



Serializing the UMLS into RDFS

Janos G. Hajagos

Stony Brook School of Medicine

9/29/2010

RDFS

`<http://www.w3.org/2000/01/rdf-schema#>`

- A light weight schema for Resource Description Framework (RDF)
- Class and properties
 - `rdfs:Class` & `rdf:Property`
- Basic inference
 - `rdfs:subClassOf` & `rdfs:subPropertyOf`
 - `rdfs:domain` & `rdfs:range`
- Human readable labels
 - `rdfs:label` & `rdfs:comment`

UMLS

Unified Medical Language System

- A long term (1986) research project of NIH's National Library of Medicine (NLM)
- A metathesaurus connecting different medical/biomedical vocabularies together with a concept unique identifiers (CUI)
- A semantic network of 54 broad types
 - “Neoplastic process” isa “Disease or Syndrome”
- `rdfs:seeAlso`
 - [“Ontologies for Data Integration: A Semantic Web Perspective”](#) for more in depth material on the UMLS
 - [Initial publication of UMLS in 2007 in a semantic format.](#)

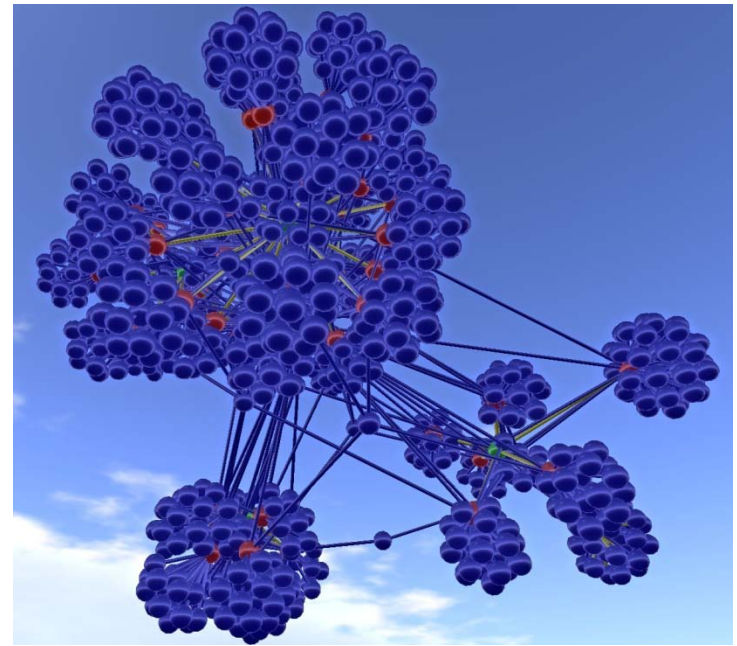
UMLS 2010AA Sources in Conversion

Sorted by decreasing order of the total number of terms in source vocabulary

- SNOMED Clinical Terms
- Metathesaurus Names
- MeSH
- RXNORM
- Read Codes
- LOINC
- MEDCIN
- NCBI Taxonomy
- MedDRA
- NCI Thesaurus
- Gene Ontology
- ICD-10-PCS
- Digital Anatomist
- Foundational Model of Anatomy
- National Drug File - Reference Terminology
- FDA National Drug Code Directory
- HUGO
- SNOMED Intl 1998
- Multum
- OMIM
- UMDNS
- National Drug File
- ICPC2-ICD10 Thesaurus
- Medical Entities Dictionary
- FDA Structured Product Labels
- Alternative Billing Concepts
- ICD-9-CM
- WHOART
- CRISP Thesaurus
- Alcohol and Other Drug Thesaurus
- CPT in HCPCS
- National Drug Data File
- ICNP
- Nursing Outcomes Classification
- SNOMED 1982
- Clinical Problem Statements
- MedlinePlus
- PDQ
- Clinical Classifications Software
- HCPCS
- Psychological Index Terms
- Gold Standard Alchemy
- ICPC-2 PLUS
- CPT
- HL7 Version 3.0
