# GHGA Metadata: Quick Submission Guide

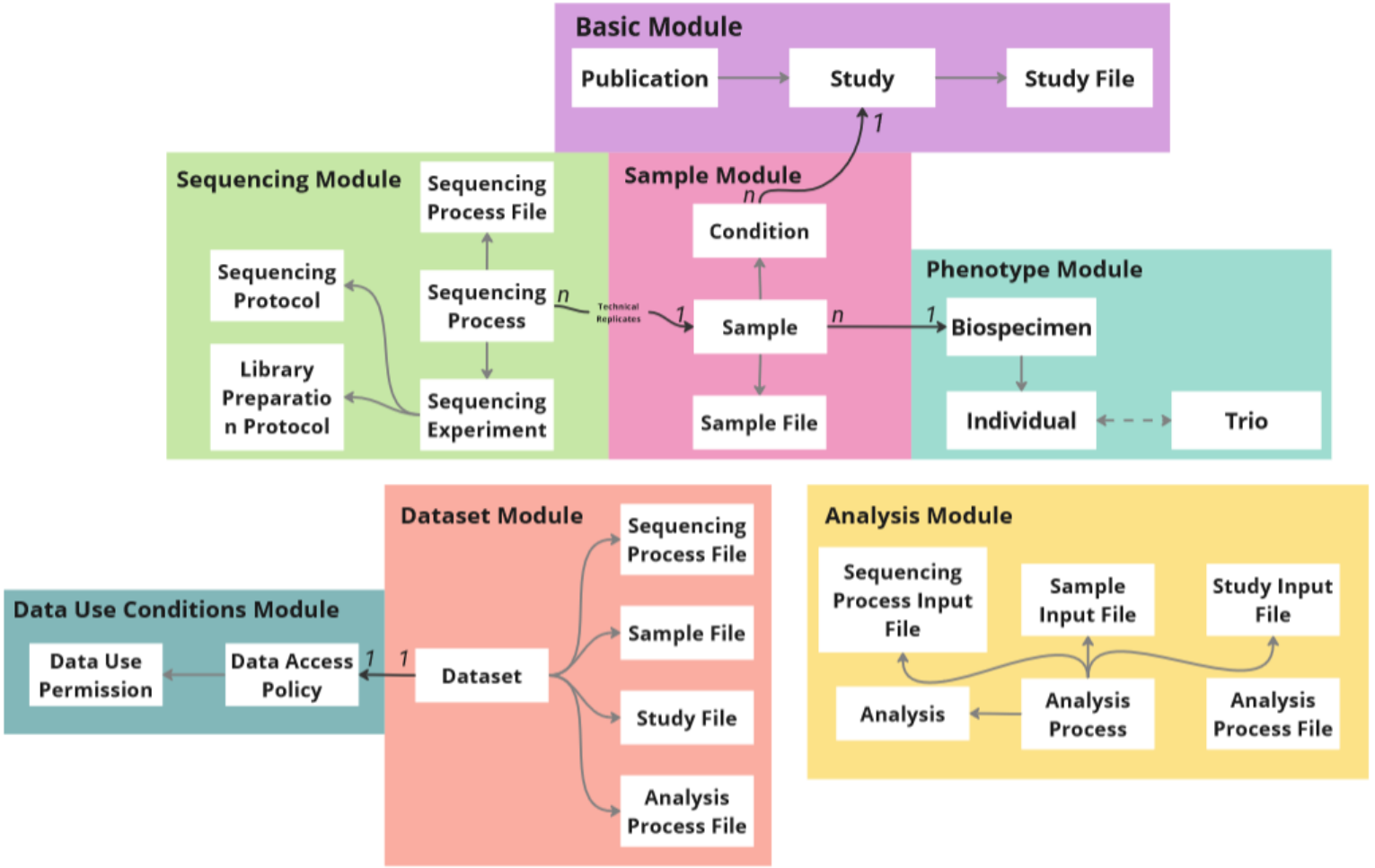
> v.1.1.1

# Resources

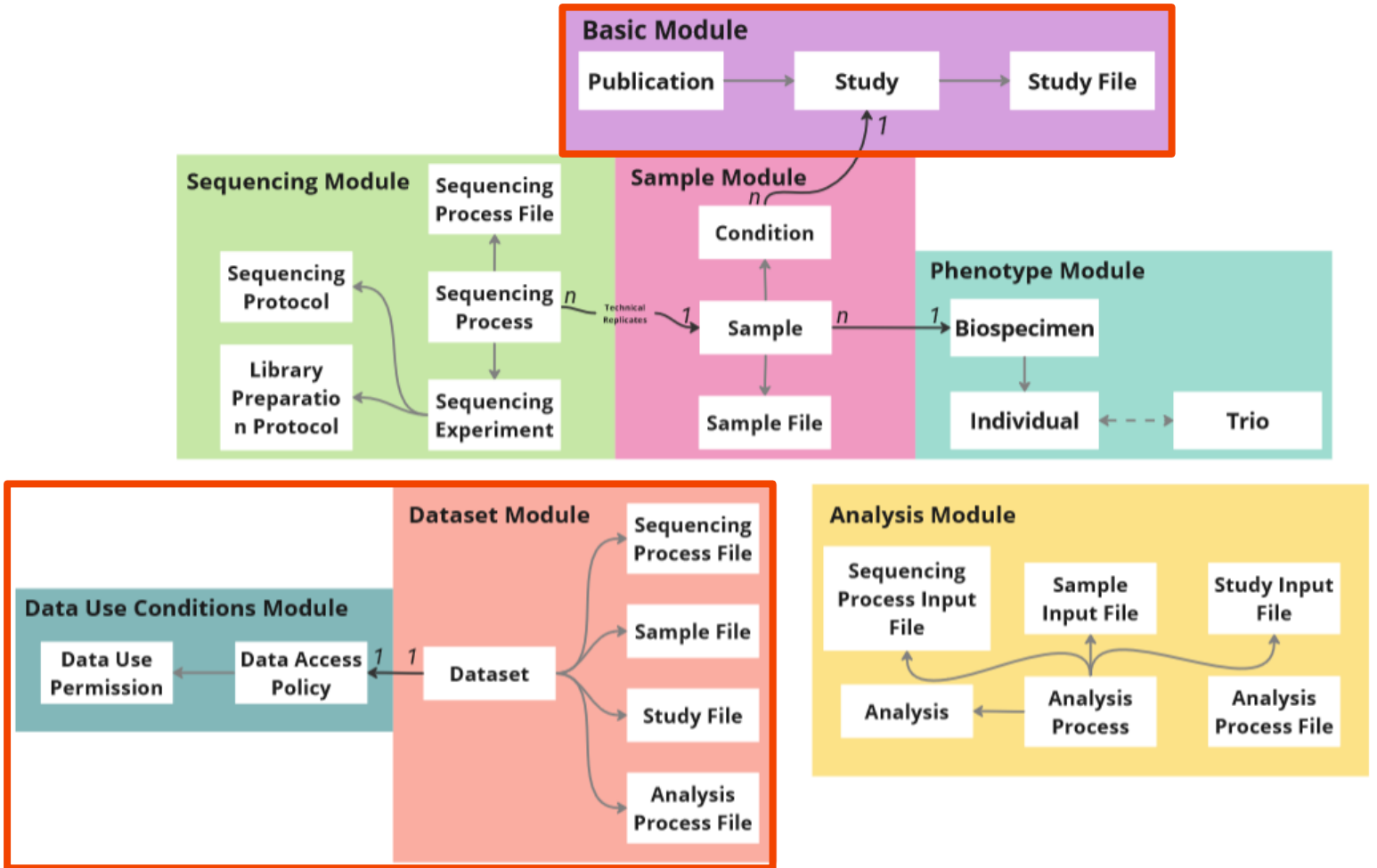
- Submission spreadsheets: https://github.com/ghga-de/ghga-metadata-schema/tree/main/spreadsheets
- JSON:
  https://github.com/ghga-de/ghga-metadata-schema/blob/main/artifacts/jsonschema/ghga.schema.json (Needs updated link!)
- Example spreadsheet/JSON:
  https://github.com/ghga-de/example-data
- Full documentation:
  https://ghga-de.github.io/docs/metadata/overview/

GHGA

# Metadata Schema > v.1.1.1



# Basic / Dataset module



# Basic / Dataset module

- *Basic:* Specifies study, file and publication metadata

- *Dataset:* Collects files in one or more datasets. Datasets should be curtated to contain files with similar experimental or phenotypic features (such as containing all WGS files of a submission or all related to the same tumor type)

- *Data Use Conditions/Policy:* Specifies the responsible contact persons for granting access to the data in form of a committee and under which conditions access is granted.

