PheKnowLator Tutorial
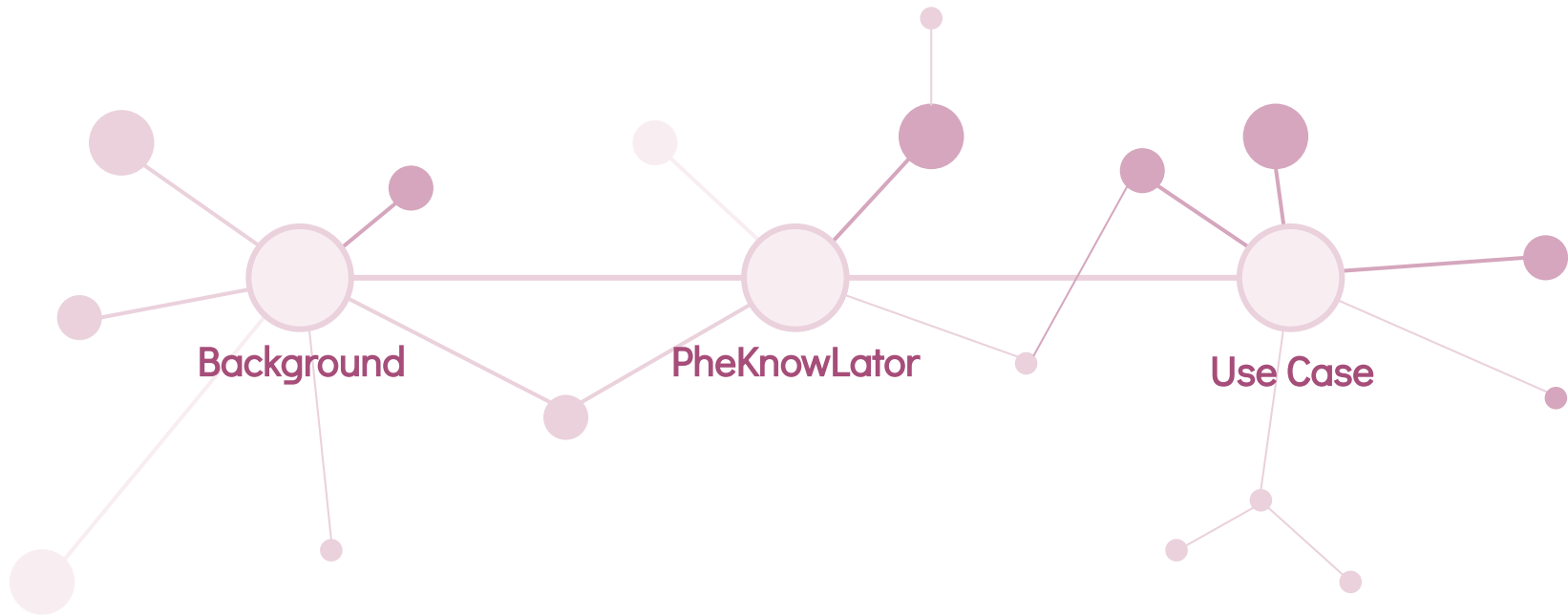# Entity Search

Tiffany J. Callahan, PhD

June 21, 2022
