Unified Biomedical Knowledge Graph (UBKG)

- The UBKG, a knowledge graph, represents a set of interrelated concepts from biomedical ontologies and vocabularies.
- The UBKG integrates data from the National Library of Medicine's Unified Medical Language System (UMLS) with collections of assertions derived from ontologies or vocabularies that are external to the UMLS.
  - Assertions ≡ triples, or subject-predicate-object relationships.
- Sources and Source Abbreviations (SABs):
  - Source (e.g. DCC, owner) of a set of assertions is identified with a SAB
  - SAB examples:
    - PUBCHEM
    - IDG
    - UBERON
- The Rich Release Format (RRF) files contain information on source vocabularies or ontologies, codes, terms, and relationships both with other codes in the same vocabularies and with UMLS concepts.

Source: https://ubkg.docs.xconsortia.org/

Unified Biomedical Knowledge Graph (UBKG)

- For the purposes of the UBKG, an ontology is a collection of relationships that can be identified between the entities of a particular domain of knowledge.
  - The terms ontology and set of assertions can be used interchangeably for the purposes of the UBKG.
- An assertion includes two entities and a relationship.
  - The two entities are usually described as the **subject** and the **object**.
    - A particular entity may be part of more than one ontology. E.g., genes figure in a number of biomedical ontologies.
  - The relationship is expressed as the **predicate** of the assertion.
  - An assertion is thus a statement in the form: ***subject-predicate-object***
- The domain for an ontology is also known as its namespace.
  - Domains are identified with acronyms called Source Abbreviations (SABs).

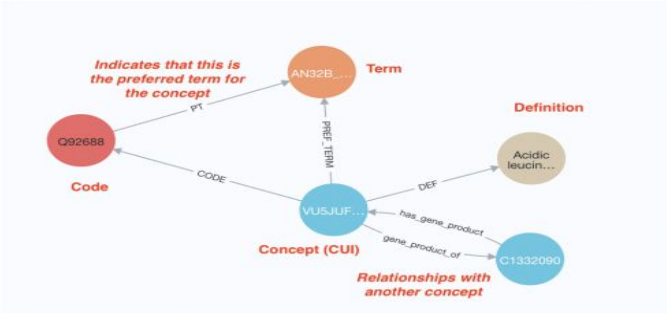
Source: https://ubkg.docs.xconsortia.org/

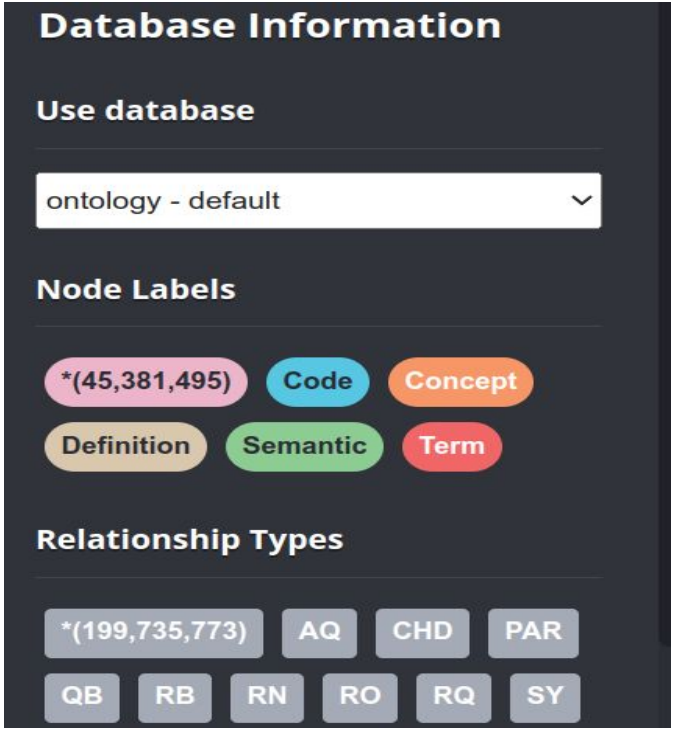
CFDE Data Distillery Data Dictionary

- The purpose of the Data Distillery Knowledge Graph (DDKG) is to link multiple sources of expertly curated data, thus providing data integration across multiple Common Fund data coordinating centers (DCCs).
- For the first phase of the project, the participating DCCs have submitted 29 different datasets for integration into the DDKG
- List of all DCCs and details of their data can be found here: https://github.com/nih-cfde/data-distillery/blob/main/DataDistilleryDataDictionary.md