GHGA

# Study

| alias | title | description | type | affiliations | attributes |
|---|---|---|---|---|---|
| The alias for an entity at the time of submission. | A comprehensive title for the study. | A detailed description (abstract) that describes the goals of this Study. | The type of Study. For example, 'Cancer Genomics', 'Epigenetics', 'Exome Sequencing'. | The Institution(s) associated with an entity. | Custom key/value pairs that further characterizes the Study. (e.g.: approaches - single-cell,_bulk_etc) |
| type: string | type: string | type: string | type: string | type: string | type: string |
| single value | single value | single value | single value | multiple values | multiple values |
| unrestricted | unrestricted | unrestricted | controlled vocabulary | unrestricted | unrestricted |
| required | required | required | required | required | optional |
| STUDY_A | The A Study | A study that is the A study | SYNTHETIC_GENOMICS | Some Institute; Some other Institute | budget=3.5M;funding=EU |

- In order to describe a *Study*, data submitters are required to provide information about the study affiliation(s), title, description and type.
- An alias to link the study to different objects in the schema has to be provided
- A meaningful title and description of the study should be set up. Briefly explain what data is deposited why. Similarly, affiliations to the contributing institutions should be listed here separated by semicolons.
- Study.type is a broadly categorizes the study using a controlled vocabulary.

GHGA

# Study.File

| alias | study | name | format | size |
|---|---|---|---|---|
| The alias for an entity at the time of submission. | The study associated with an entity. | The given filename. | The format of the file: BAM, SAM, CRAM, BAI, etc. | The size of a file in bytes. |
| type: string | type: string | type: string | type: string | type: integer |
| single value | single value | single value | single value | single value |
| unrestricted | restriction: value from Study.alias | unrestricted | controlled vocabulary | unrestricted |
| required | required | required | required | required |
| FILE_1 | STUDY_A | SEQ_FILE_A_R1.fastq.gz | FASTQ | 1048599 |

| size | checksum | forward_or_reverse | checksum_type | dataset |
|---|---|---|---|---|
| The size of a file in bytes. | A computed value which depends on the contents of a block of data and which is transmitted or stored along with the data in order to detect corruption of the data. The receiving system recomputes the checksum based upon the received data and compares this value with the one sent with the data. If the two values are the same, the receiver has some confidence that the data was received correctly. | Denotes whether a submitted FASTQ file contains forward (R1) or reverse (R2) reads for paired-end sequencing. The number that identifies each read direction in a paired-end nucleotide sequencing reaction. | The type of algorithm used to generate the checksum of a file. | The Dataset associated with an entity. |
| type: integer | type: string | type: string | type: string | type: string |
| single value | single value | single value | single value | single value |
| unrestricted | unrestricted | controlled vocabulary | unrestricted | restriction: value from Dataset.alias |
| required | required | recommended | required | required |
| 1048599 | 6d893451b8ed4159c9c31693894fdf6bb4e37c40705b977f315e00d75190d71e | FORWARD | SHA256 | DS_1 |

- At the core of GHGA is the deposition of raw files that have been generated while carrying out an experiment. These files also have to be annotated with metadata, in order to give data requesters more information on what files have been deposited at GHGA by the data submitter.
- Study.File contains broad, file-centric metadata, such as the checksum, size and name to identify files correctly and a dataset alias that links it to a dataset in file.dataset.
- Study.File can be used to add files without further linked metadata-entities to a dataset.
- Format and forward_or_reverse use controlled vocabularies.

# Publication

| alias | title | abstract | author | year | journal | doi | study | xref |
|---|---|---|---|---|---|---|---|---|
| The alias for an entity at the time of submission. | The title for the Publication. | The study abstract that describes the goals. Can also hold abstract from a publication related to this study. | The individual who is responsible for the content of a document version. | Year in which the paper was published. | Name of the journal. | DOI identifier of the Publication. | The Study entity associated with this Publication. | One or more cross-references for this Publication. |
| type: string | type: string | type: string | type: string | type: integer | type: string | type: string | type: string | type: string |
| single value | single value | single value | single value | single value | single value | single value | single value | multiple values |
| unrestricted | unrestricted | unrestricted | unrestricted | unrestricted | unrestricted | unrestricted | restriction: value from Study.alias | unrestricted |
| required | optional | optional | optional | optional | optional | required | required | optional |
| PUB_1 | A paper of a study | This study aims finding findings. | John Doe | 1965 | Journal of Studies | 10.1234/abcd.5678 | STUDY_A | abcd_pubmed_link,pubmed |