- ICD-9-CM Entry Terms (UMLS)
- CPT Hierarchical Terms (UMLS)
- Micromedex
- ICD-10
- Nursing Interventions Classification
- HL7 Version 2.5
- International Classification of Functioning, Disability and Health
- COSTART
- Standard Product Nomenclature
- Read Codes Synth
- Neuronames Brain Hierarchy
- Minimum Data Set, 2.0
- USP Model Guidelines
- Read Codes Am Engl
- International Classification of Functioning, Disability and Health for Children and Youth
- HCPCS Hierarchical Terms (UMLS)
- Minimal Standard Terminology (UMLS)
- Master Drug Data Base
- Minimum Data Set, 3.0
- DXplain
- Clinical Concepts by R A Miller
- Congenital Mental Retardation Syndromes
- Current Dental Terminology in HCPCS
- Beth Israel Problem List
- Outcome and Assessment Information Set
- Patient Care Data Set
- Library of Congress Subject Headings
- Omaha System
- ICPC
- Home Health Care Classification
- ICPC2E
- Authorized Osteopathic Thesaurus
- COSTAR
- Pharmacy Practice Activity Classification
- AI/RHEUM
- DSM-IV
- Patient Health Questionnaire
- PNDS
- Patient Monitoring Guidelines for HIV care and antiretroviral therapy (ART)
- Read Codes Am Synth
- DSM-III-R
- ICD-10 Am Engl
- Source Terminology Names (UMLS)
- Classification of Nursing Diagnoses
- Quick Medical Reference
- ICPC2-ICD10 Thesaurus, 7-bit
- ICPC2-ICD10 Thesaurus, Am Engl
- HL7 Version 2.5, 7-bit equivalents
- Diseases Database
- Confusion Assessment Method (CAM)
- UltraSTAR
- Braden Scale for Predicting Pressure Sore Risk
- Faces, Legs, Activity, Cry, and Consolability (FLACC) Scale
- Routine Health Outcomes Ltd. (RHO)
- ICPC2E Am Engl
- Glossary of Clinical Epidemiologic Terms

Relation of RxNorm to UMLS

- RxNorm's data table structure is very close to the UMLS
- RxNorm sources are contained within the UMLS
- RxNorm is updated more frequently so as to better capture new drugs entering the market



[RxNorm RDF data of drugs containing Lithium visualized in 3D by Nexus](#)