Unified Biomedical Knowledge Graph (UBKG) database

- Latest UBKG database (10 sep 2023) information:
  - URL: http://chiltepin.health.unm.edu:5000/browser/
  - Neo4j port: 5500 and password: Hello@001
- UBKG comprises the following nodes and edges:





Unified Biomedical Knowledge Graph (UBKG) database

Number of nodes in UBKG: 45,381,495

```
MATCH ()
RETURN count(*) as count
```

Number of edges in UBKG: 199,735,773

```
MATCH ()-[r]->()
RETURN count(r) as count
```

Count per Node Label in UBKG:

```
MATCH (n)
RETURN labels(n) as NodeLabel, count(labels(n)) as
LabelCount;
```

"NodeLabel"	"LabelCount"
["Concept"]	15639530
["Semantic"]	127
["Definition"]	705122
["Term"]	12462662
["Code"]	16574054

Count per Relationship Type in UBKG:

```
MATCH ()-[r]->()
RETURN TYPE(r) AS relationshipType, COUNT(r) AS
relationshipCount;
```

There are 2101 Relationship Types in UBKG. List of relationship type

Nodes and Edges in UBKG

- Assertion is identified by SAB
- List of SAB: https://ubkg.docs.xconsortia.org/contexts/#data-distillery-context
- “Nodes” table on right shows the corresponding element in HuBMAP for a given field in nodes file. E.g. To access node_label, you need to use Term node with PT relationship.

Nodes

Field	Corresponding element in HuBMAP graph
node_id	**Code** node
node_label	• **Term** node • **PT** relationship PT (preferred term)
node_definition	• **Definition**
(optional)	node • **DEF** relationship
node_synonyms (optional)	• **Term** node • **SYN** relationship
node_dbxrefs (optional)	Cross-references

Edges

Field	Corresponding element in UBKG
subject	**Code** node
predicate	relationships
object	**Code** node

Cypher queries: UBKG vs. Local KG

- The cypher queries to fetch the similar information from Local (UNM) KG and UBKG will be different as one needs to use SAB to select a particular domain in UBKG.
    - The following cypher queries are some examples.

UBKG

```
MATCH (uniprot_code:Code
{SAB:'UNIPROTKB'})-[:PT]-(term:Term)-[r]-(concept:Concept)-[:D
EF]-(def:Definition)
WHERE uniprot_code.CODE = "Q02318"
RETURN uniprot_code.CodeID, term.name, def.DEF,
concept.CUI
```

Local KG

```
1 MATCH (n:Protein)
2 WHERE n.node_id ="UNIPROTKB Q02318"
3 RETURN n
```

Graph | Table | Text | Code

n


```
{
  "identity": 255804,
  "labels": [
    "Protein"
  ],
  "properties": {
    "node_namespace": "IDG",
    "node_synonyms": "Sterol 26-hydroxylase, mitochondrial",
    "node_dbxrefs": "ENSP00000258415",
    "node_label": "CYP27A1",
    "node_definition": "MK-24",
    "node_id": "UNIPROTKB Q02318"
  },
  "elementId": "255804"
}
```

Cypher queries

- The following GitHub page has several cypher queries to explore the Data Distillery Knowledge Graph (DDKG): https://github.com/TaylorResearchLab/CFDE_DataDistillery/blob/main/user_guide/CFDE_DataDistillery_UserGuide.md

