FAANG metadata - sample specification

This document describes the specification for all sample metadata. You can find an overview of our metadata and archival plans in the overview document. The experiment and analysis documents are also in this git repo. Further guidance can be found on the FAANG wiki pages.

In the sample context, we consider donor animals, tissue samples, primary cells or other biological material to be samples. All Samples must be registered in BioSamples at EMBL-EBI as this samples archive has the best support for 'child of' and 'derived from' sample relationships. The NCBI BioSample database is a peer of the EMBL-EBI BioSamples, and they exchange data regularly. FAANG samples should be registered in the EMBL-EBI BioSamples prior to data submission. This document describes the attributes which must be associated with any BioSamples submission.

Sample metadata requirements

Most requirements are laid out like this:

- `attribute name` (*data type*) a brief description

The data types are described later in this document.

Common

These attributes should be present on every sample record.

Required:

- `Sample name` (*text*) sample names should follow the naming rules listed below. Each name must be unique.

- `Material` (*ontology term*) the type of material being described. This will be used to decide what metadata are required and must be one of the expected terms:
  - organism
  - specimen from organism
  - cell specimen

  - cell culture
  - pool of specimens
  - cell line

- `project` (*text*) project name - this should always be 'FAANG'. This will allow the DCC to identify FAANG samples

Optional:

- `Sample Description` (*text*) a brief description of the sample including the species name
- `availability` (*URL*) either a link to a web page giving information on sample availability (who to contact and if the sample is available), or a e-mail address to contact about availability. E-mail addresses should be prefixed with 'mailto:', e.g. 'mailto:samples@example.ac.uk'. In either case, long term support of the web page or e-mail address is necessary. Group e-mail addresses are preferable to indiviudal.

Links to other records:

- `Same as` (*sample*) BioSample ID for an equivalent sample record, created before the FAANG metadata specification was available. This is optional and not intended for general use, please contact the data coordination centre (faang-dcc@ebi.ac.uk) before using it.

Animal

An animal sampled for FAANG. The following attributes are in addition to the attributes listed in the 'Common' section above. The `material` should be reported as organism.

Required:

- `Organism` (*NCBI taxon ID*)
- `Sex` (*ontology term*) animal sex, described using any child term of PATO_0000047
- `breed` (*ontology term*) animal breed, described using the FAANG breed description guidelines and Livestock Breed Ontology

Recommended:

- `birth date` (*date*) birth date, in the format YYYY-MM-DD, or YYYY-MM where only the month is known. For embryo samples, record 'not applicable'
- `health status` (*ontology term*) Healthy animals should have the term normal, otherwise use the as many disease terms as necessary from EFO

Optional:

- `birth location` (*text*) name of the birth location
- `birth location latitude` (*number*) latitude of the birth location in decimal degrees. Units should be specified as 'decimal degrees'
- `birth location longitude` (*number*) longitude of the birth location in decimal degrees. Units should be specified as 'decimal degrees'
- `birth weight` (*number*) weight, in kilograms or grams. Units must be specified
- `placental weight` (*number*) weight, in kilograms or grams. Units must be specified.
- `pregnancy length` (*number*) length of time, in days, weeks or months
- `delivery timing` (*ontology term*) possible values
  - early parturition
  - full-term parturition
  - delayed parturition
- `delivery ease` (*text*) possible values
  - normal autonomous delivery
  - c-section

  - vetinarian assisted

- `pedigree` (*URL*) a link to pedigree information for the animal

Phenotypic, physiological and environmental information can be recorded using as many terms as necessary from the
VT, ATOL, EOL and MP ontologies. Select an appropriate term and units for the information you wish to record. For example, for growth rate you could use the attribute name `post natal growth rate` and units of 'grams per day'.

Links to other records:

- `Child of` (*sample*) sample name or Biosample ID for sire/dam. Required if related animals are part of FAANG, e.g. quads.

Specimen

A piece of tissue taken from an animal. The following attributes are in addition to the attributes listed in the 'Common' section above. The `material` should be reported as specimen from organism.

Required:

- `specimen collection date`(*date*) date at which the specimen was collected

- `animal age at collection` (*number*) animal age at the point of collection, in years, months, weeks or days. Units must be specified. An estimate is acceptable where the age is not precisely known.

