory, but can be filled with "missing" if too sensitive to release.

State / Province / Region - This field is optional, but can be filled with "missing" if too sensitive to release.

City / Landmark - This field is optional, but can be filled with "missing" if too sensitive to release.

Postal / Zip - This field is optional, but can be filled with "missing" if too sensitive to release.

Latitude - A decimal number, +/- 180 degrees: North = positive value; South = negative value. NCBI requests 4 digit precision and this is calculated based on the above. This field is mandatory, but can be filled with "missing" if too sensitive to release.

Longitude - A decimal number, +/- 180 degrees; East = positive value; West = negative value. NCBI requests 4 digit precision and this is calculated based on the above. This field is mandatory, but can be filled with "missing", if too sensitive to release.

### Under the category of *Isolate Details*:

**Source material category** (brown field) – *There is a drop down on the data entry spreadsheet to select:* options are as follows: Clinical; Domesticated, Avian; Domesticated, Livestock; Environmental, Swab or Sampling; Food; Laboratory Animal Model; Wild, Animal; or Wild, Avian

**Collection matrix** (brown field) - The type of sample taken.

**Host organism type** (yellow field) - For Clinical or Animal isolates, only. The natural (as opposed to laboratory) host for the organism from which the sample was obtained. Use the full taxonomic name, eg, "Homo sapiens"; "Bos taurus", etc.

**Sample body part, product or tissue type** (yellow field) – For Clinical or Animal isolates, only. Type of host tissue or material from which the isolate was extracted. Two links are provided to the BRENDA Ontology animal tissue and plant tissue trees where you can select an appropriate ontology term if it isn't on included list.

**OR Food category**: (brown field) - For Food isolates, only. Select from the drop-down controlled vocabulary listing of food categories.

**and sub-category** (brown field) - For Food isolates, only. *Select from the drop-down controlled vocabulary listing of food sub-categories.*

**and preservation method** For Food isolates, only. *Select from the drop-down controlled vocabulary listing of food preservation methods.*

**OR environmental sample** (brown field) - Environmental & other material (e.g. manure, sediment, water), but excluding animal, plant, and food sources. *Select from the drop-down controlled vocabulary listing of categories*.

**Sample location type** (brown field) - Built or natural environmental features and locations where the sample material was taken from (e.g. abattoirs, cages, food processing facility, river, irrigation ditch, *etc*.). *Select from the drop-down controlled vocabulary listing of categories.*

**Isolation source** (yellow field) - Describes the physical, environmental and/or local geographical source of the biological sample from which the sample was derived. Summary of attributes specified in the above searchable fields will be concatenated as the Isolate Source description and exported for “Isolation Source” in Biosample.

Under the category of *Clinical Details*

Describes attributes of the host at the time of sample collection, if known. Entries below are optional except for **host individual's details** (required for clinical isolates only).

**\*host individual's details** (required for clinical isolate; free text field) - Mandatory entry for clinical isolates. Additional information not included in other defined vocabulary fields, but not to include specific patient identifiers or other private information Entries should use semicolons to separate distinct descriptor phrases.

Host subject Id –A unique identifier by which each subject can be referred to, de-identified, e.g. #131. [Optional.entry. Not essential for NCBI submission].

Host age - Age of host at the time of sampling. Optional entry.

Host disease outcome - Final outcome of disease, if known. e.g., acute gastroenteritis, death, chronic disease, recovery. Optional entry.

Host disease stage - Stage of disease at the time of sampling, if known. Optional entry.

Host health state - Information regarding the health state of the individual sampled at the time of sampling, if known. Optional entry.

Host sex - Gender or physical sex of the host, if known. [For some animal hosts, this may likely not be known]. Optional entry.

Host description (misc details) - Additional useful host information not included in other defined vocabulary fields. Entries should use semicolons to separate distinct descriptor phrases.

