Conceptual Description of the Level 1 C2M2

This is a conceptual and narrative description of the Level 1 C2M2. It covers the things (proper nouns) in the Level 1 C2M2 and their relationships, along with the a description of the tables used to represent them. The last section covers the internal controlled vocabularies used for a few attributes. These notes don't go heavily into things like the format (syntax) of the columns or the specific primary key and foreign key relationships.

Things (Proper Nouns) Described

The Level 1 C2M2 includes tables to describe the following things (entities), and the relationships between them.

- Namespaces
- Project
- File
- Subject
- Biosample
- Collection

This section has descriptions of each thing and a list of its attributes (fields).

Namespaces

These are logical groupings of things so that the names DCCs use don't collide. Unless we think a DCC will have names for things like biosamples or files that could collide, then a single namespace for DCC is all that's needed. The unique names for things in the C2M2 are a combination of namespace and a local id, like (namespace, id) for collections.

Attributes

- **namespace** A globally unique ID representing this namespace
- **abbreviation** A very short display label for this namespace
- **name** A short, human-readable, machine-read-friendly label for this namespace
- **description** A human-readable description of this namespace

Project

There can be a single project for each DCC, or things like studies can be represented as subprojects. The field persistent id could be a website for project, or a DOI for a paper. When we get to collections to describe datasets or cohorts, we'll show what project they were part of.

- **abbreviation** A very short display label for this project
- **name** A short, human-readable, machine-read-friendly label for this project
- **description** A human-readable description of this project
- **persistent id** A persistent, resolvable (not nec. retrievable) URI generated by a DCC and attached to this project

File

- **file id** The unique name for this file, compromised of
    - **namespace** namespace for the DCC or file creator
    - **id** An ID representing this file, unique within this namespace

- **project** Which project or subproject created this file
- **persistent id** A persistent, resolvable (not nec. retrievable) URI generated by a DCC (using, e.g., our minid server) and attached to this file
- **creation_time** An ISO 8601 -- RFC 3339 (subset)-compliant timestamp documenting this file's creation time: YYYY-MM-DDTHH:MM:SS±NN:NN
- **size_in_bytes** The size of a file in bytes
- **sha256** The output of the SHA-256 cryptographic hash function after being run on this file: one or both of sha256 and md5 is required; sha256 is preferred
- **md5** The output of the MD5 message-digest algorithm after being run as a cryptographic hash function on this file: one or both of
- **filename** A filename with no prepended PATH information
- **file_format** An EDAM CV term ID identifying the digital format of this file (e.g. TSV or FASTQ)
- **data_type** An EDAM CV term ID identifying the type of information stored in this file (e.g. RNA sequence reads)

Subject

- **subject id** The unique name for this subject, compromised of
    - **namespace** namespace for the DCC or subject provider
    - **id** An ID representing this subject, unique within this namespace
- **project** Which project or subproject created this file
- **persistent id** A persistent, resolvable (not nec. retrievable) URI generated by a DCC and attached to this subject
- **creation_time** An ISO 8601 -- RFC 3339 (subset)-compliant timestamp documenting this subject record's creation time: YYYY-MM-DDTHH:MM:SS±NN:NN
- **granularity** A CFDE CV term categorizing this subject by multiplicity (see Subject Granularity under Controlled Vocabularies). One of
    - single organism
    - symbiont system
    - host-pathogen system
    - microbiome
    - cell line
    - synthetic

Biosample

- **biosample id** The unique name for this biosample, compromised of
    - **namespace** namespace for the DCC or biosample owner
    - **id** An ID representing this biosample, unique within this namespace
- **project** Which project or subproject created this biosample
- **persistent id** A persistent, resolvable (not nec. retrievable) URI generated by a DCC and attached to this biosample
- **creation_time** An ISO 8601 -- RFC 3339 (subset)-compliant timestamp documenting this biosample's creation time: YYYY-MM-DDTHH:MM:SS±NN:NN
- **assay_type** An OBI CV term ID describing the type of material represented by this biosample
- **anatomy** An UBERON CV term ID used to locate the origin of this biosample within the physiology of its source or host organism

Collection

Like projects, collections can have subcollections. Collections can hold files, biosamples, or subjects, which is done using a relationship.

