Data Dictionary

OVERVIEW

This database was created to store hundreds of thousands of human and machine annotations.
In this context "human" annotations refer to annotations created by an expert reviewer using a program such as protege, whereas "machine" annotations are automatically labeled. Medical dictionaries and journal publications are parsed and stored in this database.

LIST OF TABLES

Table Name	Used for Scrubbing	Medical Concepts (UMLS)	Used for Publication Analysis
Feature_matrix_test	YES	YES	
Feature_matrix_train	YES	YES	
Human_annotations_test	YES		
Human_annotations_train	YES		
Machine_annotations_test	YES		
Machine_annotations_train	YES		
Lookup_dictionary	YES	YES	
Lookup_term_frequency	YES		
Lookup_umls	YES		
Pubs_authors			YES
Pubs_keywords			YES
Pubs			YES
Pubs_refs			YES

TABLE DESCRIPTIONS

Feature_matrix_*

Stores feature matrix that is built from the Machine_annotations_* and Human_annotations_* tables. This is the rolled up feature set used for classification.

Human_annotations_*

Stores all annotations created by humans as part of a manual annotations effort.

Machine_annotations_*

Stores all annotations created by the UIMA pipeline.

Lookup_dictionary

Contains names from the 1990 US census that are used in

Lookup_term_frequency

Contains term frequency calculated across a random selection of 10,000 open access medical publications. Raw open access publications are available for free through NIH/NLM.

Lookup_umls

Contains terms from UMLS subset that was used Scrubber.

This DOES NOT include the UMLS CUIDs due to licensing restrictions.

Medical Vocabularies

Vocabularies	#Concepts

COSTAR	3,461
HL7V2.5	5,020
HL7V3.0	8,062
ICD10CM	102,048
ICD10PCS	253,708
ICD9CM	40,491
LOINC	327,181
MESH	739,161
RXNORM	437,307
SNOMEDCT	1,170,855