- *Publication* is an optional metadata entity. If it is submitted, its properties become mandatory/optional.
- If no publication is present at the time of the submission, leave the entity empty.
- Publication captures general information about the related journal and publication specifics.
- A publication has to be linked to a study by entering the study.alias in the field publication.study

GHGA

# Data Access Committee (“DAC”)

| alias | email | institute |
|---|---|---|
| The alias for an entity at the time of submission. | Email of a person. | The institute a person is affiliated with. |
| type: string | type: string | type: string |
| single value | single value | single value |
| unrestricted | unrestricted | unrestricted |
| required | required | required |
| DAC_1 | dac@dac.dac | The DAC institute |

- The *DAC* entity bundles necessary information that is required to identify the Data Controller of the deposited data. Therefore a name and description for the *DAC*, and the main contact have to be provided upon submission. The information about a contact includes the email address and the associated affiliation.

- *DO NOT USE A PERSONAL EMAIL!*
- Submissions are stored for long periods of time. If a sole person is responsible the risk of deposited submissions with an unresponsive DAC is too high. Please, use an institutional/functional email to forward mails to members of the DAC.

# Data Access Policy (“DAP”)

| alias | name | description | policy_text | policy_url | data_access_committee | data_use_permission | data_use_modifiers |
|---|---|---|---|---|---|---|---|
| The alias for an entity at the time of submission. | A name for the Data Access Policy. | A short description for the Data Access Policy. | The terms of data use and policy verbiage should be captured here. | URL for the policy, if available. This is useful if the terms of the policy is made available online at a resolvable URL. | The Data Access Committee linked to this policy. | Data use permission associated with a policy. Typically one or more terms from DUO and should be descendants of 'DUO:0000001 data use permission'. | Modifier for Data use permission associated with a policy. Should be descendants of 'DUO:0000017 data use modifier' |
| type: string | type: string | type: string | type: string | type: string | type: string | type: string | type: string |
| single value | single value | single value | single value | single value | single value | single value | multiple values |
| unrestricted | unrestricted | unrestricted | unrestricted | unrestricted | restriction: value from DataAccessCommittee.alias | controlled vocabulary | controlled vocabulary |
| required | required | required | required | recommended | required | required | recommended |
| DAP_1 | DAP 1 | A Data Access Policy 1 | This is a very permissible DAP | http://some/policy | DAC_1 | disease specific research | clinical care use |

- A *DAP* is directly linked to the *DAC* and *Dataset* entity, thus providing the condition under which the data deposited at GHGA can be re-used by a data requester. The submitter must provide an alias, name, description and either the policy text for the *DAP* or the URL where the *DAP* is stored. The *DAP* needs to be linked to the *DAC* and *Dataset* by entering the related aliases in policy.data_access_committee and dataset.data_access_policy.

- To systematically and semantically identify the conditions under which deposited data can be reused, data submitters can optionally provide DUO terms that are used to identify the research purpose under which the data can be requested, e.g. General Research Use (DUO:0000042), research specific restrictions (DUO:0000012).
- The controlled vocabulary for DUO permission and modifier can be found here

