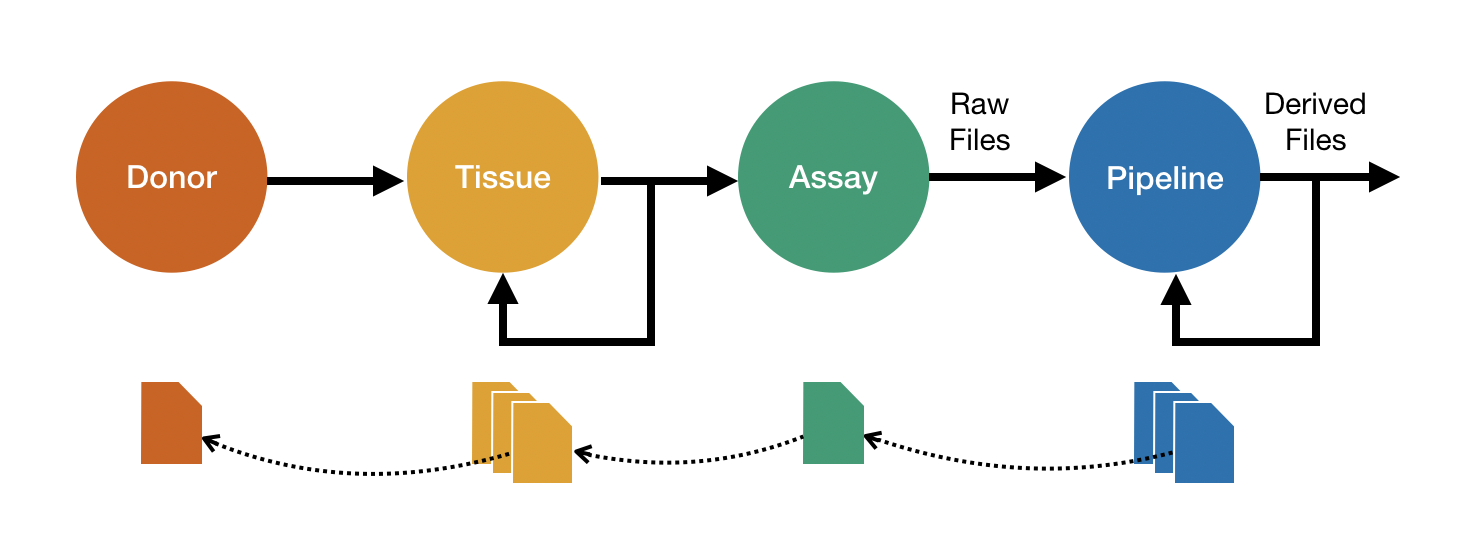
HuBMAP   
Assay Metadata Submission Format

Draft Specification

*Chuck McCallum, Christine Briggs, Nils Gehlenborg*

# Overview

The approach described below is based on this simplified model of HuBMAP data entities (the file icons are representing metadata files):



Metadata for HuBMAP assays are submitted as tabular text files with the following *sections*:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tissue**  **References** | **Level 1**  **Attributes** | **Level 2**  **Attributes** | **Level 3**  **Attributes** | **Level 4**  **Reference** | **Output Data Reference** |
|  |  |  |  |  |  |
| … | … | … | … | … | … |

Sections general include *multiple columns* and are used to capture the metadata attributes associated with the assay (“Level 1”, “Level 2”, “Level 3”, “Level 4”, see below for definitions), to reference the input material for the assay (“Sample Reference”) as well as to reference the *raw* data files generated by the assay (“Output Data Reference”).

In this context “reference” describes a column that contains a pointer to a HuBMAP ID, to a file (could also be a ZIP archive), or to a directory containing multiple files. The term “attribute” corresponds to a column in the tabular file.

A row in the metadata file can either refer to a single data file or to a directory of data files (which might contain further directories). In the latter case, the attributes provided in that row would apply to all files contained within the directory.

# Metadata Levels for Assays

## Definition of Levels

* **Level 1:** These 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:** These are 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:** These are 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:** This is information that might be unique to a lab or is 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).

