GDC DATA MODEL

GDC Data Model Components

Graph Representation of the GDC Data Model

The GDC data model is represented as a graph with nodes and edges, and this graph is the store of record for the GDC. It maintains the critical relationship between projects, cases, clinical data and molecular data and insures that this data is linked correctly to the actual data file objects themselves, by means of unique identifiers. The graph is designed in terms of the "property graph" model, in which nodes represent entities, edges between nodes represent relationships between entities, and properties on both nodes and edges represent additional data which describe entities and their relationships. Relationships are encoded as edges of a given type which associate exactly two nodes. Properties of nodes or relationships are sets of key-value pairs.

Original metadata as submitted by external users is extracted and loaded first into the graph. Representations of the data provided by the other GDC components are derived from the authoritative graph model. Note that file and archive objects are not stored in the graph, but rather in an external object store. The node/edge structure of the graph is depicted below.

GDC-DataModel-Aug2017

Image not found or type unknown
[1]

Data Dictionary and Validation

Item semantics (names, accepted values) and their interrelationships can be changed or updated easily within the GDC data model. However, they cannot be completely free to vary at any time, since users depend on the stability of the graph and its vocabulary. Therefore, a means to encode a schema for the metadata is also an integral part of the GDC data model. The GDC maintains standard terms, their definitions and references to public ontologies and vocabularies in dictionary files. Dictionaries are expressed in YAML, utilizing JSON Schema [2] conventions to be computable and are intended to provide both the internal GDC standard vocabulary and the publicly-accessible source of GDC term information. Each node is represented by a dictionary file that specifies node properties and values, as well as allowable edges to other node types. Dictionaries are kept in version control with periodic releases.

As submitted metadata is processed, the associations among cases, biospecimens, and data files are extracted and stored in the graph. To prevent errors in those associations from entering the graph, a validation system is implemented at the application level. Primary validation rules

(such as data type and accepted value checking) are defined in JSON Schema as part of the GDC data dictionaries. Secondary validations (those that, for example, confirm data consistency among values of different entries in a submission) are implemented as custom modules and referenced in relevant dictionaries. Validation is executed on incoming metadata and must pass before they are allowed to enter the graph itself.

Data Type and Subtype Definition and Tagging

To categorize files from legacy programs on initial import, GDC developed and implemented a system of associating data type, data subtype, platform, experimental strategy, and access type assignments to files based on file name pattern matching. To facilitate user search and download of desired data, a system of short tokens or tags has also been developed to classify externally available files into related groups outside the set of defined data typing dimensions listed above. The data typing facets and tags provide the basis for categorizing files submitted or generated in the future. See the TCGA Tags Guide [3] for more details.

GDC Data Model Management

The GDC has designed the data model to be able to accommodate changes in structure and content with minimal required changes to underlying datastore configuration or to existing content. In particular, using a graph oriented design allows additions of items, relationships, and properties that leave preexisting queries in the codebase and external workflows more or less unaffected. At the same time however, a data management system that is easily modified can also more easily grow unsystematically, creating a risk that the existence of certain items in the database, and means for finding them, can be unknown to all stakeholders except for those who made the modifications. Modifications to a graph may also unintentionally remove or change links to items, creating "orphans" that no longer can be retrieved through existing queries. The risk of creating unmanageable datastores can be mitigated by establishing processes of technical communication and change control. There are three main groups for this purpose at GDC.

Data Model Working Group

The Data Model Working Group is a small group of GDC and Leidos bioinformaticists and engineers with PO representation. It meets weekly to work through design and import questions and issues. In the design phases, this group has actively made design and implementation decisions. Those decisions are recorded in GDC Data Model Working Group meeting notes.

Data Model Change Control Board (CCB)

The Data Model Change Control Board (CCB) is a group that meets at need to formally review requests to change key data model configuration items, and establishes consensus among GDC stakeholders to approve or deny these requests. CCB is responsible for changes to the following configuration items:

- The addition, modification, or deletion of data model node and edge classes, and their allowable relationships and semantic content;
- The graph RDBMS schema;
- The property graph schema (instantiated in *gdcdatamodel*); as well as
- GDC dictionary JSON documents.

Technical implementation details of design changes approved by the CCB are generally out of the CCB scope of discussion. Implementation of data model changes is handled by the GDC SDLC process. Change requests may be submitted by any stakeholder.

Data Model Advisory Group

The Data Model Advisory Group includes external NCI and other CCG-invited experts in cancer clinical and genomic data organization and management, high level CCG representatives, and GDC and Leidos management, bioinformaticists and engineers. This group meets at need to provide an external perspective on the ongoing development of the data model, and advises on strategic issues of design that will enable the GDC data model to

- Expand to support new NCI genomics projects;
- Develop and increase interoperability among peer level databases and key GDC collaborators; and
- Maintain semantic consistency to maximize data usability in clinical and epidemiological studies.

Source URL: https://gdc.cancer.gov/content/gdc-data-model-0

Links:
[1] https://gdc.cancer.gov/files/public/image/FullDataModel_July2017.png
[2] http://json-schema.org/
[3] https://gdc.cancer.gov/resources-tcga-users/legacy-archive-tcga-tag-descriptions