**Submitting a Dataset to HuBMAP**

Every dataset submission has two *required components* and one *optional component* submitted together in a data ***submission directory***.

Required components:

* One ***data directory*** per dataset
* One ***assay*** ***metadata.tsv*** per assay type

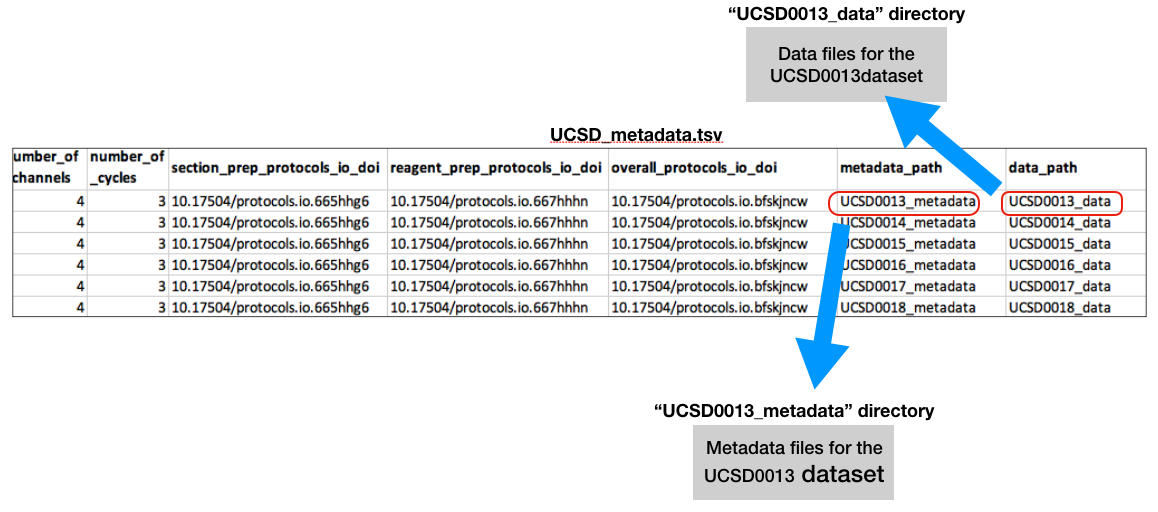
Optional component, dependent upon the assay type:

* One ***metadata directory*** per dataset

If multiple datasets have been generated with the same assay\_type, they may be submitted together in a single data ***submission directory*** with a single ***assay metadata.tsv*** *l*isting all datasets, one per row (**Figure 1**). Each ***data directory***contains the data files (eg. *image files, fastq files, etc)* for a single dataset.

Each ***metadata directory*** contains optional relevant unstructured metadata files (eg. QC reports, instrument metadata in a *json, csv, txt, xml. etc*) for a single dataset.