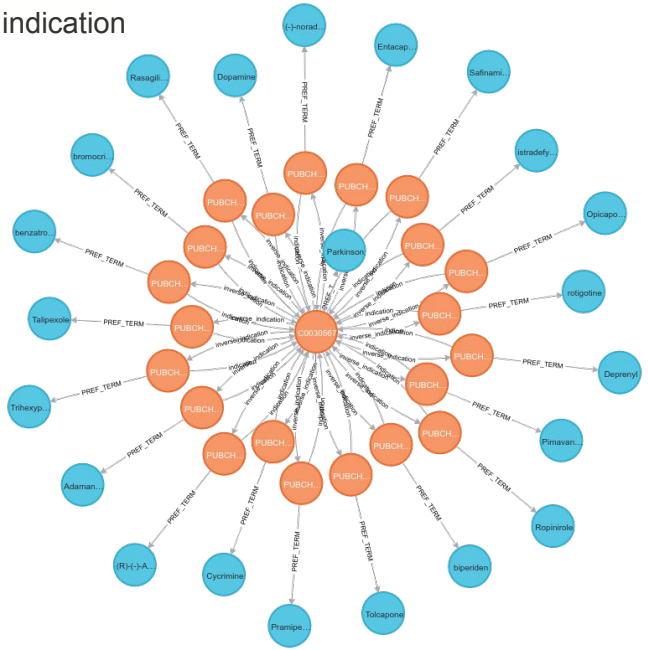
- The following GitHub page has some cypher queries to fetch information from UBKG using IDG as SAB: https://github.com/unmtransinfo/cfde-distillery/blob/main/cql/UBKG_cypher_examples_sept13.md

- Neo4j performs query caching.
  - The subsequent run of the identical query will return results much more quickly.

Cypher queries

- Compounds associated with “Parkinson’s Disease” via the IDG “indication” relationship:

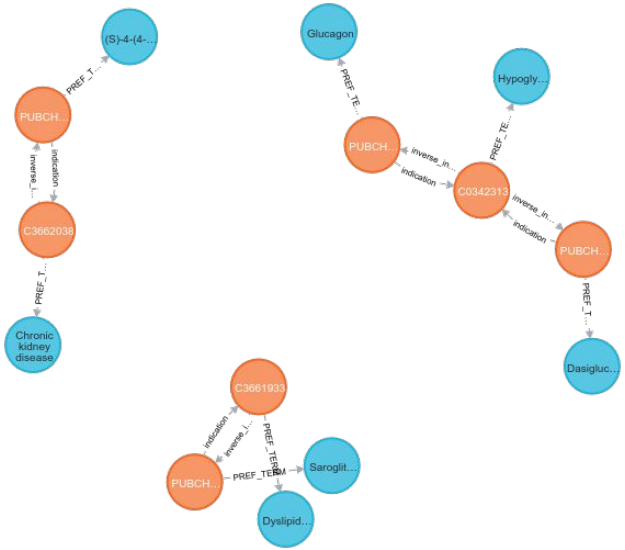
```
MATCH (compound_term:Term)-[:PREF_TERM]-(compound_concept:Concept)-[:indication
{SAB:'IDGD'}]-(disease_concept:Concept)-[:PREF_TERM]-(disease_term:Term)
WHERE disease_term.name = "Parkinson Disease"
RETURN *
```



Cypher queries

- Disease terms containing the string “diabetes” and linked via the IDG indication relationship:

```
MATCH (t1:Term)<-[:PREF_TERM]-(p1:Concept)-[:indication]-(c2:Concept)--(t2:Term)
WHERE t1.name
CONTAINS 'diabetes'
RETURN * ;
```



Cypher queries

- Showing compounds associated with proteins via the IDG “bioactivity” relationship:

```
MATCH (compound_term:Term)-[:PREF_TERM]-(compound:Concept)-[r:bioactivity
{SAB:"IDGP"}]-(protein:Concept)-[:PREF_TERM]-(protein_term:Term)
RETURN *
LIMIT 10
```