- `developmental stage` (*ontology term*) a child term of life cycle stage

- `organism part` (UBERON term preferred)

- `specimen collection protocol` (*protocol*) a link to the protocol followed when taking the specimen

Recommended:

- `health status at collection` (*ontology term*) Animal disease / health status at point of collection. Healthy animals should have the term normal, otherwise use the as many disease terms as necessary from EFO

Optional:

- `fasted status` - (*text*) One of the following values, for which the criteria *must* be specified in the protocol:
  - fed
  - fasted
  - unknown

- `number of pieces` (*number*) Units must be specified as 'count'

- `specimen volume` (*number*) Units must be specified as either 'square centimeters', 'liters' or 'milliliters'

- `specimen size`(*number*) Units must be specified as either 'meters', 'centimeters', 'millimeters', 'square meters', 'square centimeters', or 'square millimeters'

- `specimen weight` (*number*) Units must be specified as either 'grams', 'kilograms'

- `specimen picture url` (*URL*) Link to a picture of the specimen

- `gestational age at sample collection` (*number*) If the animal was pregnant when the specimen was taken, state how long had it been pregnant for. Units must be specified as 'days' or 'weeks'.

Phenotypic, physiological and environmental information can be recorded using as many terms as necessary from the
VT, ATOL, EOL and MP ontologies. Select an appropriate term and units for the information you wish to record. For example, for growth rate you could use the attribute name `post natal growth rate` and units of 'grams per day'.

Links to other records:

- `Derived from` (*sample*) sample name or BioSample ID for an *animal* record (required).

Pool of specimens

Each specimen within the pool should have its own complete specimen record. The sample names (if detailed in the same file), or BioSample IDs if they already exists in the BioSamples database, are recorded in multiple 'derived from' fields. As many 'derived from' fields as are required to record all of the specimens that are part of the pool can be included.

Required:

- `pool creation date`(*date*) date at which the pool of specimens was created
- `pool creation protocol` (*protocol*) a link to the protocol followed when creating the pool

Optional:

- `specimen volume` (*number*) Units must be specified as either 'square centimeters', 'liters' or 'milliliters'
- `specimen size`(*number*) Units must be specified as either 'meters', 'centimeters', 'millimeters', 'square meters', 'square centimeters', or 'square millimeters'
- `specimen weight` (*number*) Units must be specified as either 'grams', 'kilograms'
- `specimen picture url` (*URL*) Link to a picture of the specimen

Links to other records:

- `Derived from` (*sample*) specimen name or BioSample ID for a *specimen* record (required), multiple allowed.

Purified cells

Cells purified from a specimen. The following attributes are in addition to the attributes listed in the 'Common' section above. The `material` should be reported as cell specimen.

Required:

- `cell type` (*ontology term*) as many terms as necessary from the CL ontology

- `purification protocol` (*protocol*) protocol describing how the cells were purified

Optional:

- `markers` (*text*) markers used to isolate and identify the cell type (e.g. for FACS sorted cells)

Links to other records:

- `Derived from` (*sample*) sample name or BioSample ID for a *specimen* record (required).

Cell culture

Cells cultured from a specimen or purified cells. The following attributes are in addition to the attributes listed in the 'Common' section above. The `material` should be reported as cell culture.

Required:

- `culture type`(*ontology term*) a child term of BTO_0000214
- `cell type` (*ontology term*) a term from the CL ontology
- `cell culture protocol` (*protocol*) protocol describing how the cells were purified
- `culture conditions` (*text*) brief description of culture conditions (e.g. 'on feeder cells', 'E8 media')
- `number of passages` (*number*) number of times the cell line has been re-plated and allowed to grow back to confluency or to some maximum density if using suspension cultures

Links to other records:

- `Derived from` (*sample*) sample name or BioSample ID for a *specimen* or *purified cell* record (required).

Cell line

A cultured cell population that represents a genetically stable and homogenous population of cultured cells that shares a common propagation history. The metadata requirements for this sample type are less stringent than for others, to allow for the level of detail normally available for established cell lines. The following attributes are in addition to the attributes listed in the 'Common' section above. The `material` should be reported as cell line

Required:

- `Organism` (*NCBI taxon ID*)
- `Sex` (*ontology term*) animal sex, described using any child term of PATO_0000047
- `cell line` (*text*) name of the cell line
- `biomaterial provider` (*text*) name of company or lab that supplied the cell line

Recommended:

- `catalogue number` (*text*) Identifier for the cell line in the suppliers catalogue. E.g. 'ACC 701' for IPEC-J2 from DSMZ.
- `passage number` (*number*) The number of times the cell line has been re-plated and allowed to grow back to confluency or to some maximum density if using suspension cultures.
- `date established` (*date*) date the line was established/re-established
- `publication` (*URL*) a publication where the cell line has been fully described including. This should include details such as doubling time and adhesion preference.

Optional:

- `breed` (*ontology term*) animal breed, described using the FAANG breed description guidelines and Livestock Breed Ontology
- `disease` (*ontology term*) a child term of either PATO_0000461 or EFO_0000408

- `cell type`(*ontology term*) a child term of either CL_0000000 or BTO_0000000
- `culture conditions` (*text*) brief description of culture conditions (e.g. 'on feeder cells', 'E8 media')
- `culture protocol` (*protocol*) protocol describing the maintenance of the culture
- `karyotype` (*text*) karyotype of the cell line

Links to other records:

- `Derived from` (*sample*) sample name or BioSample ID for the sample or animal the cell line was dervived from, where this is known and can be described within the FAANG standards (optional).