### Under the category of *Processing*:

Sequencing facility – For example, for GRDI-FWS samples, all sequencing was carried out by the National Microbiology Laboratory. Please define as “NML”.

Sequencing platform - Defines the sequencing instrumentation and associated chemistry e.g. Illumina SBS (Illumina sequencing by synthesis), please enter if known. Otherwise consult with the sequencing facility.

Under the category of Isolate Environmental Details - Describes the habitat, season, bacterial counts and water chemistry associated with environmental isolates. e.g fecal indicator bacteria number (cfu/mL) and turbidity, when such data is collected. Terrain and biome are mandatory entries for environmental isolates.

**Environmental terrain** - Describes the land surface e.g. desert, forest from which the specimen was isolated, when known. *Select from the drop-down controlled vocabulary listing of environmental terrain categories*. Mandatory entry for environmental isolates.

**Environment – biome** - Describes the land use and/or ecosystem from which the specimen was isolated. Can overlap with Environmental terrain; but is broader and includes aquatic and anthropogenic habitats. Biomes are defined based on factors such as plant structures, leaf types, plant spacing, and other factors like climate. Biome should be treated as the descriptor of the broad ecological context of a sample. Examples include: desert, taiga, deciduous woodland, or coral reef. EnvO (v 2013-06-14) terms can be found via the link: [www.environmentontology.org/Browse-EnvO](http://www.environmentontology.org/Browse-EnvO). *Select from the drop-down controlled vocabulary listing of environmental terrain categories*. Mandatory entry for environmental isolates.

**Season classification** - Division of the year, marked by changes in weather, ecology, and hours of daylight, as experienced in North America. *Select from the drop-down controlled vocabulary listing of environmental terrain categories*. Optional entry; but desirable for environmental samples.

The remaining entries are optional entries for NCBI submission, but may be included if desired.

* Fecal indicator bacteria number, in cfu/ml
* Collection depth, in meters - The vertical distance below local surface for the collected isolate, e.g. for sediment or soil samples depth is measured from sediment or soil surface, respectively, in metres.
* Temperature (sample) in degrees Celcius - Temperature of the sample (or local environment) at time of sampling.
* Average daily temp. (site) in degrees Celcius - Usually looked up in weather service for date 3 days prior to sample collection date. Enter if known. Otherwise enter Not Collected.
* Precipitation, in millimetres per day
* Water velocity, in metres cubed per second -
* pH, 1 – 14 -
* Turbidity, in formazin turbidity / nephelometric unit(s) -
* Conductivity, in milliSiemens per cm -
* Redox potential, in millivolt -
* Dissolved oxygen concentration, in umol/kg -
* Nitrate concentration, in micromole per litre -
* Nitrite concentration, in micromole per litre -
* Phosphorous concentration, in micromole per litre-
* Soluble inorganic material (name) in \_\_\_\_\_ concentration \_\_\_\_\_ umol/kg. A named substance and its concentration and units measurement for substances such as ammonia, road-salt, sea-salt, cyanide, hydrogen sulfide, thiocyanates, thiosulfates, *etc..*

1. Curating Data and Viewing in the “NCBI BIOSAMPLE” Worksheet:

*It is recommended that after you manually enter all data, and that you then inspect and curate the entry in the NCBI Biosample worksheet to ensure that all information was captured accurately.*

Notes:

