

## Metadata Template User Guide

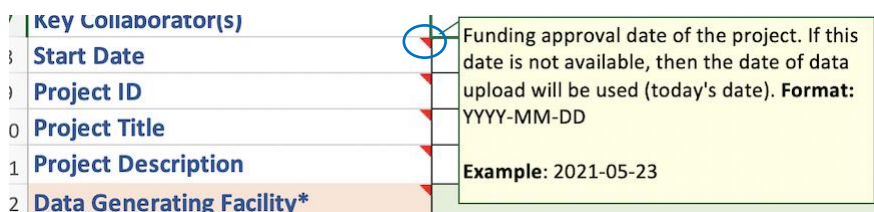
The fields on this form comprise the curation-level metadata needed for optimal deposition of master copies of genomic data into the Data Management Environment (DME). Sample level metadata can be added or updated after initial submission. In addition to minimizing the risk of data loss, provision of these fields assures that data are interoperable with the basic data management capabilities provided by DME and data platforms linked to DME, including group-level and IRP-level data catalog functionality. Users who wish to leverage advanced data management and analysis capabilities available in the environment, including support in NIDAP for data submission to the GDC, CDS, GEO, and/or dbGaP data repositories, are encouraged to review the “Recommended Fields” tabs and complete the additional fields.

### 1. Template Functionality

The metadata template is a form built within Microsoft Excel. It is a dynamic form, as selecting a particular field will auto-populate corresponding fields within the form, e.g. selecting TRUE for “Is Cell Line?” will auto-populate fields to capture Cell Line Name, whereas selecting FALSE will auto-populate fields that would be needed for non-cell line samples.

#### 1.1. Hover-over Notes

Fields on the form with a red triangle in the upper right-hand corner have the functionality of a hover-over note. If a user hovers their mouse cursor over the field, then a note will appear that will give the user more information on that field. This information within the note can also be found in the data dictionary. Information within the hover-over note can contain field descriptions, formatting directions, and examples to help users understand how to fill out the template. An example of the hover-over note is as follows, with the red triangle indicator circled in blue:



#### 1.2. Field Colors

There are currently several field colors on the template that indicate different functionality on the form. Users can refer to the following explanations to better understand what the different colors require from their data.

- 1.2.1. **Pink/Peach-** The fields that are a peach/pink color indicate Project-level Metadata that need to be addressed before moving on to the Sample-level Metadata. The data collected in these fields will auto-populate fields for

corresponding sample data. Please address these pink fields before entering anything in the sample section.

## 2 Data Generating Facility\*

- 1.2.2. **Green-** If a field on the form is highlighted in green, then these fields contain a drop-down selection. Clicking on the drop-down area when a green field is selected will display a list of choices to choose from for this data. Please note that some drop downs will indicate to make an entry instead of a selection. In these cases, you can overwrite the drop-down selection with free text of your choosing. If a cell is not colored green, then users can enter text freely.



- 1.2.3. **Blue-** Blue fields are found in the Experimental and Technical Conditions section of the template. These fields are for users to include data needed for analysis, such as Tumor Stage, Treatment, Drug, Response, etc. Batch is included in this blue section because it is required for analysis of study data and for tracking samples. To enter data into the blue fields, simply select one of the fields labeled as Condition and overwrite with your preferred field name. You can now enter the additional sample-level data that you have deemed vital to your project.

Experimental and Technical Conditions -	
Batch	Condition

### 1.3. Drop-downs vs Free Text

As indicated in 1.1.2, the green colored fields contain drop-down menus for users to pick a data choice instead of entering in the text themselves. This allows for managing data for the value domain, as well as standardizing the template for certain data types. Some drop-downs may direct the user to enter free text as the only choice for their data entry. Users can type in those template cells, as they would be able to type in any other cell of the form.

There following is an example of drop-down menu with choices displayed after clicking on the gray arrow (circled in red):

Project Name	Organism	Cancer Type	Strain
		▼	

The following is an example where a user can overwrite the drop-down selections and use free text:

Project Name	Organism	Strain
		▼
		Please enter Organism Strain

If a cell is not colored green, then users can enter text freely.

#### 1.4. Dynamic fields

This template has the capability of auto-populating fields based on drop-down choices. For example, if a user were to choose Organism as “Human” at the project level, then the sample level will populate fields that correspond to human data. The non-human choices (mouse, rat, etc.) will populate sample fields that correspond to non-human data (i.e., stain). Please fill out the dynamic fields highlighted in pink/peach color in the project level before moving on to sample level.

#### 1.5. Data Dictionary

The Data Dictionary tab contains descriptions and examples for all fields within the template. Descriptions give users further details about a given field on the template, while Examples give users potential data entries for a given field. This tab also contains the mappings for NIDAP to other repositories.

#### 1.6. Disease, Diagnoses, Antibodies tab

The Disease, Diagnoses, Antibodies tab contains reference material for the template. There are fields on this template that require users to choose values that cannot fit into drop-down menus. For example, Primary Diagnosis has over 2000 potential values to choose from, therefore this tab acts as a reference guide for certain fields in the template. It contains Disease options for NIDAP, Disease Type & Primary Diagnosis for GDC and CDS, and ChIP Antibody Targets for ChIP-seq data. It also contains links to where these vocabularies originate from and are managed for future reference.