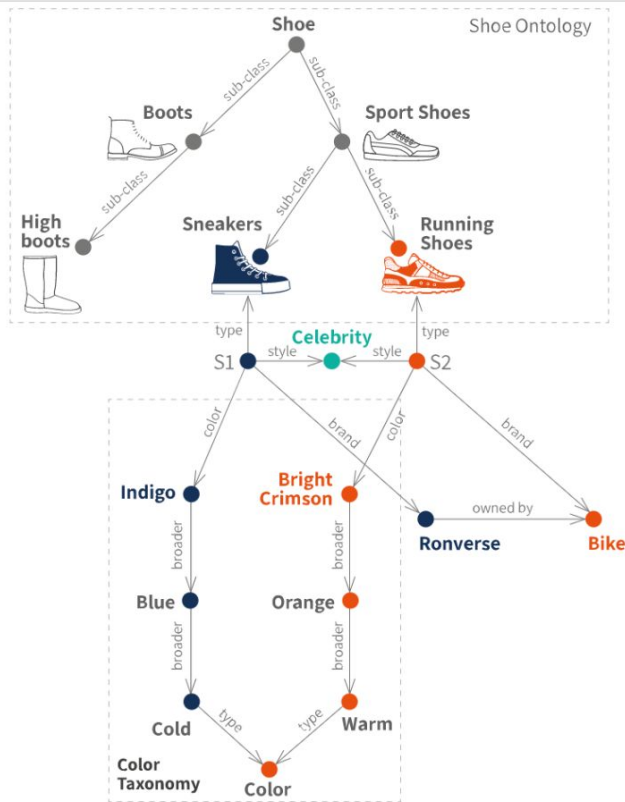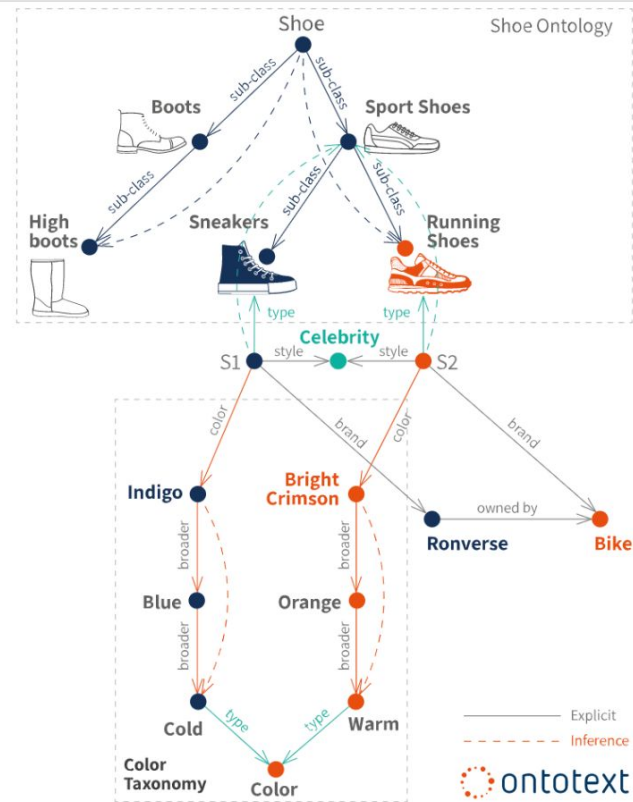
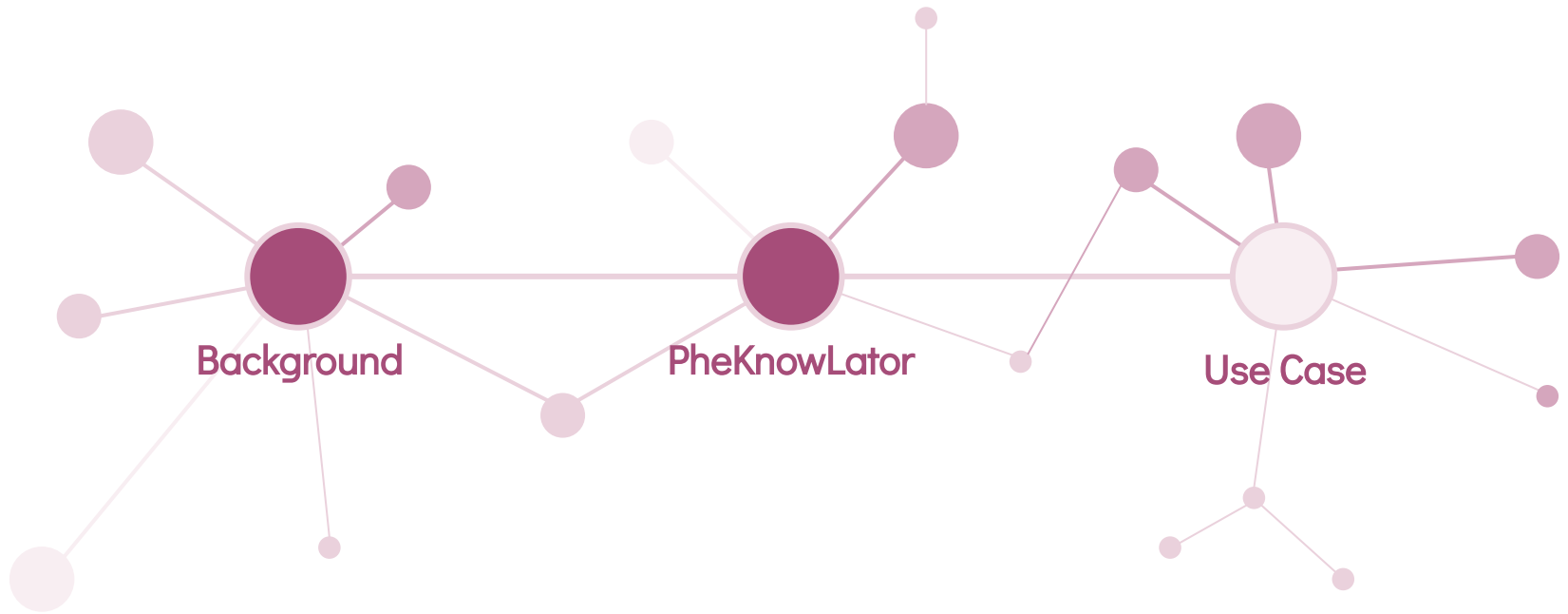Background   PheKnowLator   Use Case

