KG-Microbe: a reference knowledge-graph and platform for harmonized microbial information

Marcin P. Joachimiak[1], Harshad Hegde[1], William D. Duncan[1], Luca Cappelletti[2], Justin T. Reese[1], Anne E. Thessen[3], Christopher J. Mungall[1]

[1] Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory
[2] Università degli Studi di Milano, Milan, Italy
[3] College of Agricultural Sciences, Oregon State University

github.com/Knowledge-Graph-Hub/kg-microbe

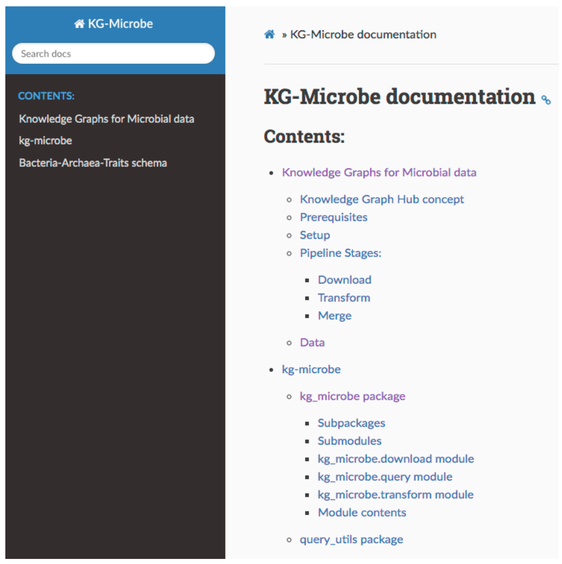
Knowledge Graph construction template and build pipeline

knowledge-graph-hub.github.io/
github.com/Knowledge-Graph-Hub/kg-dtm-template

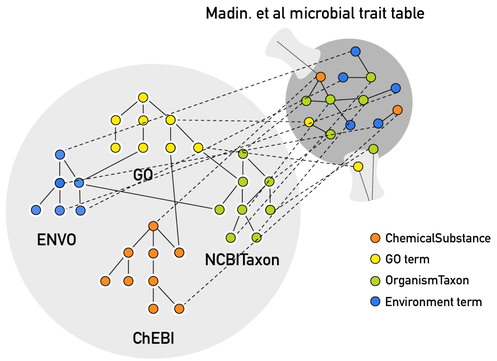
We used a KG construction pipeline template, covering standard cases for data ingestion, Named Entity Recognition (NER) tool integration for data standardization, tools for interacting with ontologies, as well as KG construction software for merging and summarizing KG data sources.

Auto-generated documentation

knowledge-graph-hub.github.io/kg-microbe

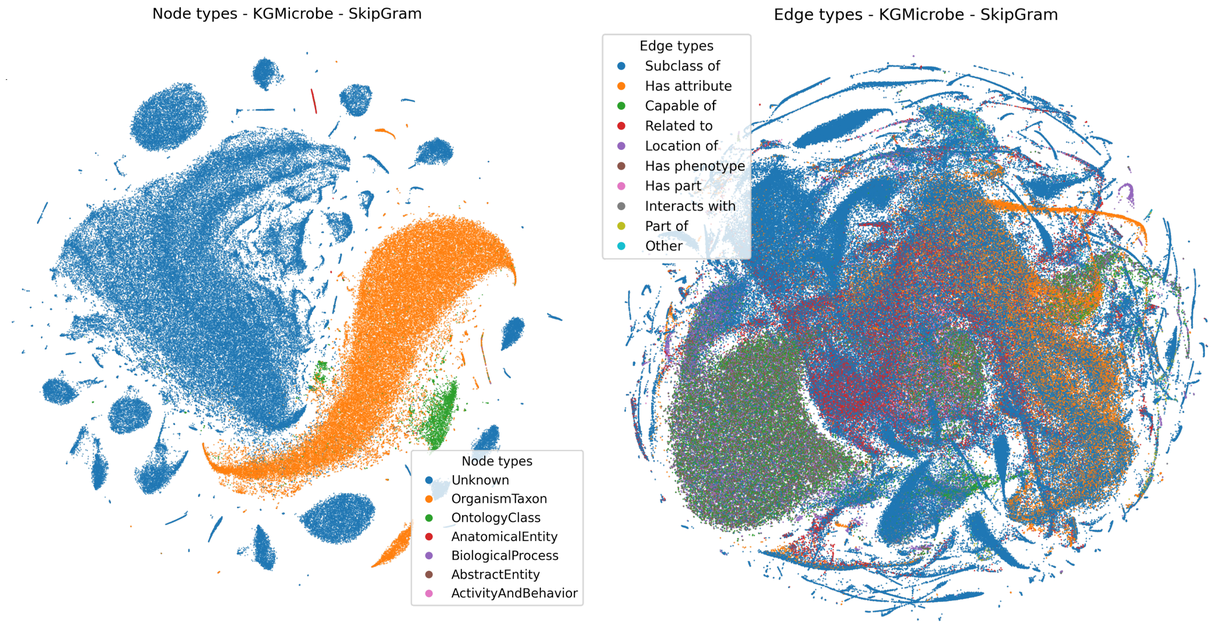


KG-Microbe data sources



Data sources were selected for relevance to microbial trait information. Supporting ontologies for biological functions, environmental terms, chemical entities, and organismal taxonomy were selected. Further data source ingests are prioritized by their potential for inter-linking data types and coverage of knowledge about microbial traits.