Pooled samples

Where samples are pooled, a new sample record should be created, containing

- `pooling protocol` (*protocol*)

Links to other records:

- `Derived from` (*sample*) sample name or BioSample ID for a *specimen*, *purified cells* or *cell culture* record (required).

Data types for sample attributes

BioSamples takes sample records with a set of attributes. Each attribute has a name and a value. It can also have 'Units', or a 'Term Source' and a 'Term Source ID'. The Term Source and ID allow us to refer to entries in other databases or ontologies. This is fully described on the BioSamples help pages. The following section describes the expectations for each data type within FAANG.

date

Dates should be reported in an ISO 8601 format, YYYY-MM-DD for dates or YYYY-MM for months. To ensure clarity, the format must be reported as the 'units'.

NCBI taxon ID

A species name and identifier from the NCBI Taxonomy database. For example, a human would be described with a value of 'Homo sapiens', a term source of 'NCBI Taxonomy' and a term source ID of 9606.

number

A number, with units specified. BioSamples recommends that units are given without abbreviations .For example, a birth weight could have a value of 1.3 and the units specified as 'kilograms'.

protocol

A URL link to a protocol document on the FAANG FTP site. Please contact the FAANG data coordination centre to have your protocol documents added to the FTP site.

text

Text, using US English spellings.

URL

A URL, such as 'http://faang.org/'. Depending on the context, http, ftp, mailto links may be appropriate. Examples:

- ftp, ftp://ftp.faang.ebi.ac.uk/ftp/README
- http, http://faang.org/
- mailto, mailto:bob@example.org

ontology term

A reference to an ontology term. The attribute value should be the term label. The term source should be the ontology used, and the term source ID should be an ID from that ontology. For example, cerebral cortex could be described with a term source of 'UBERON', a term source ID of 'UBERON:0000956' and a value of 'cerebral cortex'.

location

A location should be reported as using three attributes:

- `location` (*text*) name of the location
- `location latitude` (*number*) latitude in decimal degrees. Units should be reported as 'decimal degrees'
- `location longitude`(*number*) longitude in decimal degrees. Units should be reported as 'decimal degrees'

sample

Samples can be referred to in two ways. If the sample you need to reference is in the submission, use the sample name. If the sample was already submitted, use the BioSample ID (e.g. SAMEA2821491).

Missing data

Where data cannot be included in a submission, submit one of these text values instead

- 'not applicable' (i.e. does no apply to this sample)
- 'not collected' (i.e. will always be missing)
- 'not provided' (i.e. may be added later)
- 'restricted access' (i.e. it isn't missing, we just can't include it in a public document)

The use of these values will interact with the metadata validation system as follows:

- attribute is required
  - not applicable, not collected, not provided - validation will regard these as an error
  - restricted access - validation will generate a warning

- attribute is recommended
  - not collected, not provided - validation will generate a warning
  - restricted access, not applicable - pass
- attribute is optional
  - validation will pass with any of missing values terms

Sample naming

We propose a sample naming scheme comprising the following elements:

- short species code
- lab or institute short name
- alpha numeric sample ID from LIMS

The purpose is to ensure that samples are uniquely and clearly identified, with reasonably short names.

Short species codes:

- *Bubalus bubalis* BBU
- *Bos taurus* BTA
- *Sus scrofa* SSC
- *Ovis aries* OAR
- *Gallus gallus* GGA
- *Equus caballus* ECA
- *Capra hircus* CHR

Submission

Your submission should be prepared following the guidance on the FAANG wiki pages. This will guide you through:

- Downloading the empty Excel template to record your metadata
- Completing the template following the instructions and referring to the latest metadata rules specification. The rules for each attribute define if it is mandatory or optional, what sort of data is expected (numeric, date, text, etc.), what units are permitted, and whether or not an ontology term is required.
- Visiting the FAANG validation service where you can validate that your Excel complies with the metadata specifications.
- Resolving any errors or warnings that it provides, referring to the instructions and referring to the latest metadata rules specification for advice.
- Converting your template into SampleTab ready for submission using the FAANG conversion tool
- Samples should be submitted to BioSamples@EBI. All samples tagged with a `project` of 'FAANG' will be added to the FAANG BioSamples group. Samples in this group will be synced to BioSample@NCBI periodically. Samples in BioSamples@EBI/BioSample@NCBI can be referenced in submissions to SRA at EBI and NCBI.

The DCC team at EMBL-EBI will further check the submitted metadata against the specification. Samples that do not meet the minimum requirements will be not be included in FAANG data releases and will be marked as such in the FAANG data portal

Further guidance can be found on the FAANG wiki pages.