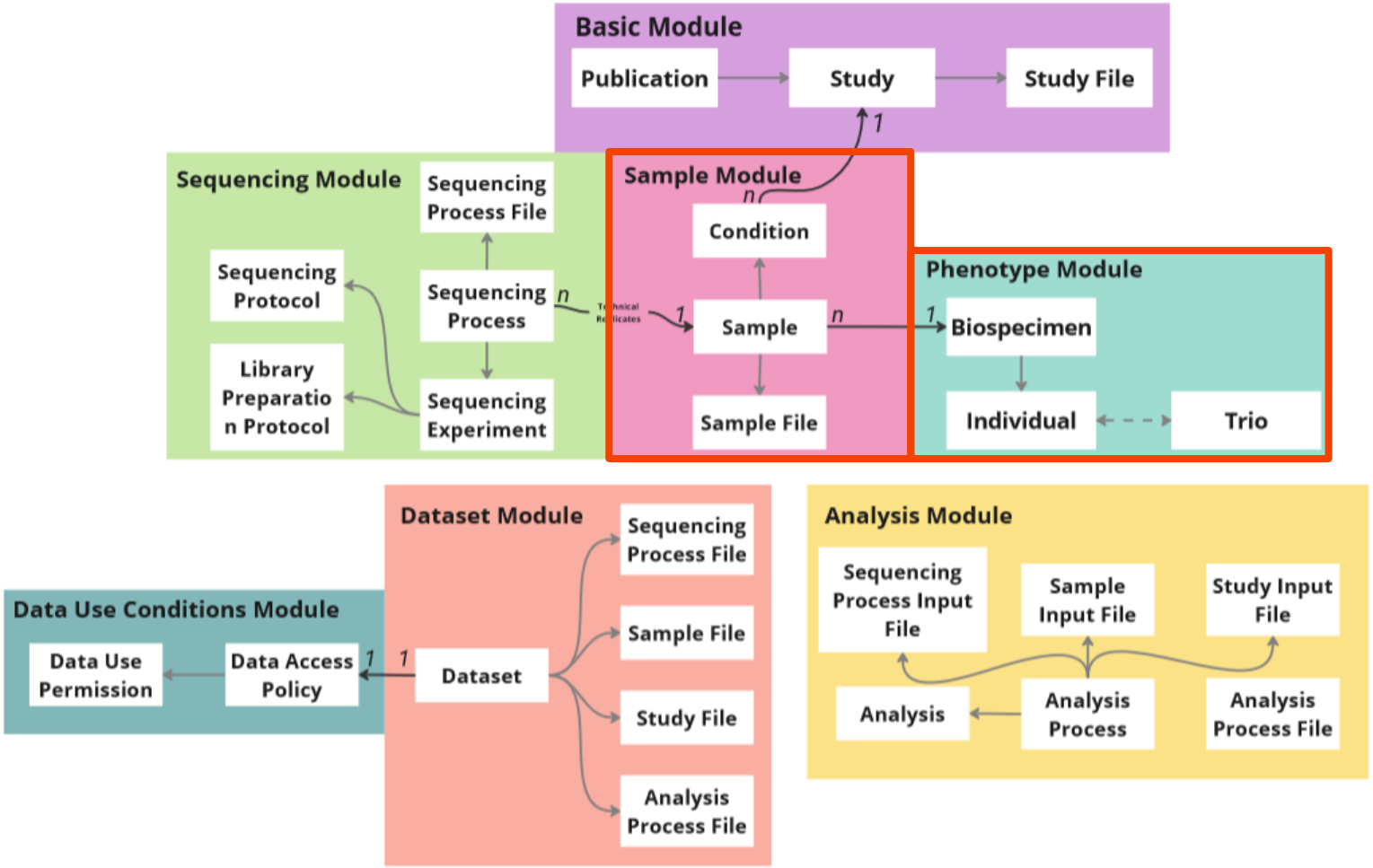
# Dataset

| alias | title | description | types | data_access_policy |
|---|---|---|---|---|
| The alias for an entity at the time of submission. | A title for the submitted Dataset. | Description of an entity. | The type of a dataset. | The Data Access Policy that applies to this Dataset. |
| type: string | type: string | type: string | type: string | type: string |
| single value | single value | single value | multiple values | single value |
| unrestricted | unrestricted | unrestricted | unrestricted | restriction: value from DataAccessPolicy.alias |
| required | required | required | required | required |
| DS_1 | The A dataset | An interesting dataset A | A Type; Another Type | DAP_1 |

- GHGA presents its content to potential data requesters with the *Dataset* entity, which focuses on sharing functionality by describing the contents at a high level. Each dataset is linked to a *Data Access Policy*, which builds the legal basis for the sharing of data. One dataset has links to *Experiment* and / or *Analysis* entities to bundle all relevant data that makes a dataset by the definition of the GHGA Metadata Schema.

- A minimal submission can be created as standalone, based on the modules *Basic*, *Data Use Conditions* and *Dataset*. This is meant to store legacy, unannotated data or files that do not fit anywhere else!

# Sample / Phenotype module




GHGA

# Sample / Phenotype module

- GHGAs *Sample* metadata can be separated into three distinct entities: *Sample*, *Biospecimen* and *Condition*.

- Both the *Sample* and *Biospecimen* entities provide the data submitter with options to deposit metadata that allows for deeper insight into the characteristics of samples and biospecimen.

- The *Condition* allows to further define the state of the samples and to group samples within a study accordingly. The following paragraph gives a definition of what a sample, biospecimen or condition is in the context of GHGAs metadata schema.