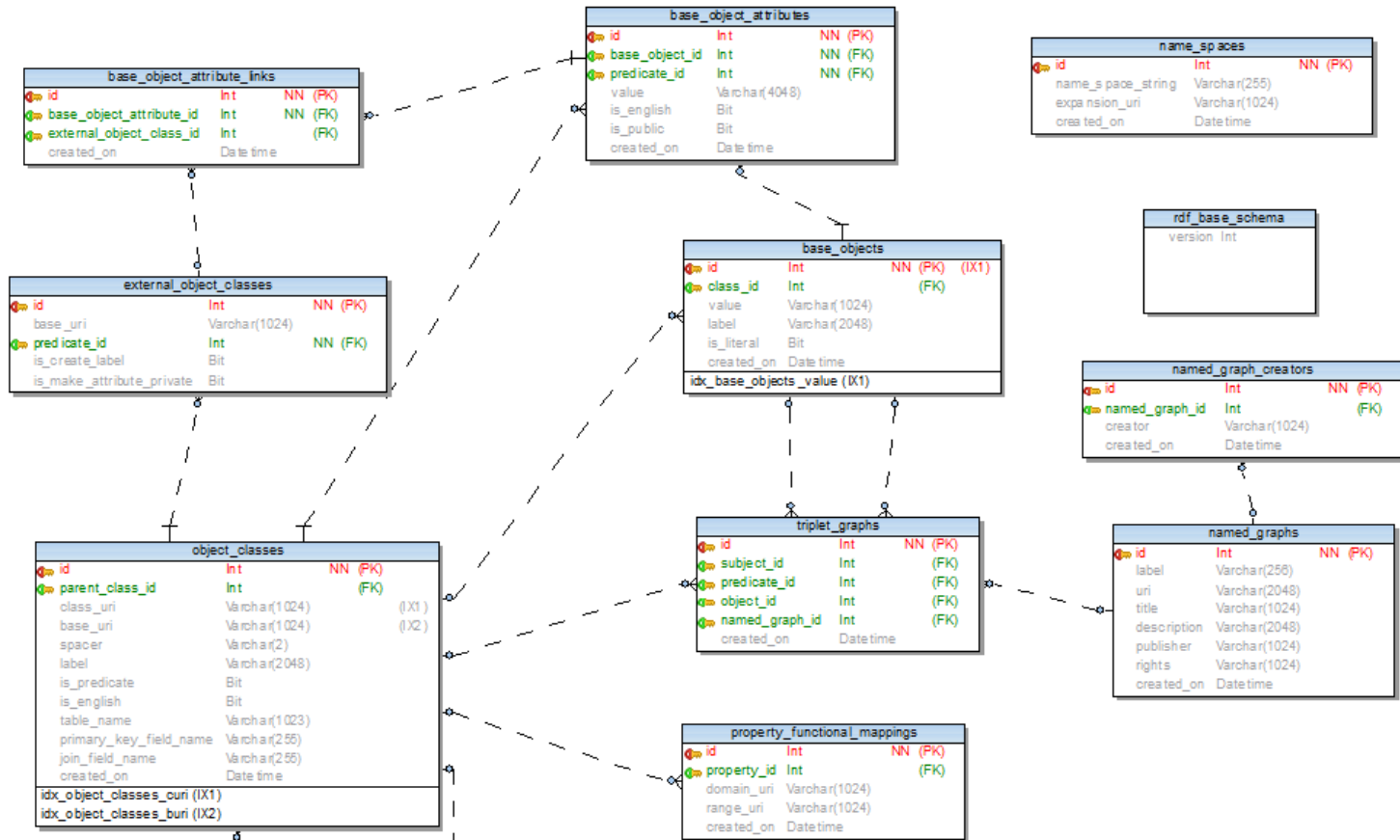
Comparison of the UMLS to RxNorm

	UMLS	RxNorm
Developed by	NLM	NLM
Focus	Genes, Drugs, Disease, Anatomy	Drugs
Identifiers	CUI AUI	RXCUI RXAUI
Update frequency	2-4 times a year	Full monthly with partial weekly updates
Release format	.NLM with Java program to generate site RRF	RRF (Rich Release Format)
Database load scripts	MySQL and Oracle	MySQL and Oracle
License	UMLS Metathesaurus Site License Each individual source is licensed separately	UMLS Metathesaurus Site License Each individual source is licensed separately

RdfBase

- Targets “one time” conversions of relational databases into RDFS
- RDFS is serialized into ntriples format
- Used to publish RxNorm, UMLS, and a data warehouse schema with reference table values
- Python based library using Microsoft SQL Server on the backend


