How to Submit Metadata and Data files to the 4DN-DCIC

In order to make your data accessible, searchable and assessable you should submit as much metadata as possible to the 4DN system as well as the raw fastq files that you have generated. This guide will show you how to find out what kind of metadata we collect for your particular type of experiment and the mechanisms by which you can submit your metadata and data to the 4DN system.

Before you can submit data to the 4DN system you must be a registered user of the site and have the appropriate access credentials. You must be designated as a submitter for the lab for which you want to submit files and metadata. To get set up with an account with the correct access contact the data wranglers.

A quick note on metadata and data accessibility. For most metadata items the default permission will be that the data will only be viewable by the members of the submitting lab and will only be editable by users who have been designated as submitters for that lab. The metadata will also be accessible to data wranglers who can help you review the data and alert you to any issues as the submission is ongoing. Once the data and metadata are complete and quality controlled they will be released according to the data release policy adopted by the 4DN consortium.

Using a Microsoft Excel WorkBook as a template for submission

Preparing the data
Based on the type of experiment(s) that you plan to submit data for the data wranglers will provide you with a Microsoft Excel WorkBook containing several WorkSheets. Each sheet corresponds to an Item type in our metadata database. The workbook provided should contain all the sheets that you may need for your submission. Each sheet should also contain all the data fields that can be submitted for that Item type. Depending on if you have submitted data before or you are possibly using shared reagents you may not need to provide information on every sheet.

Generally, it makes sense to begin with the left most sheet in the workbook as the sheets in a workbook are ordered so that Items that have fields that take a reference to another Item as their value appear ‘after’ i.e. to the right of that Item’s sheet in the workbook.

Excel Headers
1) Field name
2) Field description
3) Choices for controlled vocabulary (some fields only accept a value from a list of selection, like experiment type)
A sheet for an Item starts with a row of field names. The second row of the sheet includes a description of each of the fields (as it appears on the web pages for that Item). In some cases the values that you can submit for a particular field are constrained to a specific set of terms and when this is the case the possible values are shown in the third row.
*The first entry will start from row 4, and column 2.

Entering Values
Each field can be a certain type; string, number/integer, list. If the type is integer, number or array, it will be indicated with the fields name; field:number, fields:int, field:array. If the field is a string, you will only see the field name.
If the field is an array (field:list), you may enter a single item, or multiple items separated by comma.

field:array
description
enums
item1,item2,item2,item4

Most field values are strings - either a term from a list, a string that identifies an Item, or a text description. However, there are some fields values that require specific formatting. These cases and how to identify them are described below.

When the string must conform to a certain format
In some cases a field value must be formatted in a certain way or the Item will fail validation. Examples of these are Date fields (YYYY-MM-DD format) and URLs (checked for proper URI syntax).
In other cases a field value must match a certain pattern. For example, if a field requires a DNA sequence then the submitted value must contain only the characters A, T, G, C or N.

Database Cross Reference (DBxref) fields that contain identifiers that refer to external databases are another special formatting case. In many cases the values of these fields need to be in 'database_name':ID format. For example, an SRA experiment identifier would need to be submitted in the form 'SRA:SRX1234567'. Note that in a few cases where the field takes only identifiers for one or two specific databases the ID alone can be entered - for example, when entering gene symbols in the 'targeted_genes' field of the Target Item you can enter only the gene symbols i.e. PARK2, DLG1.

When a field is a list
Some fields allow you to enter more than one value of the same type. For example you can have a Biosample that contains multiple modifications. In these cases the field name is appended with ':array' to indicate that multiple values are allowed. Enter the values as a comma separated list i.e. value1,value2,value3 …

When a field specifies a linked item
Some fields in an Item may contain references to a different Item. These may be of the same type or different types. Examples of this type of fields include the 'biosource' field in Biosample or the 'experiments_in_set' field in the ExperimentSet Item. Note that the latter is also an example of a list field that can take multiple values.

When field(s) indicate an embedded object
Some Items can contain embedded sub-objects that are stored under a single Item field name but that contain multiple sub-fields that remain grouped together. These are indicated in the Item spreadsheet using a '.' (dot) notation. For example the "experiment_relations" field has 2 sub-fields called "relationship_type", and "experiment". In the spreadsheet field names you will see experiment_relations.relationship_type and experiment_relations.experiment.

If the Item field is designed to store a list of embedded sub-objects, you can enter multiple sub-objects by manually creating new columns and appending incremented integers to the fields names for each new sub-object.

For example, to submit a total of three related experiments to an Experiment-HiC Item you would find the experiment_relations.relationship_type and experiment_relations.experiment columns, copy them and add four new columns to the sheet named:
experiment_relations.relationship_type-1
experiment_relations.experiment-1
experiment_relations.relationship_type-2
experiment_relations.experiment-2

and enter a valid 'relationship_type' term and experiment identifier to each of the three pairs of columns.

When field is "attachment"
If there are any files (pdf, png, …) that you would like to submit with the object "document" or "image", you need to enter the path of the file under the fields "attachment'. The patch can be the full path, or the relative path to the excel file.

Using aliases
Every item in our database is assigned a "uuid" as upon entry. Some items like experiments and files also have "accession". These two are the default identifying term of any item. Besides these two, there can be object specific identifying terms, like award number for awards, or lab name for labs. All these values can be used for referencing existing items in the excel sheets. However, when you enter information in the excel sheets that needs to reference each other you can not use uuid or accession since they are not assigned yet. To overcome this problem, we will be using aliases.

An alias is a lab specific identifier that you can assign to an item. Aliases take the form of lab:id_string. An alias must be unique within all items. Once you submit an alias for an Item then that alias can be used as an identifier for that Item in the submission you are working on (if the Item will be referenced by another field in the submission) or a subsequent submission.