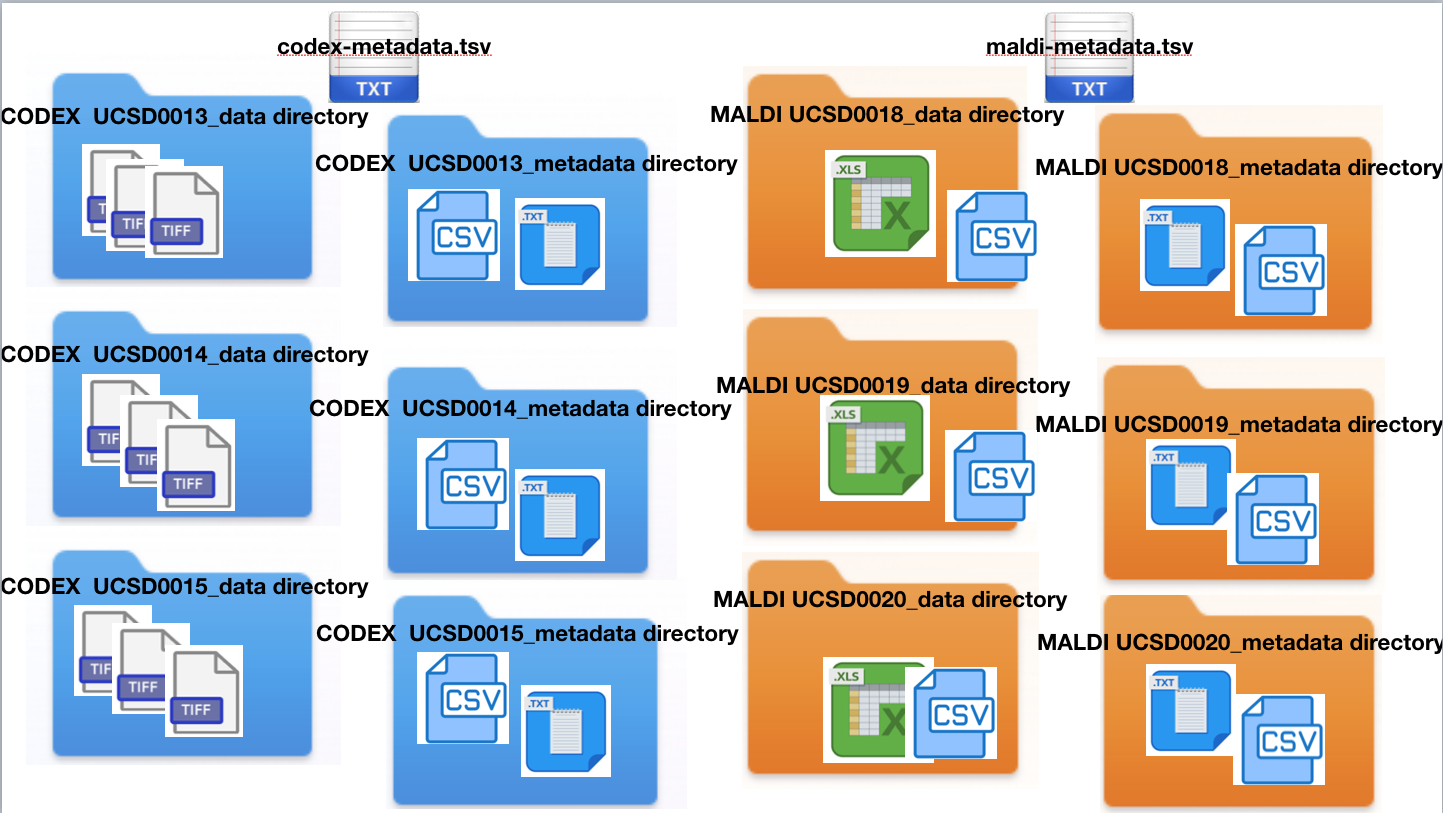
The purpose of the ***assay metadata.tsv*** is to provide levels 1, 2 & 3 assay metadata features (eg. *protocol links, instruments in addition to assay-specific parameters*) for each dataset. Datasets are listed one per row. Since different assays involve different assay parameters, HuBMAP provides several assay-type specific metadata.tsv templates (eg. CODEX, MALDI-IMS, autofluorescence microscopy, LC-MS, RNAseq, etc). The last 2 fields of every ***assay metadata.tsv*** (Figure 1: *metadata\_path* & *data\_path* are circled in red) contain the relative paths to the corresponding *data directory* and optional *metadata directory* for each dataset.



**Figure 1: An assay-type specific *assay metadata.tsv* lists datasets in the submission directory for that assay-type. Datasets are listed one per row. The data\_path fields point to the corresponding dataset directories in the data submission directory. The optional metadata\_path fields point to the corresponding metadata directories.**

**Figure 2** below shows the general directory structure of a data submission. Note that a submission directory may contain multiple ***data directories*** and ***metadata directories***, each directory corresponding to one dataset/row in a single assay-specific ***assay metadata.tsv***.