RdfBase Relational DB Schema

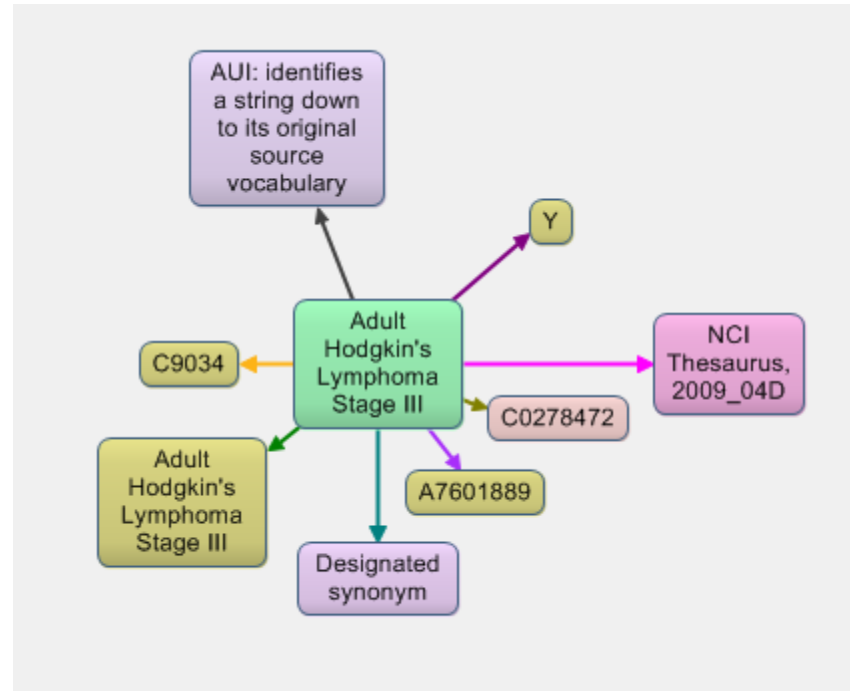


RdfBase Code Snippet

```
86 class m004PublishMRCONSO(MigrationClass):
87     def up(self):
88
89         self.cursor.execute("""
90             select CUI, AUI, TTY, CUI, CODE, ISPREF, cast(STR as varchar(4000)) as [STR]
91                 into MRCONSO_cleansed
92             from MRCONSO where SUPPRESS = 'N'
93         """)
94
95         umls_obj = RdfBase(self.cursor)
96         aui_obj = RdfBaseTable("MRCONSO_cleansed", umls_obj)
97         aui_obj.set_class_name("umls:AUI", "AUI: identifies a string down to its original source vocabulary", spacer="/")
98         aui_obj.set_label_field_name("STR")
99         aui_obj.set_value_field_name("AUI")
100        aui_obj.set_primary_key_field_name("AUI")
101        aui_obj.map_field_name("SAB")
102        aui_obj.map_field_name("CODE")
103        aui_obj.map_field_name("AUI")
104        aui_obj.map_field_name("CUI")
105        aui_obj.map_field_name("TTY")
106        aui_obj.map_field_name("CUI")
107        aui_obj.map_field_name("ISPREF")
108        aui_obj.publish()
109
110    def down(self):
111        self.cursor.execute("drop table MRCONSO_cleansed")
112
```


MRCONSO (objects)