## Use in HuBMAP Portal

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Indexed for Faceted Browsing** | **Indexed for Free-Text Search** | **Displayed in Assay Over-**  **view** | **Displayed in Assay Detail View** | **Available for Download** |
| **Level 1** | ✓ | ✓ | ✓ | ✓ | ✓ |
| **Level 2** | ✓ | ✓ | ✓ | ✓ | ✓ |
| **Level 3** | (✓) | ✓ |  | ✓ | ✓ |
| **Level 4** |  |  |  |  | ✓ |

# Example for CODEX Data

We have prepared a sample spreadsheet (with mostly fictional data) for a submission of a single CODEX data set that illustrates how the definitions below would be implemented.

<https://docs.google.com/spreadsheets/d/1GaBbtXb0jWCcx_G0RWrBp3ME4yyqglGbQJGC6xCkkHs/edit#gid=1784808680>

# Format Definition (here: “CODEX”)

## Tissue Reference Section

This section links back to the tissue that the assay was executed on. The donor ID is redundant but can be used for sanity checks against the tissue ID.

### donor\_id

**Description:** HuBMAP ID of the donor of the assayed tissue.

**Allowed Values:** HuBMAP ID (must reference a donor)

### tissue\_id

**Description:** HuBMAP ID of the assayed tissue.

**Allowed Values:** HuBMAP ID (must reference a tissue)

## Level 1 Attributes Section

### execution\_datetime

**Description:** Start date and time of assay.

**Allowed Values:** A date and time string formatted as “YYYY-MM-DD hh:mm ZZZ”, where YYYY is the year, MM is the month with leading 0s, and DD is the day with leading 0s, hh is the hour with leading zeros, mm are the minutes with leading zeros, and ZZZ is the three letter timezone abbreviation.

### protocols\_io\_doi

**Description:** DOI for protocols.io referring to the protocol for this assay. Note: any parameters of this protocol should be captured as Level 3 fields. In exceptional cases, some parameters might be included in Level 4 information.

**Allowed Values:** DOI (e.g. "10.17504/protocols.io.basdiea6")

### operator

**Description:** Name of the person responsible for executing the assay.

**Allowed Values:** Any string.

### operator\_email

**Description:** Email address for the operator.

**Allowed Values:** Syntactically correct email address.

### pi

**Description:** Name of the principal investigator responsible for the data.

**Allowed Values:** Any string.

### pi\_email

**Description:** Email address for the principal investigator.

**Allowed Values:** Syntactically correct email address.

### assay\_category

**Description:** Each assay is placed into one of the following 3 general categories: generation of images of microscopic entities, identification & quantitation of molecules by mass spectrometry, and determination of nucleotide sequence.

**Allowed Values:** "microscopy", "mass-spectrometry", "sequencing"

### assay\_type

**Description:** The specific type of assay being executed.

**Allowed Values:** Will vary by assay\_category:

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

### analyte\_class

**Description:** Analytes are the target molecules being measured with the assay. Some assays target more than one analyte. For example, mass spectrometry assays such as MALDI IMS will detect peptides and lipids within the same experiment, as will DESI IMS.

**Allowed values:** A single analyte or a comma-delimited list of analytes. Categories of analytes : “peptide”, "protein", "RNA", "DNA", "lipid", "metabolite" .

### is\_targeted

**Description:** Specifies whether or not a specific molecule(s) is/are targeted for detection/measurement by the assay.

**Allowed Values:** "True" or "False".

## Level 2 Attributes Section (here: “imaging”)

Different categories of assays (*imaging*, *mass spectrometry*, *sequencing*) will use different sets of Level 2 attributes. **For this preview, we only cover those attributes shared by all *imaging* assays.**

### acquisition\_instrument\_vendor

**Description:** An acquisition\_instrument is the device that contains the signal detection hardware and signal processing software. Assays generate signals such as light of various intensities or color or signals representing molecular mass.

**Allowed Values:** Akoya, Zeiss,10xGenomics, ...

### acquisition\_instrument\_model

**Description:** Manufacturers of an acquisition instrument may offer various versions (models) of that instrument with different features or sensitivities. Differences in features or sensitivities may be relevant to processing or interpretation of the data.

