Sample information

Content

There is a minimal amount of information that is required for human sample information. Basically the sample information is key/value pair information.

The information is divided to mandatory, recommended, and user defined information:

Mandatory information

sample_id: This an internal id for your sample record. It must be unique to one sample record for a provider. Example is ‘s123’. The value must single word.

patient_id: Each patient needs to have an unique id that doesn’t contain their name, patient_id shall be only identifiable by the provider. Example is ‘pa123’ . The value must single word.

provider: This is the data provider. Example is ‘bornagene’.

organism: In this case the value is ‘homo sapiens’ or ‘human’.

sex: Gender or sex of the patient. The values are ‘male’ or ‘female’. The value must single word.

race: Race of the patients need to be recorded. Example is ‘kurd’. The value must single word.

title: A title that represent the sample.
description: Any description related to patient, sample record, and to be performed experiment.

organism_part: The part of organism that the sample has been taken. Example is ‘blood’.

age: age of host at the time of sampling. Example is ’55 years and 3 months’.

health_state: Health or disease status of the sample at the time of collection. If the data isn’t available for this please use ‘NA’

tested_disease: List of diseases that has screened; can include multiple diagnoses. The value of the field depends on host; for humans the terms should be chosen from DO (Disease Ontology), free text for non-human. The list shall be separated by ‘,’. Example is ‘disease1, disease2’.

disease: List of diseases diagnosed; can include multiple diagnoses. The value of the field depends on host; for humans the terms should be chosen from DO (Disease Ontology), free text for non-human. The list shall be separated by ‘,’. Example is ‘disease1, disease2’. If there is no disease available use ‘NA’.

phenotype: Phenotype of sampled organism. If it is not available, use ‘NA’.

collection_date: The date of sampling, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated i.e. all of these are valid ISO8601 compliant times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008.

smoker: specification of smoking status. Can be ‘NO’, ‘YES’, or ‘YES-NUMBER’ (to provide number of cigarette per day). Example is ‘YES-12’, means this person smoke 12 cigarettes per day.

family_id: Each family needs to have an unique id that doesn’t contain their name, family_id shall be only identifiable by the provider. Example is ‘fam123’. The value must single word. Family_id has been requested to be able to group all the samples that are part of a family and have some sort of family relationship.

family_relationship: Family_relationship has been requested to uncover the family relationships among the samples for further and future data analysis. Please use one of the family members as ‘subject’, ‘subject’ is the main person and we would strongly recommend using the first person that has been sequenced as ‘subject’. Then for all other samples in ‘subject’ family use the relationship to the subject. Example:

sample_id	patient_id	family_id	family_relationship	sex
sam1	pa23	fam1	mother	female
sam2	pa2	fam1	subject	male
sam4	pa5	fam1	father	male
sam5	pa7	fam2	subject	female
sam7	pa8	fam2	mother	female
sam8	pa18	fam3	subject	female

In this example 6 people has been sequenced from three different family, in family 1 (fam1), father, mother, and son have been sequenced. In family 2 (fam2), a mother and daughter have been sequenced. In family 3 (fam3), at this point of time only one female has been sequenced, but in case the new member of family will be sequenced you can follow the instruction and use fam3 as this family_id and provide their sample information.

Experiment_id: Please see section 2 (Experiment information). Please provide the experiment_id from the sequencing experiment that has been performed to sequence the given sample record.

is_tumor: ‘YES’ or ‘NO’

occupation: most frequent job performed by subject. It must be Single-line text.

Recommended information

lat_lon: It is geographical coordinates of the location where the specimen was collected. It must be Single-line text.
Examples are:

lat_lon:"47.94 N 28.12 W"

lat_lon:"45.0123 S 4.1234 E"

Degrees latitude and longitude in format "d[d.dddd] N|S d[dd.dddd] W|E"

collected_by: name of persons or institute who collected the specimen collecting institution.

User defined information

Based on the your lab requirement, you can add your defined attribute to the sample record. Example, some may would want to record family history and add '**family_history**' as a new key to the records the family history information in Single-line text.

Format

The accepted format is tab-delimitated file. The first line shall contain the list of keys. From the second line the values for corresponding keys needs to be provided. Each line shall represent one sample record.

A reduced keys example is:

sample_id	patient_id	organism	sex	age	tested_disease
sam12	p11	human	male	38	clone cancer, heart disease
s18	pa165	human	female	55	breast cancer

A complete example has been provided in **sample.txt** file.