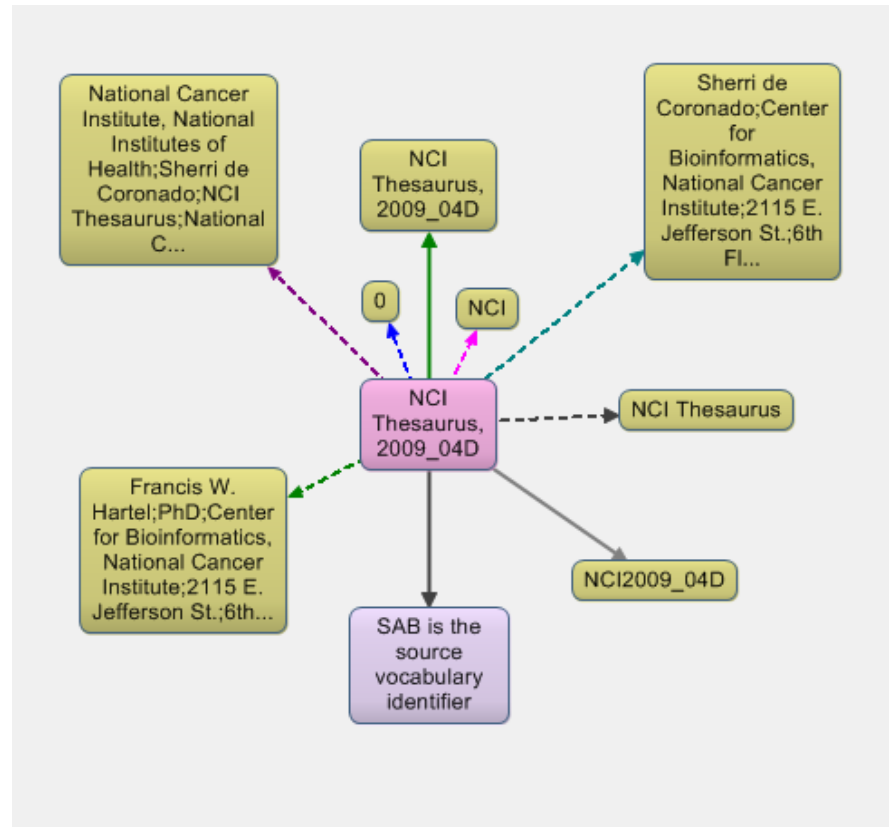
MRCONSO		
CUI	Varchar(8)	
LAT	Varchar(3)	
TS	Varchar(1)	
LUI	Varchar(10)	
STT	Varchar(3)	
SUI	Varchar(10)	
ISPREF	Varchar(1)	
 AUI	Varchar(9)	NN (PK)
SAUI	Varchar(50)	
SCUI	Varchar(50)	
SDUI	Varchar(50)	
SAB	Varchar(20)	
TTY	Varchar(20)	
CODE	Varchar(50)	
STR	Text	
SRL	Numeric(10,0)	
SUPPRESS	Varchar(1)	
CVF	Numeric(10,0)	
created_on	Datetime	



<<http://link.informatics.stonybrook.edu/umls/AUI/A7601889>>


MRSAB (metadata)

MRSAB		
VCUI	Varchar(8)	
RCUI	Varchar(8)	
 VSAB	Varchar(20)	NN (PK)
RSAB	Varchar(20)	
SON	Text	
SF	Varchar(20)	
SVER	Varchar(20)	
VSTART	Varchar(8)	
VEND	Varchar(8)	
IMETA	Varchar(10)	
RMETA	Varchar(10)	
SLC	Text	
SCC	Text	
SRL	Numeric(10,0)	
TFR	Numeric(10,0)	
CFR	Numeric(10,0)	
CXTY	Varchar(50)	
TTYL	Varchar(300)	
ATNL	Text	
LAT	Varchar(3)	
CENC	Varchar(20)	
CURVER	Varchar(1)	
SABIN	Varchar(1)	
SSN	Text	
SCIT	Text	
created_on	Datetime	



<<http://link.informatics.stonybrook.edu/umls/SAB/NCI>>

MRSAT (attributes)


MRSAT		
CUI	Varchar(8)	
LUI	Varchar(10)	
SUI	Varchar(10)	
METAUI	Varchar(50)	
STYPE	Varchar(50)	
CODE	Varchar(50)	
 ATUI	Varchar(11)	NN (PK)
SATUI	Varchar(50)	
ATN	Varchar(50)	
SAB	Varchar(20)	
ATV	Text	
SUPPRESS	Varchar(1)	
CVF	Numeric(10,0)	
created_on	Datetime	

Defines the subject

Defines the predicate

Attribute value

MRREL (relationships)

MRREL		
CUI1	Varchar(8)	
AUI1	Varchar(9)	
STYPE1	Varchar(50)	
REL	Varchar(4)	
CUI2	Varchar(8)	
AUI2	Varchar(9)	
STYPE2	Varchar(50)	
RELA	Varchar(100)	
 RUI	Varchar(10)	NN (PK)
SRUI	Varchar(50)	
SAB	Varchar(20)	
SL	Varchar(20)	
RG	Varchar(10)	
DIR	Varchar(1)	
SUPPRESS	Varchar(1)	
CVF	Numeric(10,0)	
created_on	Datetime	