Background

PheKnowLator

Use Case

Image: https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/

Background
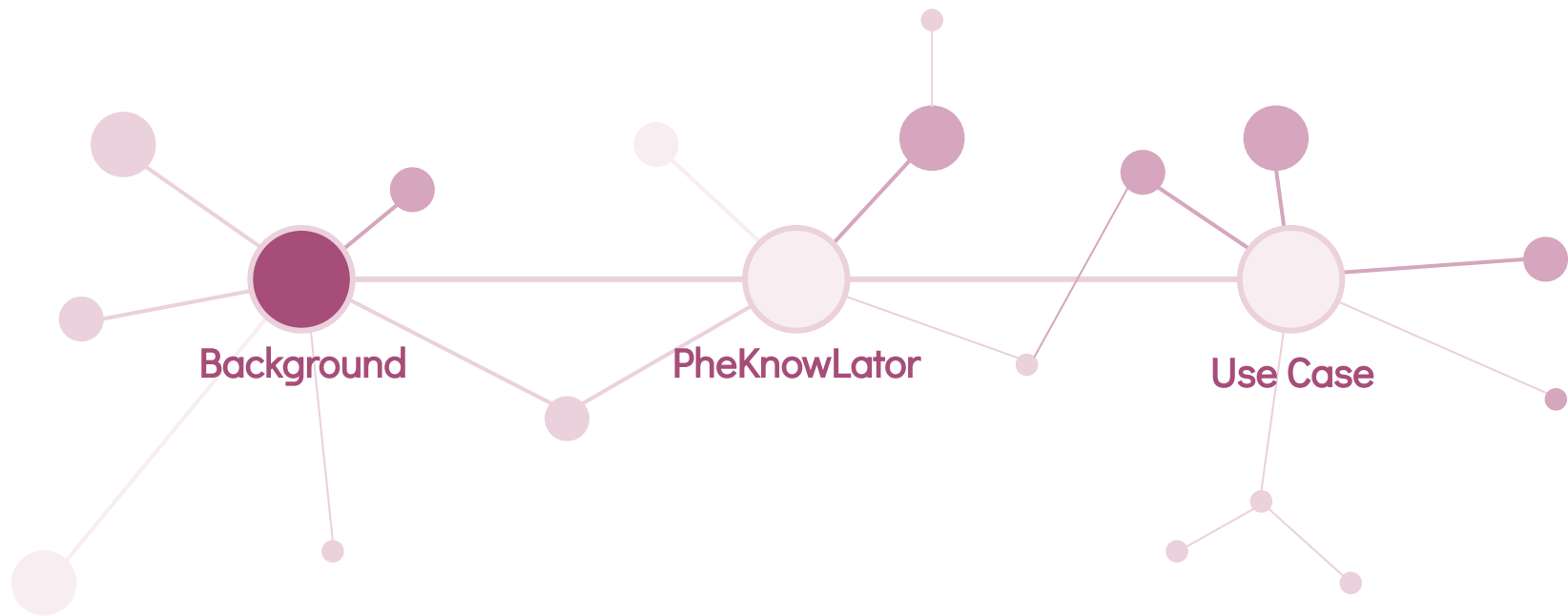
PheKnowLator

Use Case

What is PheKnowLator (PKT)

Knowledge graphs integrate disparate data, can help decipher complex processes, and have frequently been used to systematically interrogate complex systems[1]

PheKnowLator (Phenotype Knowledge TransLator) addresses unsolved challenges for constructing open-source knowledge graphs[2,3]

What is PheKnowLator and How do I access it?

Software: Python 3 library designed to build large-scale FAIR biomedical knowledge graphs that are grounded in the Open Biomedical Ontology (OBO) Foundry ontologies

Benchmarks: Monthly knowledge graphs to represent Human Disease Mechanisms (Google Cloud Storage)

Papers and Presentations: Zenodo Community that contains preprints, PowerPoint Slides, and recorded presentations

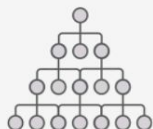[1]PMID:21414488; [2]PMID:32637040; [3]PMID:33954284

Data Download

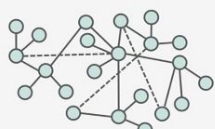Clinical Records

Experimental Results

Biomedical Ontologies

Linked Open Data

Knowledge Graphs

Data Preparation

Data Cleaning

Identifier Mapping

ID 1   ID 3
ID 2

Metadata Generation

Entity Identifier — Labels / Definitions / Synonyms
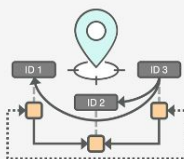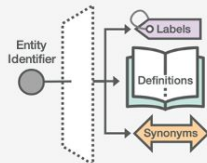
Concept Annotation

sample	DDX11L21	OR4G4P	FAM138A
Mammary tissue – hTERT-HME1	0.1326	1.3369	2.3568
...	...	...	...

