# JSON data hub informal decription

A formal definition is at: https://github.com/IHEC/ihec-ecosystems/blob/master/JSON_Data_Hub_Validator/data_hub_schema.json

A JSON data hub contains three main sections:

```
{
    "hub_description": { ... },
    "datasets": { ... },
    "samples": { ... }
}
```

## 1. hub_description

Gives general information about the content of the hub.

```
    "hub_description": {
        "taxon_id": ...,
        "assembly": "...",
        "publishing_group": "...",
        "email": "...",
        "date": "...",
        "description": "...",
        "description_url": "...",
    }
```

- **taxon_id**: Species taxonomy id. (e.g. human = 9606)
- **assembly**: UCSC Reference genome assembly ID (e.g. human = h19 or hg38)
- **publishing_group**: IHEC member consortium that published this data hub.
  - Controlled vocabulary: `["Blueprint", "CEEHRC", "CREST", "DEEP", "ENCODE", "KNIH", "NIH Roadmap"]`
- **email**: Contact email for this data hub publishing group.
- **date**: Data hub release date, in ISO 8601 format.
- **description**: *(optional)* A description of the hub content.
- **description_url**: *(optional)* A link to an HTML document describing the hub content.

## 2. samples

Defines the list of samples contained in this data hub.

Metadata properties are specified in the IHEC Metadata specification.
Required attributes are defined in the IHEC Ecosystem track hub specification.

```
"samples": {
    "sample_id_1": {
        "sample_ontology_uri": "...",
        "molecule": "...",
        "disease": "...",
        "disease_ontology_uri": "...",
        "biomaterial_type": "...",

        *Additional metadata depending on biomaterial type may be required. Please refer
    },
    "sample_id_2": {
        ...
    }
}
```

- **sample_id**: An identifying designation for the biological sample.
- **sample_ontology_uri**: Ontology term that links to sample ontology information. Depending on the biomaterial_type, will be either an UBERON or CL ontology term.
- **molecule**: The type of molecule that was extracted from the biological material. Include one of the following: total RNA, polyA RNA, cytoplasmic RNA, nuclear RNA, genomic DNA, protein, or other.
  - Controlled vocabulary: `["total RNA", "polyA RNA", "cytoplasmic RNA", "nuclear RNA", "genomic DNA", "protein", "other"]`
- **disease**: Free form field for more specific disease information. If dealing with a rare disease consider identifiability issues.
- **disease_ontology_uri**: Ontology term that links to disease ontology information. If dealing with a rare disease consider identifiability issues. The NCI metathesaurus term C0277545 "Disease type AND/OR category unknown" should be used for unknown diseases. Phenotypes associated with the disease should be submitted as DISEASE_ONTOLOGY_URIs (if available) or in the free form DISEASE attribute.
- **biomaterial_type**:
  - Controlled vocabulary: `["Cell Line", "Primary Cell", "Primary Cell Culture", "Primary Tissue"]`

## 3. datasets

Describes datasets obtained experimentally.

```
"datasets": {
    "experiment_1": {
        "sample_id": "...",
        "experiment_attributes": { ... },
        "analysis_attributes": { ... },
        "browser": { ... }
```

```
    },
    "experiment_2": {
        ...
    },
}
```

### 3.1 sample__id

Links to the sample object defined in section 2. Value can be either a string or an array of strings. Each string should match a key in the samples hash.

### 3.2 experiment__attributes

Defines the experiment-related metadata properties for this dataset.

Metadata properties are specified in the IHEC Metadata specification.
Required attributes are defined in the IHEC Ecosystem track hub specification.

```
"experiment_attributes": {
    "experiment_type": "...",
    "assay_type": "...",
    "experiment_ontology_uri": "...",
    "reference_registry_id": "..."
}
```

- **experiment__type**: Must be one of the experiment types defined in the Metadata Standards document. (e.g. “DNA Methylation”, “mRNA-Seq”, “ChIP-Seq Input”)
- **assay__type**: As described in the experiment_ontology_uri term. (e.g. ‘DNA Methylation’)
- **experiment__ontology__uri**: Ontology term that links to experiment ontology information.
- **reference__registry__id**: The IHEC Reference Epigenome registry ID for this dataset, assigned after submitting to EpiRR.

### 3.3 analysis__attributes

Defines the bioinformatics analysis-related metadata properties for this dataset.

Metadata properties are specified in the IHEC Metadata specification.
Required attributes are defined in the IHEC Ecosystem track hub specification.

```
"analysis_attributes": {
    "analysis_group": "...",
    "alignment_software": "...",
    "alignment_software_version": "...",
```

```
            "analysis_software": "...",
            "analysis_software_version": "..."
        }
```

- **analysis_group**: The group that ran the bioformatics analysis to produce this dataset tracks
- **alignment_software**: The name of the software used for mapping.
- **alignment_software_version**: The version of the software used for mapping.
- **analysis_software**: The name of the software used for determining signal (read density).
- **analysis_software_version**: The version of the software used for determining signal (read density).

### 3.4 browser:

Points to data tracks for the experiment. Object keys represent the track type.

```
"browser": {
    "signal_forward": [
        {
            "big_data_url": "...",
            "description_url": "...",
            "md5sum": "...",
            "subtype": "...",
            "sample_source": "...",
            "primary":
        },
        {
            ...
        }
    ],
    "signal_reverse": [
        {
            ...
        }
    ]
}
```

- For each 'track type' key under 'browser', the following properties can be provided:
  - **big_data_url**: The URL from where this dataset track can be obtained online.
  - **description_url**: The URL of the document giving more information about this dataset track.
  - **md5sum**: The checksum for this track.

  - **subtype**: *(optional if there's only one track for this track_type)* If there are multiple files for this track type, use this free text field to put more information about what kind of information this track represents.
  - **primary**: *(optional if there's only one track for this track_type)* When there are multiple tracks for this track type, set this field to 'true' to express that this is the primary track to represent this track type.
  - **sample_source**: *(optional)* Use this field if the track belongs to only one sample of a merged dataset.

- Any key is supported, however, only keys corresponding to required track types as defined in the minimum required track types document are read.
- If a track is stranded (forward or reverse), the opposite strand track also needs to be provided.
- If multiple tracks are of the same track type, the first track in the list shall be the "main" one, meaning it is the one that best represents this sample for this track type.

**Notes:**

- Note that the the format is extensible. You can annotate your data, and include data way beyond the specification.
- For examples, see:
  - https://github.com/IHEC/ihec-ecosystems/tree/master/BCGSC_CEMT/Templater/examples
  - https://github.com/IHEC/ihec-ecosystems/blob/master/JSON_Data_Hub_Validator/example1.json