Defines the object

Defines the subject

Defines the predicate

Example: Montelukast Sodium in NCI

- NCI Thesaurus is not part of RxNorm
- NCI Thesaurus is well structured in RDFS
 - [available separately as OWL](#)
- [Allegrograph's Gruff](#) is a tool being used to explore the UMLS
- CUIs allow linking across different source vocabularies

Gruff - An AllegroGraph Browser - C:/rdfstore/umls/ (read / write)

File View Add Link Remove Layout Select Inclusion Options Layout Options Drawing Options Table & Query Options Help

Montelukast Sodium Revisit ← → Show All Triples

Property	Value
AUI	A10791891
CAS REGISTRY	151767-02-1
CODE	C47625
Comment	The orally bioavailable monosodium salt of montelukast, a selective cysteinyl leukotriene receptor antagonist with anti-inflammatory and bronchodilating activities. Montelukast selectively and competitively blocks the cysteinyl leukotriene 1 (CysLT1) receptor, preventing binding of the inflammatory mediator leukotriene D4 (LTD4). Inhibition of LTD4 activity results in inhibition of leukotriene-mediated inflammatory events including: migration of eosinophils and neutrophils; adhesion of leukocytes to vascular endothelium, monocyte and neutrophil aggregation; increased airway edema; increased capillary permeability; and bronchoconstriction. The CysLT1 receptor is found in a number of tissues including
Concept in subset	A12794052
CONTRIBUTING SOURCE	FDA
FDA UNII CODE	U103J18SFL
Has CUI	C0380447
Has free acid or base form	A12815994
Has SAB	NCI
Has Term Type	PT
Has tradename	A10770412
Isa	A7664876
ISPREF	Y
Label	Montelukast Sodium
PDQ CLOSED TRIAL SEARCH ID	593502
PDQ OPEN TRIAL SEARCH ID	593502
PREFERRED NAME	Montelukast Sodium
SUPPRESS	N
Type	AUI
is Has salt form of	A12815994
is Inverse isa of	A7664876
is Subset includes concept of	A12794052
is Tradename of of	A10770412

Click the righthand column to visit that resource in the table view AND add the triple to the graph view. Shift-click the righthand column to ONLY add the node to the graph. Control-click to ONLY visit the resource in the table. Control-shift-click a a URL to visit it in your web browser. Shift-click the left column to add every node und

"151767-02-1"

AUI	→
CAS REGISTRY	→
CODE	→
Comment	→
Concept in subset	→
Contraindicated drug	→
CONTRIBUTING SOURCE	→
FDA UNII CODE	→
Has CUI	→
Has free acid or base form	→
Has ingredient	→
Has SAB	→
Has STN	→
Has Term Type	→
Has tradename	→
Isa	→
ISPREF	→
MESH UI	→
NDC	→
Va product component of	→

