Relational schema

The core of our metadata repository exhibits a star-like relational schema, illustrated in Relational_Schema_Figure, centered on the Item table; it physically implements the Genomic Conceptual Model, as each table corresponds to a GCM entity. The core schema is extended by two subschemas representing, respectively, the original unstructured metadata – in the form of key-value pairs – and the semantic enrichment for specific attributes of four core tables (Knowledge Base).

Core schema

The core schema is a classic data mart, with a central fact table describing Items (or data files) and four dimensions:

- The *Biological dimension* describes the biological material and process observed in the experiment that generated the genomic item. It includes Donor, Biosample and Replicate entity tables, and the Replicate2Item bridge table.
- The *Management dimension* describes the organizations or projects that are behind the production of each experiment. It includes the Project and CaseStudy entity tables, and the Case2Item bridge table.
- The *Technology dimension* describes the process used for the production of the experimental or annotation item and includes the ExperimentType entity table.
- The *Extraction dimension* describes the containers available in the repository for storing items that are homogeneous for data analysis; it includes the Dataset entity table.

All core tables have a numerical sequential primary key, conventionally named <table_name>_id and indicated as PK in Relational_Schema_Figure. Tables Donor, Biosample, Replicate, Item, and CaseStudy have, in addition, a secondary unique key <table_name>_source_id that refers to the original source; such secondary key is used for providing backward links to the data source (and for direct comparison of source contents with the ones in the repository during periodic updates/reloads).

Core tables have two kinds of foreign keys (FKs): the FKs that uniquely identify a row of another table of the core schema (in red in Relational_Schema_Figure) and FKs that reference concepts in the Knowledge Base from the core attributes that are semantically enriched (in blue in Relational_Schema_Figure). Nullable attributes are indicated in Relational_Schema_Figure with N.

Relationships in the core schema from the Item outward are functional (i.e., one Item has one ExperimentType, while an ExperimentType may be the same for multiple Items), with the exception of two many-to-many relationships: each Item derives from one or more Replicates and belongs to one or more CaseStudies.

We next discuss every table of the core schema.

Item. Each item corresponds to a processed data file that contains genomic region data. It references the ExperimentType and Dataset tables with FKs, whereas item_id is directly used in bridge tables Replicate2Item and Case2Item. Size, date, and checksum denote properties of the corresponding genomic data file; source_url, local_url, file_name, and source_page include information useful to locate and download the physical data file and associated information. The content_type describes the type of genomic regions in the file (such as gene segments, introns, transcripts, etc.) and is enriched using concepts in the NCIT (28) and SO (29) ontologies. The platform is the instrument used to sequence the raw data related to the item and is enriched using the OBI ontology (30). The pipeline includes a list of methods used for processing phases, from raw data to processed data.

Replicate. When an assay is performed multiple times on separate biological samples (or even on the same sample), multiple replicas of the same experiment are generated, each associated with a distinct item and progressive numbers (indicated as biological_replicate_number and technical_replicate_number). Multiple replicates for the same item are present in the sources ENCODE and Roadmap Epigenomics.

Replicate2Item. This bridge table, by combining item_id and replicate_id, can associate multiple Items to a single Replicate (i.e., they may have undergone different processing) and multiple Replicates to a single Item (such items are generally called "combined").

Biosample. It describes the material sample taken from a biological entity and used for the experiment. It references the Donor table with an FK. The biosample_type distinguishes between tissues, cell lines, primary cells, etc. The tissue field is enriched by concepts in the Uberon ontology (31), describing a multicellular component in its natural state, or the provenance tissue of cells. The cell field allows to specify single cells (in natural state), immortalized cell lines, or cells differentiated from specific cell types; it is enriched by concepts in the EFO (32) and CL (33) ontologies. The disease (i.e., illness investigated within the sample) is enriched by the NCIT ontology; the is_healthy field stores a Boolean condition, as the biological sample may be healthy/control/normal or non-healthy/tumoral.

Donor. It describes the donor providing the biological sample. The donor age, gender, ethnicity (enriched with terms from the NCIT ontology), and species (enriched with terms from the NCBITaxon terminology (34)) refer to the individual from which the biological sample was derived (or the cell line established).

CaseStudy. It connects the set of items that are collected together, as they participate to the same research objective (the criteria used by each source to group together such files are variable). It references the Project table with an FK. The source_site represents the physical site where the material is analyzed and experiments are physically produced (e.g., universities, biobanks, hospitals, research centers, or just laboratory contact references when a broader characterization is not available). External_reference may contain identifiers taken from the main original source and other sources that contain the same data.