# Sample

| alias | name | type | description | isolation | storage | biospecimen | condition |
|---|---|---|---|---|---|---|---|
| The alias for an entity at the time of submission. | Name of the sample (eg:GHGAS_Blood_Sample1 or GHGAS_PBMC_RNAseq_S1). | The type of sample. | Short textual description of the sample (How the sample was collected, sample source, Protocol followed for processing the sample etc). | Method or device employed for collecting/isolating a biospecimen or a sample. | Methods by which a biospecimen or a sample is stored (e.g. frozen in liquid nitrogen). | The Biospecimen from which this Sample was prepared from. | The condition associated with an entity. |
| type: string | type: string | type: string | type: string | type: string | type: string | type: string | type: string |
| single value | single value | single value | single value | single value | single value | single value | single value |
| unrestricted | unrestricted | controlled vocabulary | unrestricted | controlled vocabulary | unrestricted | restriction: value from | restriction: value from Condition.alias |
| required | required | optional | required | recommended | recommended | optional | required |
| SAMPLE_1 | GHGAS_blood_sample1 | CF_DNA | Arterial blood sample 1 | Blood collection tube holder/needle | frozen at -20 | BIOSPECIMEN_1 | COND_1 |

- A *Sample* is defined as a limited quantity of something to be used for testing, analysis, inspection, investigation, demonstration, or trial use. A sample is prepared from a biospecimen (isolate or tissue).
- Sample therefore captures metadata about the storage, isolation and type of a sample as well as a general description
- It is linked to condition and biospecimen via the related aliases in the columns sample.condition and sample.biospecimen
- The fields isolation and storage use controlled vocabularies and the SNOMED ontology, which can be found here.

GHGA

# Sample.File

| alias | sample | name | format | size | checksum | forward_or_reverse | checksum_type | dataset |
|---|---|---|---|---|---|---|---|---|
| The alias for an entity at the time of submission. | The sample associated with an entity. | The given filename. | The format of the file: BAM, SAM, CRAM, BAI, etc. | The size of a file in bytes. | A computed value which depends on the contents of a block of data and which is transmitted or stored along with the data in order to detect corruption of the data. The receiving system recomputes the checksum based upon the received data and compares this value with the one sent with the data. If the two values are the same, the receiver has some confidence that the data was received correctly. | Denotes whether a submitted FASTQ file contains forward (R1) or reverse (R2) reads for paired-end sequencing. The number that identifies each read direction in a paired-end nucleotide sequencing reaction. | The type of algorithm used to generate the checksum of a file. | The Dataset associated with an entity. |
| type: string | type: string | type: string | type: string | type: integer | type: string | type: string | type: string | type: string |
| single value | single value | single value | single value | single value | single value | single value | single value | single value |
| unrestricted | restriction: value from Sample.alias | unrestricted | controlled vocabulary | unrestricted | unrestricted | controlled vocabulary | unrestricted | restriction: value from Dataset.alias |
| required | required | required | required | required | required | recommended | required | required |

- This field is functionally identical with *Study.File*. It can be used to link files to a dataset that have sample metadata.
- Files do not need to be linked to datasets multiple times. A file that is linked to a dataset via Sample.File does not need to be added again as Study.File.

GHGA

# Condition

| alias | title | description | name | disease_or_healthy | case_control_status | mutant_or_wildtype | study |
|---|---|---|---|---|---|---|---|
| The alias for an entity at the time of submission. | The title that describes an entity. | Description of an entity. | The name for an entity. | Whether a condition corresponds to a disease or a healthy state. | Whether a condition corresponds to a treatment or a control. | Whether a condition corresponds to a mutant or a wildtype. | The study associated with an entity. |
| type: string | type: string | type: string | type: string | type: string | type: string | type: string | type: string |
| single value | single value | single value | single value | single value | single value | single value | single value |
| unrestricted | unrestricted | unrestricted | unrestricted | controlled vocabulary | controlled vocabulary | controlled vocabulary | restriction: value from Study.alias |
| required | optional | required | required | required | required | required | required |
| COND_1 | Condition A | Condition A is a condition | Condition A | DISEASE | TRUE_CASE_STATUS | MUTANT | STUDY_A |

- A *Condition* describes the state and origin of a sample. It captures actions applied to a sample that were necessary for the specific study in which the sample is used. The *Condition* links the *Sample* to a *Study* via the field condition.study.

- The fields sample.disease_or_healthy and sample.case_control_status use controlled vocabularies that can be found here.