AUI: identifies a string down to its original source vocabulary

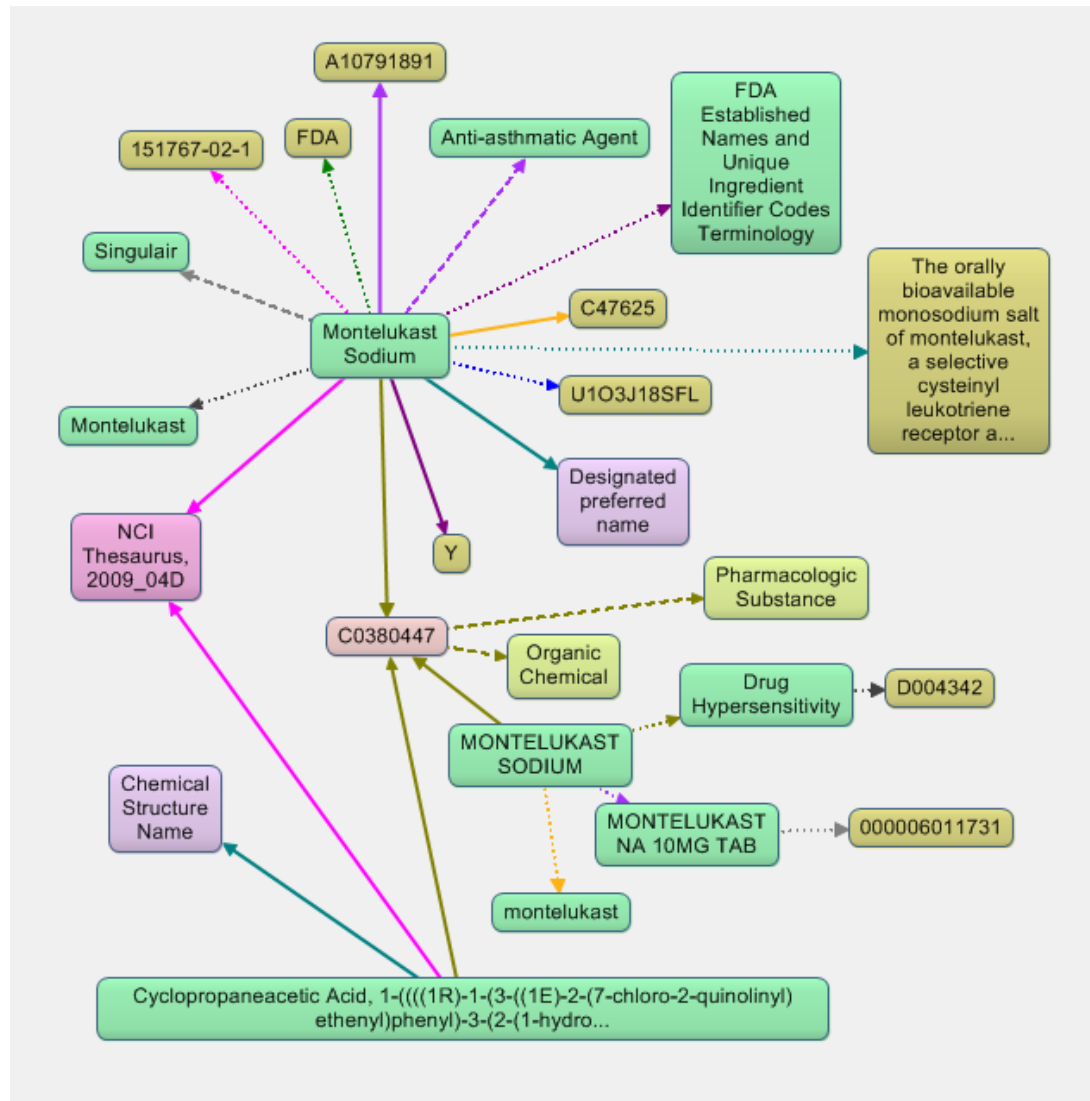
CUI: is a unique concept identifier it ties concepts across vocabularies

SAB is the source vocabulary identifier

Semantic Type

TTY (term type)

Literal



Future Steps:

Integrating with RxNorm and LODD

- Enhance RxNorm RDFS conversion
 - links to UMLS CUIs
 - Semantic types
- Define subset of the UMLS that can be published as part of the LODD
 - 110,169,095 triples in internal draft publication but includes non releasable sources
- Links from UMLS and RxNorm into other datasets in the LODD cloud

Acknowledgements

- Supreet Padhi and Jakub Pezacki
 - Worked on utilizing the UMLS for clinical rule extraction
- Erich Bremer development of RDF visualization tool