Random Notes about the Evidence Model (July 14, 2020)

The diagrams are vaguely like Entity-Relationship descriptions but that is purely a conversational framework for discussing the structure of the information. The PK (Primary Key) is a global identifier for a given node/entity; the FK refers to the PK of another node.

The "*provided_by*" and the citation identifier should be pushed out of the Biolink Model *edge* node and into an *evidence* record, unless a continuing strong argument is made for the convenience of local referencing (e.g. of keeping the "publications" property in edges). Consider "EVIDENCE" to be first class nodes in knowledge graphs, independent of the EDGES which they support.

The mandatory core properties of an EVIDENCE node are kept simple. However, given the open-ended nature of JSON, Neo4j node properties, etc., it might not be unreasonable to consider the evidence type-specific details permitted as extra property fields a given evidence node.

That is, the called "TEXT_ALIGNMENT" need not be a separate node but rather, instantiated as a "subclass" of EVIDENCE node, with the details provided in optional fields. GO evidence code "TAS" is used here as a signal that the extra details (of a TEXT ALIGNMENT) may be present.

The TEXT_ALIGNMENT box roughly corresponds to the contents of the Semantic Medline Database "PREDICTION_AUX" table. The TEXT_EVIDENCE box has start and end character indices for each of the S, P and O fields (I showed things just once is purely for brevity of expression)

The SENTENCE and CITATION boxes represent information corresponding to the similarly named Semantic Medline Database tables (although the corresponding CITATION needs to be augmented with full citation details from PubMed). The list of authors is also split out into two model boxes, PERSON and AUTHOR, the latter representing a "join relationship" to the CITATION.

Bill Baumgartner proposed the idea of a "hash" key for edges and evidence. The idea could be applied here mainly thinking in terms of indexing the edges by hashing of their canonical Subject-Predicate-Object ("S-P-O") CURIE or IRI's.

Anyhow, the above model is for "text mined" evidence. The intent is to keep the 'EVIDENCE" node core details agnostic as to evidence types, so that other kinds of evidence could be added to graphs in the future.

One possible additional field to add to EVIDENCE or TEXT_ALIGNMENT is a "score" field of some kind(?).

The *citation_id* might be the primary DOI, if it exists. As important as PubMed IDs are, we might not want to make them the default primary citation id.

The PERSON IRI, could perhaps be the person's ORCID if they have it. Not sure what to use otherwise?

PERSON, AUTHOR, CITATION could perhaps be hosted in a standalone NCATS Knowledge Graph or Service which tracks citations.