Mammary Gland (UBERON_0001911)
Breast (UBERON_0000310)
Mammary Gland Glandular Cell (CL_1001586)
hTERT-HME1 (CLO_0004291)
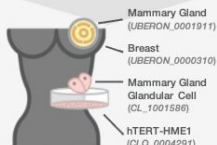
Knowledge Graph Construction

Build Edge Lists

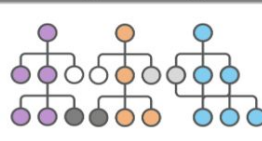| chemicalID	Symbol	GeneForms	Organism	PubMed
C534883	MYC	protein	mus musculus	21344345
C534883	MYC	protein	homo sapiens	32184358
C553100	TFEB	mRNA	homo sapiens	25716159
C553100	TFEB	protein	homo sapiens	26474267

Apply Filtering Criteria
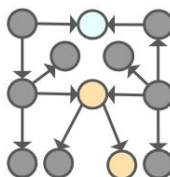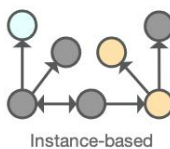
Merge Support Data

Verify Identifiers

CHEBI_105930 → ncbigene_4609
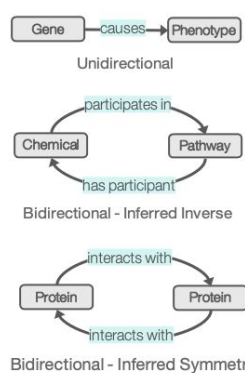CHEBI_94104 → ncbigene_7942

Merge Ontologies

OWLTools

Select Hyperparameters

Knowledge Model

Knowledge Model

Instance-based

Subclass-based

Relation Strategy

Gene —causes→ Phenotype
Unidirectional

Chemical —participates in→ Pathway
← has participant —
Bidirectional - Inferred Inverse

Protein —interacts with→ Protein
← interacts with —
Bidirectional - Inferred Symmetry

Semantic Abstraction
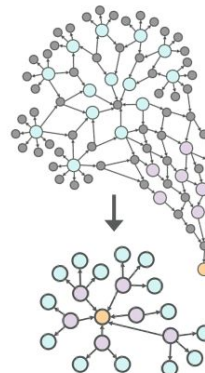
Generated Output

Quality/Provenance Reports

Knowledge Graphs

PKL   OWL NT   TXT

Edge Lists
subject   relation   object
PR_01023   RO_0003302   HP_0002511
GO_0000228   BFO_0000050   GO_0005634
VO_0001966   OBI_0000304   VO_0000570
...

Node and Relation Metadata

relation
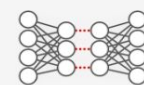Subject → Object
Triple

Metadata (.txt)
id: unique identifier (int)
label: entity name (str)
definition: description (str)
synonym: alt terms (str)

id: unique identifier (int)
label: entity name (str)
definition: description (str)
synonym: alt terms (str)

triple: UUID (str)
type: context tag (str)

Benchmarks

Knowledge Graphs

Embeddings

0.4567 0.0895 0.0084 … 0.5689
0.2356 0.1121 0.0008 … 0.3121
0.0023 0.3154 0.9999 … 0.2154
0.1122 0.8874 0.1235 … 0.1211

Endpoints

Cloud File Storage

Graph Databases

SPARQL Endpoint

PheKnowLator

Data Access

Use Case

Use Case

What: What can a PheKnowLator benchmark tell us about a drug and an outcome?
- Mechanisms of human disease (780,753 nodes; 5,072,062 edges)
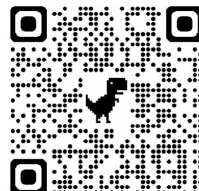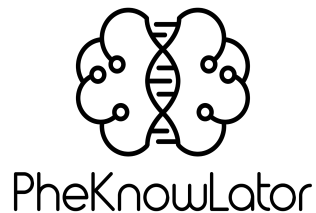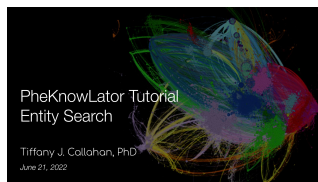- OMOP2OBO used to map OMOP concepts to OBO ontology terms

Why: Knowledge of the molecular mechanisms underlying adverse events may aid prevention
- Does the knowledge graph contain information on a known adverse event (positive)?
- What does the knowledge graph know about a drug known to not cause an outcome (negative)?

How: Self-contained Jupyter Notebook

Questions?

@Tiff_callahan   tc3206@cumc.columbia.edu