Case2Item. This bridge table, by combining item_id and case_study_id, can associate multiple Items to a single case (which is the typical scenario), but also multiple cases to a single Item (this happens when an Item appears in multiple analyses and studies).

Project. It represents the infrastructure or organization that sets the context for the experiments (or case studies). Source describes the programs or consortia responsible for the production of genomic items (currently featuring five possibilities: TCGA, ENCODE, Roadmap Epigenomics, RefSeq, GENCODE). Within a source, items may be produced within a specific initiative, specified in the project_name, which uniquely references the project; it is particularly relevant in the context of TCGA data, where items are organized based on the type of tumor analyzed in the specific project (e.g., BRCA identifies a set of items regarding the Breast Invasive Carcinoma study), or in annotation projects (such as the RefSeq reference genome annotation).

ExperimentType. It refers to the specific methods used for producing each experimental or annotation data file (hence, each item of the core schema). W.r.t. the original source, a tuple is uniquely identified by the triple technique, feature and target. The first one is enriched by the OBI or EFO ontologies and describes the assay, i.e., the investigative procedure conducted to produce the items. The second one is enriched by the NCIT ontology and describes the specific genomic aspect studied with the experiment (e.g., gene expression, mutation, histone mark). Epigenomic experiments such as ChIP-seq usually analyze a protein, which we call target; this field is enriched by concepts in the OGG ontology (35). The antibody is the protein employed against such target (values refer to The Antibody Registry, http://www.antibodyregistry.org/, or the ENCODE antibody accession, in case the first is missing).

Dataset. It gathers groups of items stored within a folder named dataset_name; dataset items are homogeneous as they share a specific data_type (e.g., peaks, expression quantifications, methylation levels), assembly (i.e., reference genome alignment – either hg19 or GRCh38), and file_format (i.e., standard data format of the items dictating the genomic region data schema, including the number and semantics of attributes, for example BED, narrowpeak or broadpeak). The Boolean variable is_annotation allows distinguishing between datasets containing experimental data and datasets storing genomic annotations (currently defined in the Item's content_type field).

Original metadata

Out of about 40 million metadata extracted from sources, around 7 million were included in the core schema. Many attributes and their respective values found within different sources cannot be mapped to the same conceptual model. We store such extra attributes in an unstructured format, using *key-value* pairs extended with the item_id of the Item which they refer to; all attributes together form the primary key, while the item_id also acts as foreign key.

Knowledge Base

Some of the attributes of the core schema have been annotated with ontological concepts using an automatic procedure following described. Enriched attributes include: the ethnicity and species describing Donors; disease, tissue and cell describing Biosamples; technique, feature and target describing ExperimentTypes; platform and content_type describing Items.

Automatic enrichment is performed by using one or two preferred bio-ontologies for each attribute (details on the annotation process are available in (36)). For a given value, when a match with an ontology term is not found, the annotation task is re-routed to a manual procedure handled by an admin user who is expert in data curation and biomedical ontologies. So far, we enriched attribute values by linking them to 1,629 terms in the 8 specified ontologies. In addition to terms that directly annotate core values (and their synonyms), we included all terms that could be reached by traversing up to three ontology levels from the base term (12,087 concepts in total); as next discussed, the use of three levels enables powerful query extensions.

The Knowledge Base is deployed using as well relational tables; in particular, we use:

1) the ***Vocabulary*** table, whose PK term identifier tid is referenced from all the core tables that contain semantically enriched attributes, with the acronym of the ontology providing the term (source, e.g., NCIT), the code used for the term in that ontology (e.g., NCIT_C4872) and its label (pref_label, e.g., Breast Carcinoma), in addition to an optional description and iri (i.e., International Resource Identifier);
2) the ***Ontology*** table, a dimension table presenting details on the specialized ontologies contained (even partially) in the knowledge base – referenced with an FK from the vocabulary table;

3) the ***Reference*** table, containing references to equivalent terms from other ontologies (in the form of a <source, code> pair) – referencing, with the FK tid, the term in the vocabulary table;
4) the ***Synonym*** table, containing alternative labels that can be used as synonyms of the preferred label along with their type (e.g., alternative syntax, related nomenclature, related adjectives) – referencing the term in the vocabulary table;
5) the ***Relationship*** table, containing ontological hierarchies between terms and the type of the relationships (either generalization *is_a* or containment *part_of*) – the primary key is composed of parent, child and type of the relationship; the first two reference the vocabulary table with FKs.

For performance issues we materialized an unfolded representation of the Relationship table and a denormalized representation of the core tables, which are used by search queries; they are rematerialized at each change of the database.