In addition, multiple assay-specific metadata.tsvs may be included together in a data **submission directory**. For example, the codex***-metadata.tsv*** below lists datasets of the CODEX assay-type (*shown in* ***blue***) in the **submission directory** while the maldi***-metadata.tsv*** lists datasets of the MALDI assay-type (*shown in* ***orange***) in the same **submission directory**:



**Figure 2: A data submission directory may contain multiple datasets for multiple assay-types. Each dataset is provided in a corresponding data directory with optional metadata provided in a corresponding metadata directory. Each assay-specific metadata.tsv *(eg.codex\_metadata.tsv, maldi\_metadata.tsv)* in the dataset submission directory lists the corresponding datasets.**

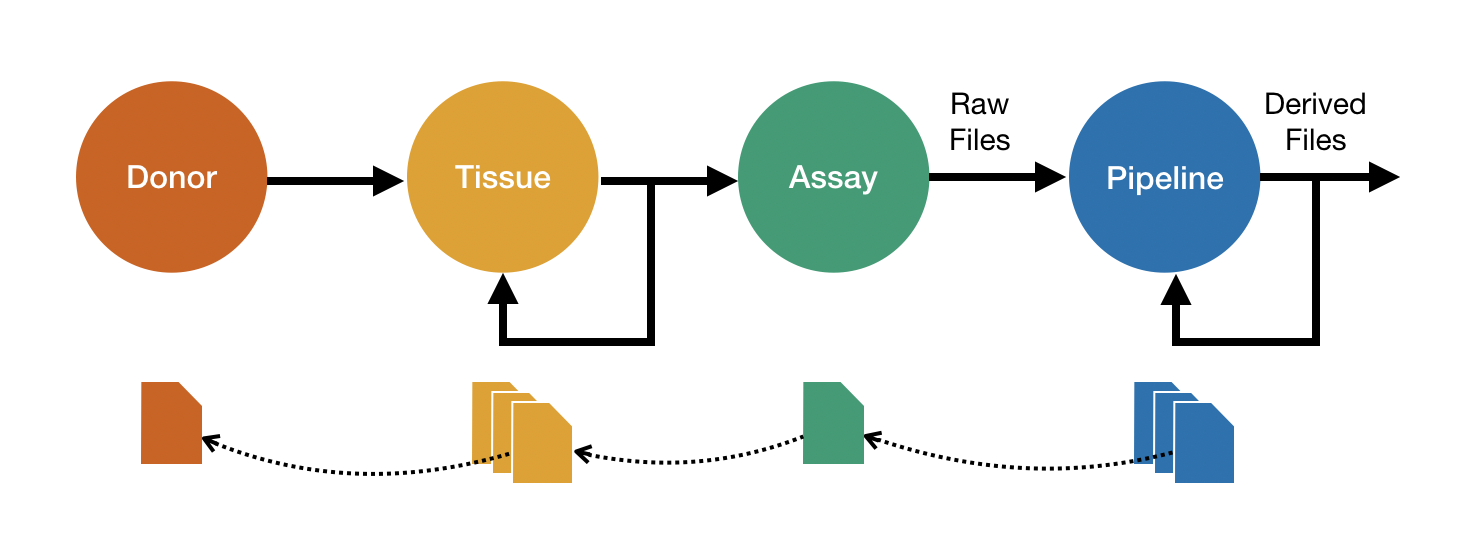
**Preparing an Assay Metadata.tsv for Data Submission**

HuBMAP supports 3 categories of assays : *imaging, nucleotide sequen*ce an*d mass spectromet*ry. Each assay category encompasses a variety of unique assay-types involving unique chemistries, platforms, data types and analysis tools. Below are examples of assay-types from each assay category in HuBMAP:

* mass\_spectrometry:
  + LC-MS
  + MS
  + TMT
* imaging:
  + AF
  + CODEX
  + Imaging Mass Cytometry
  + multiplexed IF
  + PAS microscopy
  + seqFISH
  + MALDI-IMS positive
  + MALDI-IMS negative
* sequencing:
  + bulk RNA
  + bulk ATAC
  + scRNA-Seq
  + sci-ATAC-seq
  + sci-RNA-seq
  + SNARE-SEQ2 (RNAseq & ATACseq)
  + snATAC
  + snRNA
  + SPLiT-Seq
  + WGS

