GLIOMA01 Transformation Documentation

GLIOMA01 has two transformations, one for case centric data and one for sample data. The input files for these transformations can be found in Box at <https://nih.app.box.com/folder/117853007838>.

The transformations map the columns from the input files to node.properties in the database. Mappings can be one-to-one (i.e., one column in the input file maps to a single node.property), or many-to-one (i.e., two or more columns in the input file are concatenated to map to a single node.property). In addition some columns map the values in the input file to a set of accepted values in the database for the given node.property.

# Case Transformation

## Case Node Mappings

|  |  |  |  |
| --- | --- | --- | --- |
| **PROPERTY** | **MAPPING** | **VALUE MAPPING FROM** | **VALUE MAPPING TO** |
| case\_id | icdc\_study + "-" + icdc\_patient\_id | N/A | N/A |
| patient\_id | icdc\_subject\_id | N/A | N/A |

## Demographic Node Mappings

|  |  |  |  |
| --- | --- | --- | --- |
| **PROPERTY** | **MAPPING** | **VALUE MAPPING FROM** | **VALUE MAPPING TO** |
| breed | icdc\_breed | N/A | N/A |
| patient\_age\_at\_enrollment | icdc\_age\_at\_study\_enrollment\_calendar\_years | N/A | N/A |
| sex | icdc\_sex | N/A | N/A |
| neutered\_indicator | icdc\_spayed/neutered\_status | N/A | N/A |

## Diagnosis Node Mappings

|  |  |  |  |
| --- | --- | --- | --- |
| **PROPERTY** | **MAPPING** | **VALUE MAPPING FROM** | **VALUE MAPPING TO** |
| disease\_term | icdc\_diagnosis | N/A | N/A |
| primary\_disease\_site | icdc\_primary\_disease\_site | N/A | N/A |
| stage\_of\_disease | icdc\_stage\_of\_disease | N/A | N/A |
| date\_of\_diagnosis | icdc\_date\_of\_diagnosis | N/A | N/A |
| best\_response | best\_response | N/A | N/A |
| pathology\_report | icdc\_pathology\_reports | N/A | N/A |
| treatment\_data | icdc\_treatment\_data | N/A | N/A |
| follow\_up\_data | icdc\_follow\_up\_data | N/A | N/A |

## Enrollment Node Mappings

|  |  |  |  |
| --- | --- | --- | --- |
| **PROPERTY** | **MAPPING** | **VALUE MAPPING FROM** | **VALUE MAPPING TO** |
| date\_of\_registration | icdc\_date\_of\_study\_enrollment | N/A | N/A |
| site\_short\_name | icdc\_tss | N/A | N/A |

## Additional Mappings

In addition, we include some extra columns to the transformed files. These columns are used by the data loader.

|  |  |
| --- | --- |
| **NODE.PROPERTY** | **MAPPING** |
| study.clinical\_study\_designation | icdc\_study |
| enrollment\_id | icdc\_study + "-" + icdc\_patient\_id |
| diagnosis\_id | icdc\_study + "-" + icdc\_patient\_id |

## Input Columns Not Used

Submitters sometimes include columns with data that are not going to be included in our database. We leave these columns in the input file, but the transformation ignores them. For the case input file, these columns include:

* icdc\_tumor\_morphology
* icdc\_tumor\_grade
* TSS\_Subject\_ID
* TSS\_Sample\_ID

# Sample Transformation

## Sample Node Mappings

|  |  |  |  |
| --- | --- | --- | --- |
| **PROPERTY** | **MAPPING** | **VALUE MAPPING FROM** | **VALUE MAPPING TO** |
| sample\_id | icdc\_study + “-“ + icdc\_sample\_id |  |  |
| sample\_site | icdc\_sample\_site |  |  |
| physical\_sample\_type | icdc\_sample\_type |  |  |
| general\_sample\_pathology | icdc\_general\_sample\_pathology |  |  |
| tumor\_sample\_origin | icdc\_tumor\_sample\_origin |  |  |
| summarized\_sample\_type | icdc\_sample\_type + “;” + icdc\_general\_sample\_pathology + “;” + icdc\_tumor\_sample\_origin |  |  |
|  |  | Tissue;Normal;Not Applicable | Normal Tissue |
|  |  | Tissue;Malignant;Primary | Primary Malignant Tumor Tissue |
|  |  | Blood;Not Applicable;Not Applicable | Whole Blood |
|  |  | NA;NA;NA | N/A |
| specific\_sample\_pathology | icdc\_tumor\_morphology |  |  |
| date\_of\_sample\_collection | icdc\_date\_of\_sample\_acquisition |  |  |
| sample\_chronology | sample\_chronology |  |  |
| necropsy\_sample | icdc\_necropsy\_sample |  |  |
| tumor\_grade | icdc\_tumor\_grade |  |  |
| sample\_preservation | icdc\_tissue\_archival |  |  |
| comment | sample\_comments |  |  |

## Data File Manifest Mappings

For GLIOMA01 ONLY, we used values from the submitter’s sample file to help Ming create a manifest that is used to load the data files. Note that these values were used by Ming only, and were NOT saved to our database as they appear in this file. Here are the mappings we used from the submitter’s sample input file.

|  |  |  |  |
| --- | --- | --- | --- |
| **OUTPUT COLUMN NAME** | **MAPPING** | **VALUE MAPPING FROM** | **VALUE MAPPING TO** |
| SRA\_filename | file\_name |  |  |
| seq\_platform | file\_type |  |  |
|  |  | exome | Whole Exome Sequence File |
|  |  | rnaseq | RNA Sequence File |
|  |  | rrbs | DNA Methylation Analysis File |
|  |  | wgs | Whole Genome Sequence File |
| SRA\_Title | file\_description |  |  |
| md5 | original\_md5 |  |  |
| Accession | SRA\_accession |  |  |

## Additional Mappings

In addition, we include some extra columns to the transformed files. These columns are used by the data loader.

|  |  |
| --- | --- |
| **NODE.PROPERTY** | **MAPPING** |
| case.case\_id | icdc\_study + "-" + icdc\_patient\_id |

## Input Columns Not Used

Submitters sometimes include columns with data that are not going to be included in our database. We leave these columns in the input file, but the transformation ignores them. For the sample input file, these columns include:

* icdc\_tissue\_type
* BioProject
* BioSample