Annotation and Clinical Data

Table of Contents:

- Variables and Annotations
  - Variables
    - Examples
  - Annotations
    - Examples

Variables and Annotations

Clinical Data in OpenCGA is managed through what we have called *Variable Sets* and *Annotation Sets*.

Variables

A *Variable Set* is a free modelled data model. The fields of a *Variable Set* are explained below:

- **id:** Unique String that can be used to identify the defined *Variable Set*.
- **unique:** Boolean indicating whether there can only exist one single *Annotation Set* annotating the *Variable Set* per each *Annotable** entry or not. If false, many *Annotation Sets* annotating the same *Variable Set* per *Annotable* entry will be allowed.
- **confidential:** Boolean indicating whether the *Variable Set* as well as the *Annotation Sets* annotating the *Variable Set* are confidential or not. In case of confidentiality, only the users with that CONFIDENTIAL permission will be able to access it.
- **description:** String containing a description of the *Variable Set* defined.
- **variables:** List containing all the different *Variables* that will form the *Variable Set.* Explained in detail below.

* *Annotable*: We consider an entry to be *Annotable* if the entry can have *Annotation Sets.* At this stage, only *File*, *Sample*, *Individual*, *Cohort* and *Family* are *Annotable.*

** Confidential: Explained in Sharing and Permissions section !

A *Variable Set* is composed of a set of *Variables.* A *Variable* can be understood as a user-defined field that can be of any type (Boolean, String, Integer, Float, Object, List...). The different fields of a *Variable* are:

- **id:** String containing the unique identifier of the field (*Variable*) defined by the user.
- **name:**Nice identifier of the *name*. This field is intended to be used in a web application to show the field *name* in a nicer way.
- **category:** Free String that can contain anything useful for the user to group and categorise *Variables* of a same *Variable Set.*
- **type:** Type of the field (*Variable*) defined. It can be one of BOOLEAN, CATEGORICAL*, INTEGER, DOUBLE, TEXT, OBJECT.
- **defaultValue:** Object containing the default value of the *Variable* in case the user has not given any value when creating the *Annotation Set*.
- **required:** Boolean indicating whether the field is mandatory to be filled in or not.
- **multivalue:** Boolean indicating whether the field being annotated is a List of type *type* or it will only contain a single value.
- **allowedValues:** A list containing all the possible values a field could have.
- **rank:** Integer containing the order in which the annotations will be shown (only for web purposes).
- **dependsOn:** String containing the *Variable* the current *Variable* would depend on. Let's say we have defined two different *Variables* in a *Variable Set* called *country* and *city.* We can decide that we could only give a value to *city* once the *country* have been filled in, so *city* would depend on *country.*
- **description:** String containing a description for the *Variable.*
- **variableSet:** List of *Variables* that would only be used if the *Variable* being modelled is of type *Object.* Every *Variable* from the list will have the fields explained in this list.

* *Categorical*: A *Categorical* variable can be understood as an Enum object where the possible values that can be assigned are already known. Example of some categorical *Variables* are: *month,* that can only contain values from January to December, *gender,* that could only contain values from MALE, FEMALE, UNKNOWN; etc.

Examples

We are going to create two different *Variable Sets*, remember that the Variable Sets are defined at *study* level. The first one will be used to properly identify every single *Individual* created in OpenCGA. The other one will be used to store some additional metadata from the *Samples* extracted from the *Individuals.*