**Allowed Values:** Any string, but will be normalized during curation.

### resolution\_x\_value

**Description:** The width of a pixel.

**Allowed Values:** Any positive floating-point number.

### resolution\_x\_unit

**Description:** The unit of measurement of width of a pixel.

**Allowed Values:** "cm", "mm", "um", "nm", "pm" (Note: not "µm")

### resolution\_y\_value

**Description:** The height of a pixel.

**Allowed Values:** Any positive floating-point number.

### resolution\_y\_unit

**Description:** The unit of measurement of height of a pixel.

**Allowed Values:** "cm", "mm", "um", "nm", "pm" (Note: not "µm")

### resolution\_z\_value

**Description:** Optional if assay does not have multiple z-levels. Note that this is resolution within a given sample: z-pitch (resolution\_z\_value) is the increment distance between image slices. The thickness of the sample itself is sample metadata.

**Allowed Values:** Any positive floating-point number.

### resolution\_z\_unit

**Description:** The unit of incremental distance between image slices.

**Allowed Values:** "cm", "mm", "um", "nm", "pm" (Note: not "µm").

## Level 3 Attributes Section (here: “CODEX”)

Different types of assays will use different subsets of Level 3 Fields. **For this preview, we only cover those attributes shared by all CODEX assays.**

### preparation\_instrument\_vendor

**Description:**

**Allowed Values:** Akoya, ...

### preparation\_instrument\_model

**Description:**

**Allowed Values:** free text-for now

### number\_of\_antibodies

**Description:** Number of antibodies

**Allowed Values:** Integer

### number\_of\_cycles

**Description:** Number of cycles of 1. oligo application, 2. fluor application, 3. washes

**Allowed Values:** Integer

### section\_prep\_protocols\_io\_doi

**Description:** DOI for protocols.io referring to the protocol for preparing tissue sections for the assay.

**Allowed values:** A valid DOI (e.g. "10.17504/protocols.io.basdiea6")

### reagent\_prep\_protocols\_io\_doi

**Description:** DOI for protocols.io referring to the protocol for preparing reagents for the assay.

**Allowed values:** A valid DOI (e.g. "10.17504/protocols.io.basdiea6")

## Level 4 Reference Section

### metadata\_path

**Description:** Relative path to file or directory with free-form or instrument/lab specific metadata. *Optional.*

**Allowed Values:** Path required to exist on globus, and should be distinct from any data\_path.

## Output Reference Section

### data\_path

**Description:** Relative path to file or directory with instrument data. Downstream processing will depend on filename extension conventions. *Required.*

**Allowed Values:** Path required to exist on Globus, and should be distinct from any metadata\_path or other data\_path. All the files of a submission should be covered by exactly one metadata\_path or data\_path.

# Appendix: Implementation for HuBMAP Data Portal

The required columns and allowed values will be defined by a set of JSON schema, e.g. checked into the [ingest-pipeline](https://github.com/hubmapconsortium/ingest-pipeline) repository. From those we will generate human-readable documentation (which will replace this document) and TSV templates. Each type of assay will have its own template, so the schema will have some means of indicating which attribute goes with which assay type.

Allowed values should be defined via controlled vocabularies or ontologies.  
  
The [cidc-schemas](https://github.com/CIMAC-CIDC/cidc-schemas) project may provide some of the functionality we need, though it might not handle sub-typed schema like we will have.  
  
Submissions will be TSV using UTF-8. TSV because that is easier to hand edit than CSV and UTF-8 because some personnel names include non-ASCII characters.   
  
On ingest, the TSV will be translated to nested JSON: JSON schema doesn't adhere to an OO model, so a single flat space of attributes which can be extended in different directions is difficult to validate. There will be additional, programmatic checks:

* Are the referenced HuBMAP IDs valid?
* Do metadata\_path and data\_path point to valid files in the submission?
* Are all files in the submission covered, and none of them are referenced multiple times?
* Do the Level 2 and Level 3 attributes match the provided assay\_typeand assay\_category (these are Level 1 attributes)

If validation fails, error messages will indicate the coordinates of the error relative to the spreadsheet (i.e., rows and columns) and ideally suggest how the errors could be fixed.

Before being passed to Neo4J to be loaded into the provenance graph, and eventually to Elasticsearch, the JSON will go through a second transform. This transform will

* Treat each line in the TSV as its own record.
* Flatten the nested structure that was used for validation.
* Flatten down the provenance chain, so an assay record will also have information about the donor and sample it comes from.
* Include the provenance as an additional attribute, along with metadata coming from the submission event itself. (The newly generated ID, the submitter, the datetime of submission.)