Knowledge Graph embeddings and visualization



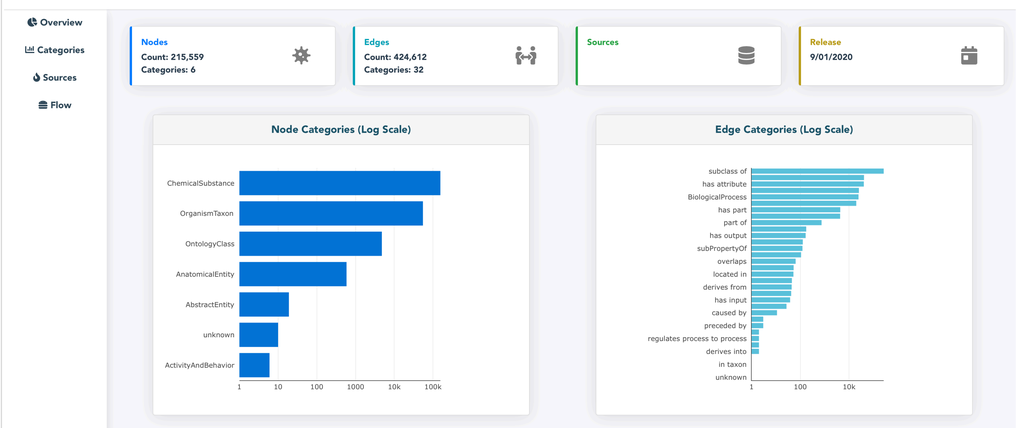
Graph embeddings were generated using a node2vec implementation from the embiggen package (github.com/monarch-initiative/embiggen). Node and edge type categories are sourced from the Biolink Model (biolink.github.io/biolink-model), and used to standardize the KG-Microbe entities and relations.

Public release directory structure

kg-hub.berkeleybop.io/kg-microbe

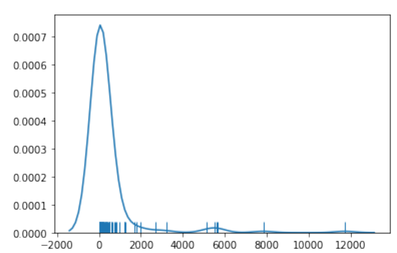
- GitHub integration with Jenkins: builds and deploys kg monthly.
- Includes raw and transformed individual data sources.
- Merged KG summary statistics.

Knowledge Graph Dashboard



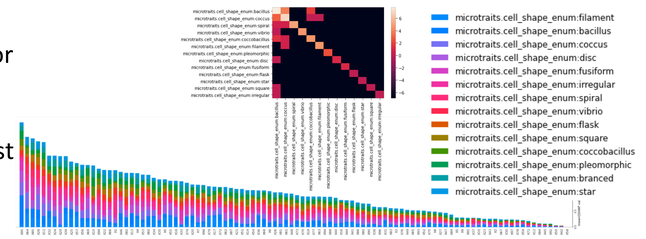
Feature frequency

A histogram of feature frequency across the collection of taxa in KG-Microbe. Note that the majority of features are sparse, hence a graph and graph embedding approach is appropriate.



Prediction of microbial shape using graph embeddings and gradient-boosted decision trees

Graph embeddings for KG-Microbe nodes were used as input features for a gradient-boosted decision tree classifier (CatBoost). Known taxa-shape associations were split into test and training, negative associations were generated via random mismatch. Performance was evaluated on 20% with-held data and embeddings generated without the held-out data.



	precision	recall	f1-score	support
microtraits.cell_shape_enum:bacillus	0.99	0.98	0.98	2343
microtraits.cell_shape_enum:coccus	0.93	0.94	0.94	103
microtraits.cell_shape_enum:spiral	0.95	0.98	0.97	1123
microtraits.cell_shape_enum:vibrio	1.00	0.75	0.86	4
microtraits.cell_shape_enum:coccobacillus	0.99	0.98	0.98	96
microtraits.cell_shape_enum:filament	1.00	1.00	1.00	1
microtraits.cell_shape_enum:pleomorphic	1.00	1.00	1.00	3
microtraits.cell_shape_enum:disc	1.00	0.50	0.67	2
microtraits.cell_shape_enum:fusiform	1.00	1.00	1.00	11
microtraits.cell_shape_enum:flask	1.00	0.97	0.99	157
microtraits.cell_shape_enum:star	1.00	0.50	0.67	2
microtraits.cell_shape_enum:square	1.00	1.00	1.00	1
microtraits.cell_shape_enum:irregular	1.00	0.99	1.00	102
accuracy			0.98	3948