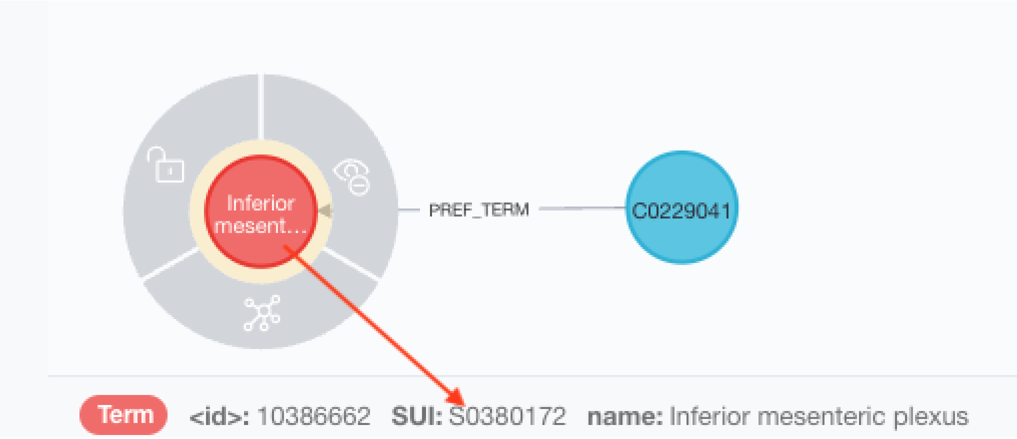
UBKG: removing SUIs - no upside, all downside

Simmons, Alan <alan.simmons@pitt.edu>
Mon 11/14/2022 7:00 AM

To: Shirey, Bill <shirey@pitt.edu>;Silverstein, Jonathan <j.c.s@pitt.edu>;Nemarich, Chris M <nemarichc@chop.edu>;Stear, Benjamin J <STEARB@chop.edu>;Taylor, Deanne M <TAYLORDM@chop.edu>;Mohseni Ahooyi, Taha <MOHSENIAHT@chop.edu>;Wenger, Eric D <WENGERE@chop.edu>

In the November 10 meeting, we discussed removing the SUI (permanent string identifier) as a property of Term nodes.



Term <id>: 10386662 **SUI:** S0380172 **name:** Inferior mesenteric plexus

The arguments for removing SUIs were:

1. For non-UMLS ontologies, we have to create base64-encoded string values for SUIs.
2. The SUI mattered less than the actual term (name property). We thought it unlikely that anyone would want to analyze by SUI.
3. The associated files (e.g., SUIs.csv, CODE-SUIs.csv, CUI-SUIs.csv), were large.

I took a look at what would be required. After my analysis, I can't see any benefit to removing SUIs; however, I do see many risks and costs to doing so.

Analysis

SUI information is a fundamental part of the current structure of the KG. The relevant CSV files are:

1. **SUIs.csv**
    1. Contains the SUI and the text of terms.
    2. Imported to build the T**erm** nodes.
    3. The file is large because the there are many terms, with long strings. We would not reduce the file much from removing a single column.
    4. Example:

SUI:ID	name
S17175117	1,2-dipalmitoylphosphatidylcholine

2. **CODE-SUIs.csv**
    a. Used to build edges between **Code** and **Term** nodes.
    b. Includes edge properties such as **Type** and **CUI. Type** is especially important, because it distinguishes the kinds of terms—e.g., synonyms (SY), preferred terms (PT), etc.
    c. Replacing SUI with term string would actually *increase* the size of the file. The majority of the SUIs (those from the UMLS) require only 9 characters.
    d. The file is large because it contains all of the many types of terms for codes.
    e. Example:

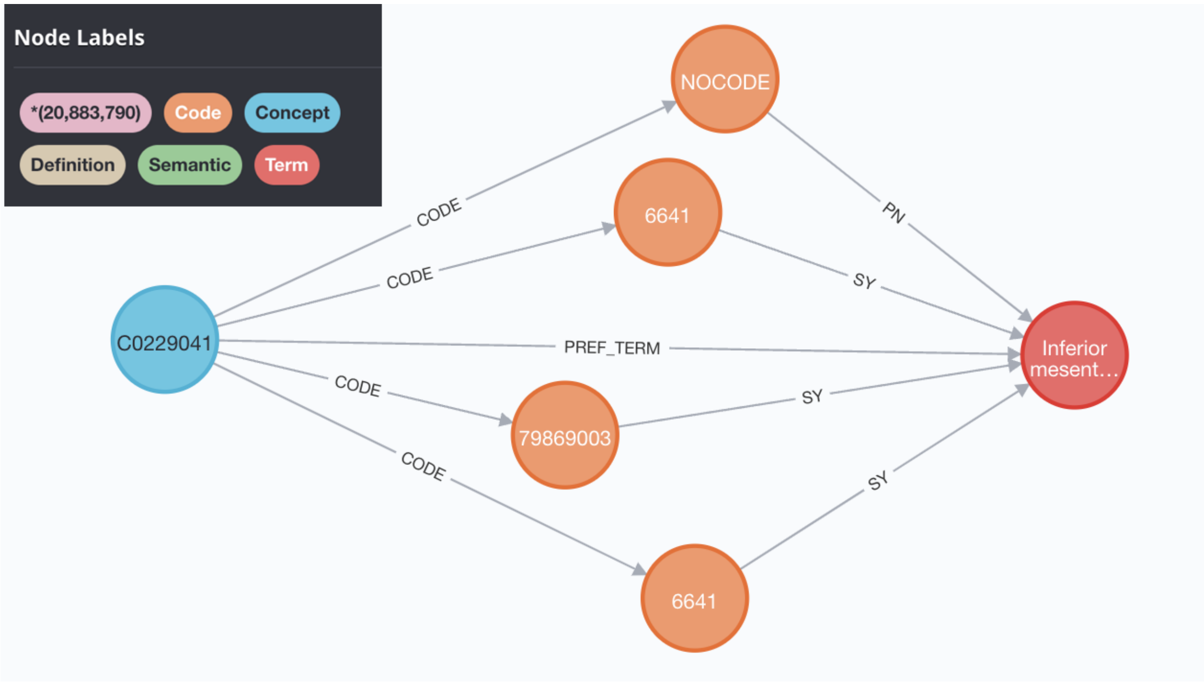
:END_ID	:START_ID	:TYPE	CUI
S17175117	RXNORM 1926948	IN	C0000039

3. **CUI-SUIS.csv**
    a. Used to build **PREF_TERM** edges between **Concept** and **Term** nodes.
    b. Replacing SUI with term string would increase the file size, as it would for CODE-SUIs.csv.
    c. Example:

:START_ID	:END_ID
C0000005	S0007492

Conclusion

1. SUIs help to reduce the size of CSV files that involve terms.
2. SUIs are integral to distinguishing terms by type.
3. SUIs may not be useful analytically, but they are useful structurally.



J. Alan Simmons
Solutions Architect

Department of Biomedical Informatics
University of Pittsburgh School of Medicine
5607 Baum Boulevard, Suite 500
Pittsburgh, PA 15206-3701

e: alan.simmons@pitt.edu
e: jas971@pitt.edu
t: (773) 220-5018