- **collection id** The unique name for this collection compromised of
  - **namespace** namespace for the DCC or collection creator
  - **id** An ID representing this collection, unique within this namespace
- **persistent id** A persistent, resolvable (not nec. retrievable) URI generated by a DCC and attached to this collection
- **abbreviation** A very short display label for this collection
- **name** A short, human-readable, machine-read-friendly label for this collection
- **description** A human-readable description of this collection

Relationships

There are several relationships between things that can be described, like which subject a biosample comes from. These a often mapping tables between the unique names (namespace, id) of different things.

Things in Collections

Collections can contain one or more files, biosamples, or subjects. A collection may contain a combination of different types. There are tables for each type that map the items into their collections. The item is identified by its namespace and id, so is the collection. Effectively, the tables look like the following:

Attributes

Files in collection

- **subject id** The unique name (namespace, id) of the subject
- **collection id** The unique name (namespace, id) of the collection

Biosamples in collection

- **biosample id** The unique name (namespace, id) of the biosample
- **collection id** The unique name (namespace, id) of the collection

Subjects in collection

- **subject id** The unique name (namespace, id) of the subject
- **collection id** The unique name (namespace, id) of the collection

Biosamples and Subjects

To allow for multiple subjects to be represented in a single biosample and vice versa, there is a mapping table between biosamples and subjects.

Attributes

- **biosample id** The unique name (namespace, id) of the biosample
- **subject id** The unique name (namespace, id) of the subject

Files Describing Subjects and Biosamples

To show a relationship between a file and a subject a or biosample, like a sequence file generated from a biosample, there are two more mapping tables.

Attributes

Files describing biosamples

- **file id** The unique name (namespace, id) of the file
- **biosample id** The unique name (namespace, id) of the biosample

Files describing subjects

- **file id** The unique name (namespace, id) of the file
- **subject id** The unique name (namespace, id) of the subject

Subject Role and Taxonomy

A table linking a subject, a subject_role (a named organism-level constituent component of a subject, like 'host', 'pathogen', 'endosymbiont', 'taxon detected inside a microbiome subject', etc.) and a taxonomic label (which is hereby assigned to this particular subject_role within this particular subject)".

Attributes

- **subject**
  - **namespace** The namespaec of the subject
  - **id** The ID of this subject
- **role** The role assigned to this organism-level constituent component of this subject (see Subject Role under Controlled Vocabularies). One of
  - single organism
  - host
  - symbiont
  - pathogen
  - microbiome taxon
  - cell line ancestor
  - synthetic
- **taxonomy_id** An NCBI Taxonomy Database ID identifying this taxon

CFDE Controlled Vocabularies

Subject Granularity

Term	Description
single organism	One organism
symbiont system	A mixed system of consisting of two or more organisms (symbionts) in symbiosis (living colocated in time and space): one such symbiont may optionally be identified as a host
host-pathogen	A special case of a symbiont system consisting of one symbiont, designated as a host, plus one or more other

system	symbionts acting to create or sustain disease within the host organism
microbiome	A symbiont system consisting of a collection of (potentially unknown or partially characterized) taxa, where the environment in which the system resides is well-characterized, but the taxonomic composition of the system may be unknown; optionally contains one symbiont specially identified as a host
cell line	A cell line derived from one or more species or strains
synthetic	A synthetic biological entity

Subject Role

Term	Description
single organism	The organism represented by a subject in the 'single organism' granularity category
host	Any organism identified as a host for a subject assigned to the 'symbiont system', 'host-pathogen system', or 'microbiome' granularity categories
microbiome taxon	A constituent taxon of either (a) a subject assigned to the 'environmental microbiome' granularity category or (b) the microbiome (non-host) portion of a subject assigned to the 'host-associated microbiome' granularity category [NB: This role is probably not appropriate for Level 1, because it necessitates the post-facto attachment of downstream analysis procedures (subject -> sample -> library prep -> sequencing -> bioinformatics -> taxonomic classification results) to a subject which was originally uncharacterized at this level]
symbiont	An organism identified as a symbiont within a subject assigned to the 'symbiont system' granularity category
pathogen	An organism identified as a pathogen symbiont in a subject assigned to the 'host-pathogen system' granularity category
cell line ancestor	A taxon identified as a source organism for a subject assigned to the 'cell line' granularity category
synthetic	A synthetic biological entity