**Assay Metadata**

Data centers provide the following 4 data types for each data submission to HubMAP:



Assay metadata, which is described in the [Assay Metadata Submission Format](https://docs.google.com/document/d/1g82GpCpFDKew60XzAO4Siaw3ZXJjwsaCpgPwhqQZxIY/edit#heading=h.qeehtnf68fas) document, is divided into 4 levels:

## Definition of assay metadata levels

* **Level 1:** Are attributes that are common to all assays, for example, the type (“CODEX”) and category of assay (“imaging”), a timestamp, and the name of the person who executed the assay.
* **Level 2:** Attributes that are common to a category of HuBMAP assays, i.e. *imaging*, *sequencing*, or *mass spectrometry*. For example, for imaging assays this includes fields such as *x resolution* and *y resolution.*
* **Level 3:** Attributes that are specific to the type of assay, for example for CODEX that would include *number of antibodies* and *number of cycles*.
* **Level 4:** Supplementary information such a QC report or information that is unique to a lab, not required for reproducibility or is otherwise not relevant for outside groups. This information is submitted in the form of a single file, a ZIP archive containing multiple files, or a directory of files. There is no formatting requirement (*although formats readable with common tools such as text editors are preferable over proprietary binary formats)*.

|  |
| --- |
| **Level 1 metadata is required for all assays:** |
| **execution\_datetime:** *A time stamp marking the beginning of data capture for the assay. This typically comes from a date-time stamped folder generated by the acquisition instrument.* |
| **protocols\_io\_doi**: *All HuBMAP protocols must be available through a protocols.io DOI link.* |
| **operator:** *The individual who conducted the assay.* |
| **operator\_email**: |
| **pi:** *The principal investigator providing the data.* |
| **pi\_email**: |
| **assay\_category:** One of 3 general assay categories- *imaging, mass spectrometry or nucleotide sequencing.* |
| **assay\_type:** *The specific type of assay being executed (eg. CODEX, immunofluorescence microscopy, LC-mass spectrometry, RNAseq, ATACseq, etc)* |
| **analyte\_class:** *Proteins, nucleic acids, metabolites, lipids.* |
| **is\_targeted:** *A boolean indicating if the assay targets a specific analyte (eg. specific mRNA(s) or protein(s).* |
|  |
| **Level 2 metadata encompasses metadata relevant to an assay category *(****CODEX Level 2 metadata is shown here as an example):* |
| **acquisition\_instrument\_vendor**: *For CODEX, the data acquisition instrument is a platform with a microscope & camera that captures an image of the sample.* |
| **acquisition\_instrument\_model:** |
| **resolution\_x\_value:** *The width of a pixel.* |
| **resolution\_x\_unit:** *The units of the resolution\_x\_value.* |
| **resolution\_y\_value:** *The height of a pixel.* |
| **resolution\_y\_unit:** *The units of the resolution\_y\_value.* |
| **resolution\_z\_value:** *Images of the sample are captured at different focal planes to generate depth of resolution. The z-resolution refers to the distance between focal planes in the sample as the microscope stage is moved vertically.* |
| **resolution\_z\_unit:** *The units of the resolution\_z\_value.* |
|  |
| **Level 3 metadata encompasses more detailed assay parameters specific to an assay\_type** *(CODEX Level 3 metadata shown here as an example) :* |
| **preparation\_instrument\_vendor:** *For CODEX, the robotic platform used to perform the cycles of application of antibodies, image capture and stripping of antibodies from the sample*. |
| **preparation\_instrument\_model** |
| **number\_of\_antibodies** |
| **number\_of\_channels** |
| **number\_of\_cycles** |
| **section\_prep\_protocols\_io\_doi** |
| **reagent\_prep\_protocols\_io\_doi** |

**Level 4 metadata encompasses unstructured metadata such as QC reports, signal to noise plots, instrument metadata jsons, csvs, etc.**