## 2. Project-level Metadata

The section of the metadata template that contains project-level metadata is required for users, as it collects vital project, contact, and study data that is necessary to accompany your samples. Please fill out all fields possible in the project-level before moving on to the

sample-level, especially the fields that are highlighted in pink/peach coloring.

Project-level Metadata		
Data Owner		
Data Owner Affiliation		
Data Curator (for the Data Owner)		
Project Scientist/Project POC		
Project Scientist/Project POC Email		
Key Collaborator(s)		
Start Date		
Project ID		
Project Title		
Project Description		
Data Generating Facility*		*fill out before Sample-level
Sequencing Platform		
Is Cell Line?*		*fill out before Sample-level
Organism*		*fill out before Sample-level

### 3. Sample-level metadata

The sample-level metadata on this template is required to collect information about your project samples. Users can utilize as many rows necessary to be included in on this template. The first five fields in the sample section are required for all samples (Sample Name, Subject ID, Disease, Library Strategy, and Analyte Type). The remaining five fields after Analyte Type will change depending on the drop-down selections made in the project-level metadata section (Is Cell Line and Organism).

The first three rows of the sample section contain the field name, a description of the field, and an example of the field to help guide users on how to fill out the information for their samples.

One of the dynamic fields on this form is within the Library Strategy field. The choice that is made in this field can auto-populate new required fields in the blue Conditions sections of the template. For example, choosing ChIP-seq as Library Strategy for a sample will populate a field in Conditions that ask for information related to ChIP antibody and ChIP file name.

Sample-level Metadata							
Sample Name	Subject ID	Disease	Library Strategy	Analyte Type	Tissue	Tissue Type	Developmental Stage
See Note above for Sample Name Description. (Hover cursor over Sample Name field)	See Note above for Subject ID Description. (Hover cursor over Subject ID field)	The type of malignant disease.	Sequencing method used for this project. For multimodal single cell experiments, please use one row per Library Strategy method. ChIP-seq, Exome-seq, and Whole Genome-seq will require additional fields in the Conditions section (blue).	The kind of molecular specimen analyte. Valid Entry: cfDNA, DNA, EBV Immortalized Normal, FFPE DNA, FFPE RNA, GenomePlex (Rubicon) Amplified DNA, Nuclei RNA, Repli-G (Qiagen) DNA, Repli-G Pooled (Qiagen) DNA, Repli-G X (Qiagen) DNA, RNA, Total RNA.	Please select name of the studied tissue or organ. NOS is "Not otherwise specified."	The kind of tissue collected with respect to disease status or proximity to tumor tissue. Valid Entry: Tumor, Normal, Abnormal, Peritumoral, Unknown, Not Reported	The developmental stage of the sample, embryo
See Note above for Sample Name Examples.	See Note above for Subject ID Examples.	Breast Cancer	RNA-Seq	DNA	Lung	Tumor	

Please note that on the far-left hand side of the samples (between rows 18 and 21), there is a “plus” sign or “minus” sign that allows user to minimize or maximize the description and example rows. This is simply a way for user to customize their view of the form and contains no other functionality.

16					
17	Sample-level Metadata				
18	Sample Name				
19	See Note above for Sample Name Examples. See Note above for Examples.				
20					
21					
22					

#### 4. Experimental and Technical Conditions

The Experimental and Technical Conditions section of the template are technically a subsection of the Sample-level metadata. These fields are for users to include data needed for analysis, such as Tumor Stage, Treatment, Drug, Response, etc. Batch is included in this blue section because it is required for analysis of study data and for tracking samples. To enter data into the blue fields, simply select one of the fields labeled as Condition and overwrite with your preferred field name. You can now enter the additional sample-level data that you have deemed vital to your project. The Description and Example rows underneath the new column heading field can be cleared out (delete text within cells).

Some of the fields in the Conditions section will be auto-populated from previous dynamic field choices. If more condition fields are needed, user can create new additional fields after the blue section.

#### 5. Additional Fields

As stated in the Experimental and Technical Conditions, some of the fields in the Conditions section will be auto-populated from previous dynamic field choices. If more fields are needed, user can create new additional fields after the blue section.

#### 6. Additional Tabs

The metadata template has several tabs. For the purposes of NIDAP, the first field labeled "Required Fields – User Form" is the only field that needs to be filled out. There are additional recommended fields in the tabs that follow. These tabs are for data submission to other repositories of interest to researchers and PIs at NCI- dbGaP, CDS, GDC, and GEO. Some fields are already mapped to these repositories from the Required Fields tab of NIDAP, but additional data needs to be collected to meet the requirements of the aforementioned repositories. These tabs are here to aid in further data sharing of project data.

Fields - User Form	Recommended Fields for dbGaP	Recommended Fields for CDS	Recommended Fields for GDC	Recommended Fields for GEO	Data Dictionary	Disease, Diagnoses, Antibodies
--------------------	------------------------------	----------------------------	----------------------------	----------------------------	-----------------	--------------------------------