Individual Variable Set

```
{
    "id": "individual_private_details",
    "unique": true,
    "confidential": true,
    "description": "Private details of the individual",
    "variables": [
    {
```

```
        "id": "full_name",
        "name": "Full name",
        "category": "Personal",
        "type": "TEXT",
        "defaultValue": "",
        "required": true,
        "multiValue": false,
        "allowedValues": [],
        "rank": 1,
        "dependsOn": "",
        "description": "Individual full name",
        "attributes": {}
    },
    {
        "id": "age",
        "name": "Age",
        "category": "Personal",
        "type": "INTEGER",
        "required": true,
        "multiValue": false,
        "allowedValues": [
        "0:120"
        ],
        "rank": 2,
        "dependsOn": "",
        "description": "Individual age",
        "attributes": {}
    },
    {
        "id": "gender",
        "name": "Gender",
        "category": "Personal",
        "type": "CATEGORICAL",
        "defaultValue": "UNKNOWN",
        "required": true,
        "multiValue": false,
        "allowedValues": [
        "MALE",
        "FEMALE",
        "UNKNOWN"
        ],
        "rank": 3,
        "dependsOn": "",
        "description": "Individual gender",
        "attributes": {}
    },
    {
        "id": "hpo",
        "name": "HPO phenotypes",
        "category": "Disease",
        "type": "TEXT",
        "defaultValue": "",
        "required": true,
        "multiValue": true,
        "allowedValues": [],
        "rank": 4,
        "dependsOn": "",
        "description": "Individual HPO terms",
        "attributes": {}
    },
    {
        "id": "address",
        "name": "Address",
        "category": "Personal",
        "type": "OBJECT",
        "required": false,
        "multiValue": false,
        "allowedValues": [],
        "rank": 5,
        "dependsOn": "",
        "description": "Individual country of birth",
```

```
        "attributes": {},
        "variableSet": [
        {
            "id": "city",
            "name": "City",
            "category": "Personal",
            "type": "TEXT",
            "defaultValue": "",
            "required": false,
            "multiValue": false,
            "allowedValues": [],
            "rank": 1,
            "dependsOn": "",
            "description": "Individual city",
            "attributes": {}
        },
        {
            "id": "zip",
            "name": "ZIP code",
            "category": "Personal",
            "type": "TEXT",
            "defaultValue": "UNKNOWN",
            "required": false,
            "multiValue": false,
            "allowedValues": [],
            "rank": 2,
            "dependsOn": "city",
            "description": "ZIP code",
            "attributes": {}
        }
        ]
    }
    ]
}
```

Sample Variable Set


```
{
  "unique": true,
  "confidential": false,
  "id": "sample_metadata",
  "description": "Sample origin",
  "variables": [
    {
      "id": "tissue",
      "name": "Tissue",
      "category": "string",
      "type": "TEXT",
      "required": false,
      "multiValue": false,
      "allowedValues": [],
      "rank": 1,
      "dependsOn": "",
      "description": "Sample tissue",
      "attributes": {}
    },
    {
      "id": "cell_line",
      "name": "Cell line",
      "category": "string",
      "type": "TEXT",
      "required": false,
      "multiValue": false,
      "allowedValues": [],
      "rank": 2,
      "dependsOn": "",
      "description": "Sample cell line",
      "attributes": {}
    },
    {
      "id": "cell_type",
      "name": "Cell type",
      "category": "string",
      "type": "TEXT",
      "required": false,
      "multiValue": false,
      "allowedValues": [],
      "rank": 3,
      "dependsOn": "",
      "description": "Sample cell type",
      "attributes": {}
    },
    {
      "id": "preparation",
      "name": "Preparation",
      "category": "string",
      "type": "TEXT",
      "required": false,
      "multiValue": false,
      "allowedValues": [],
      "rank": 4,
      "dependsOn": "",
      "description": "Sample preparation",
      "attributes": {}
    }
  ]
}
```


Annotations

An *Annotation Set* is the set of *Annotations* given for a concrete *Annotable* entry using a particular *Variable Set* template. The most important fields of an *Annotation Set* are:

- **id:** Unique name to identify the annotation set created.
- **variableSetId:**Unique value identifying the *Variable Set* the *Annotation Set* is using to define the *Annotations.*
- **annotations:** List of *Annotations* or, in other words, values assigned for each *Variable* defined in the *Variable Set* corresponding to the *variableSetId.*

The *Annotations* are just key-value objects where each key need to match any of the *Variable names* defined in the *Variable Set,* and the values will correspond to the actual *Annotation* of the *Variable.*

Every time an annotation is made, OpenCGA will make, at least, the following checks:

- The data types of the *Annotations* match the types defined for the *Variables.*
- No mandatory *Variable* is missing an *Annotation*.
- The value for a particular *Variable* matches any of the allowed values if this array is provided and non empty.

Examples

An *Annotation* example for both *Variable Sets* examples can be found below:

Individual Annotation Set

```
{
  "id": "annotation_set_id",
  "variableSetId": "individual_private_details",
  "annotations": {
    "full_name": "John Smith",
    "age": 60,
    "gender": "MALE",
    "hpo": ["HP:0000118", "HP:0000220"]
  }
}
```

Sample Annotation Set

```
{
  "id": "annotation_set_id",
  "variableSetId": "sample_metadata",
  "annotations": {
    "tissue": "        umbilical cord blood",
    "cell_type": "multipotent progenitor",
    "preparation": "100 (or less, if 100 were not available) highly
purified Haematopoietic stem and progenitor cells..."
  }
}
```