* Any *ITALICISED* column headers in the NCBI Biosample worksheet are not exported for genome submission to NCBI. However, they are extremely useful and often used to generate mandatory NCBI submission fields. Therefore please fill in to the best of your ability
* Column headers LACKING an asterisk (\*) are OPTIONAL to export to NCBI, but again are extremely useful for internal project use, for cross-referencing or for additional isolate descriptions used in manuscripts, etc...
* Multiple identifying IDs (original ID, past ID, curator ID, etc. ) are key for internally keeping track of the isolate as it changed hands - as we have clearly learned is problematic. They are not exported to NCBI. So while not required by NCBI, the NML (as the genome and metadata submitting organization) and you need these for record keeping and metadata capture/cross-referencing. It is in your and your organisation’s best interests to provide these (*if more past IDs are needed, let us know - we can enter an unlimited number of columns to capture them all...*
* To capture the “sampling or surveillance program” for the original submitting facility, use **program\_reference\_id** - the” program “ of the original\_Submittor\_lab. e.g. CIPARS, FoodNet, CANWARD, etc….
* ***The requirement for environmental samples are actually much more strict than for clinical or food samples at NCBI.*** For example, **Isolation Source** needs to be well specified for all isolates but especially for environmental samples. Even if "Terrain" etc is specified. Just putting a river name in the free text column is not sufficient. First, there's nothing populating the required field, secondly, it's not informative enough. We need to know **what** did they sample from the river? Biofilm? Sediment? Water? At minimum, we need t**he material sampled** and **where it's from**.

Although it is possible to copy and paste entries from one NCBI row to another if the information is shared between isolates. Use this with extreme caution. ***Keep in mind this is exactly how the NCBI database has become filled with propagated errors that can have a real negative impact on the research of others.*** Make all entries as straightforward and as descriptive as possible, using a previously saved manual data entry line item in the NCBI Biosample worksheet as the source.

# META DATA Short Cheat Sheet

Multiple identifying IDs (original ID, past ID, curator ID, etc. ) are key for keeping track of the isolate as it changed hands - as we have clearly learned is problematic. So while it isn't required by NCBI, the NML and you need these for record keeping and metadata capture/cross-referencing. It is in your and your organisations’ best interest to provide these and **if more past IDs are needed, let us know - we can enter an unlimited number of columns to capture them all..**. they are not exported to NCBI.

Should you wish to designate one id column as **program\_reference\_id**  please just let us know.

***The requirement for environmental samples are actually much more strict than for clinical or food samples at NCBI.*** **Isolation Source** needs to be well specified for environmental samples. Even if "Terrain" etc is specified. Just putting a river name in the free text column is not sufficient. First, there's nothing populating the required field, secondly, it's not informative enough.

We need to know **What** did they sample from the river? Biofilm? Sediment? Water?

At minimum, we need t**he material sampled** and **where it's from**.

There are **8 mandatory fields** that are required for genome submission to NCBI BioSample:

1. **Sample name** (this is your FWS (or LiDS-NG) ID and is already in the sheet)
2. **Organism** (e.g., E. coli or SE or Lmo and is already filled in for you)
3. **Attribute package** (this has been filled in for you already)
4. **Collected by** (this is your organization or an appropriate other)
5. **Collection Date** (this is the year you or the original submittor *collected* the sample - field sampling date, whatever is the closest to the isolation data)
6. **Geo\_loc** (we need the country at least, if not the province or more specific. This could be sensitive information, so consider the amount of detail you wish to include)
7. **Lat-lon** (if you don't have this or don't want to disclose, put "Missing". This is a required field so SOMETHING must be entered.) OR it can be calculated from the city location - which can be the provincial capital for the province. Better if more specific, but we appreciate there may be sensitivities
8. **Isolation source** (WE MUST PUT SOMETHING IN THIS FIELD. Not only is it required, but this is what makes the sample make sense to others. If one chooses not to employ the dropdown menus, please specify **the material (the WHAT)** (e.g. water, food type, feces, environmental swab) and **WHERE the sample was taken from** (e.g. cow rectum, cow manure, river, abattoir, retail store). We need a "what" and we need a "from where". Something that works is "Human feces" or "water - irrigation ditch".

We know looking at it all for the first time is overwhelming, but *we simply need 5 additional pieces of information per isolate.*

Anything beyond that is for the purpose of improving isolate descriptions, for example for improving the science and for the manuscripts, and for developing the ontologies - which is also an output of the FWS project. *Current pain is for future gain...when this can be made a better more streamlined process.*