BioBricks-OKG: An Open Knowledge Graph for Cheminformatics and Chemical Safety

Zakariyya Mughal

Insilica

2023 Nov 06

OpenTox Virtual Conference 2023

Award: NSF SBIR 2333728

Problem space

Before you can answer a particular cheminformatics question:

P0 Many different data sources for cheminformatics and chemical safety.

Problem space

Before you can answer a particular cheminformatics question:

P0 Many different data sources for cheminformatics and chemical safety.

P1 Retrieving data sources for local analysis.

Problem space

Before you can answer a particular cheminformatics question:

P0 Many different data sources for cheminformatics and chemical safety.

P1 Retrieving data sources for local analysis.

P2 Differences in how data are represented.

Solution: tabular data

multiple data sources

local analysis

Solution: tabular data

multiple data sources

local analysis

BioBricks

Solution: graph data

multiple data sources

local analysis

Solution: graph data

multiple data sources

local analysis

interoperable data through common vocabularies, identifiers, and ontologies

Solution: graph data

multiple data sources

local analysis

interoperable data through common vocabularies, identifiers, and ontologies

BioBricks-OKG

Working as part of NSF Proto-OKN program (Award: NSF SBIR 2333728)

Working as part of NSF Proto-OKN program (Award: NSF SBIR 2333728)

Agency partner: NICEATM

BioBricks-OKG

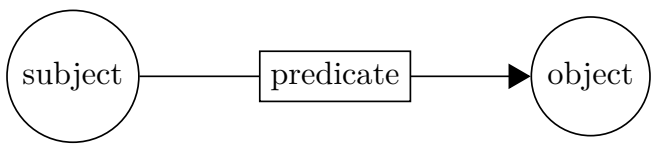
Working as part of NSF Proto-OKN program (Award: NSF SBIR 2333728)

Agency partner: NICEATM

Built on RDF metadata standard

RDF

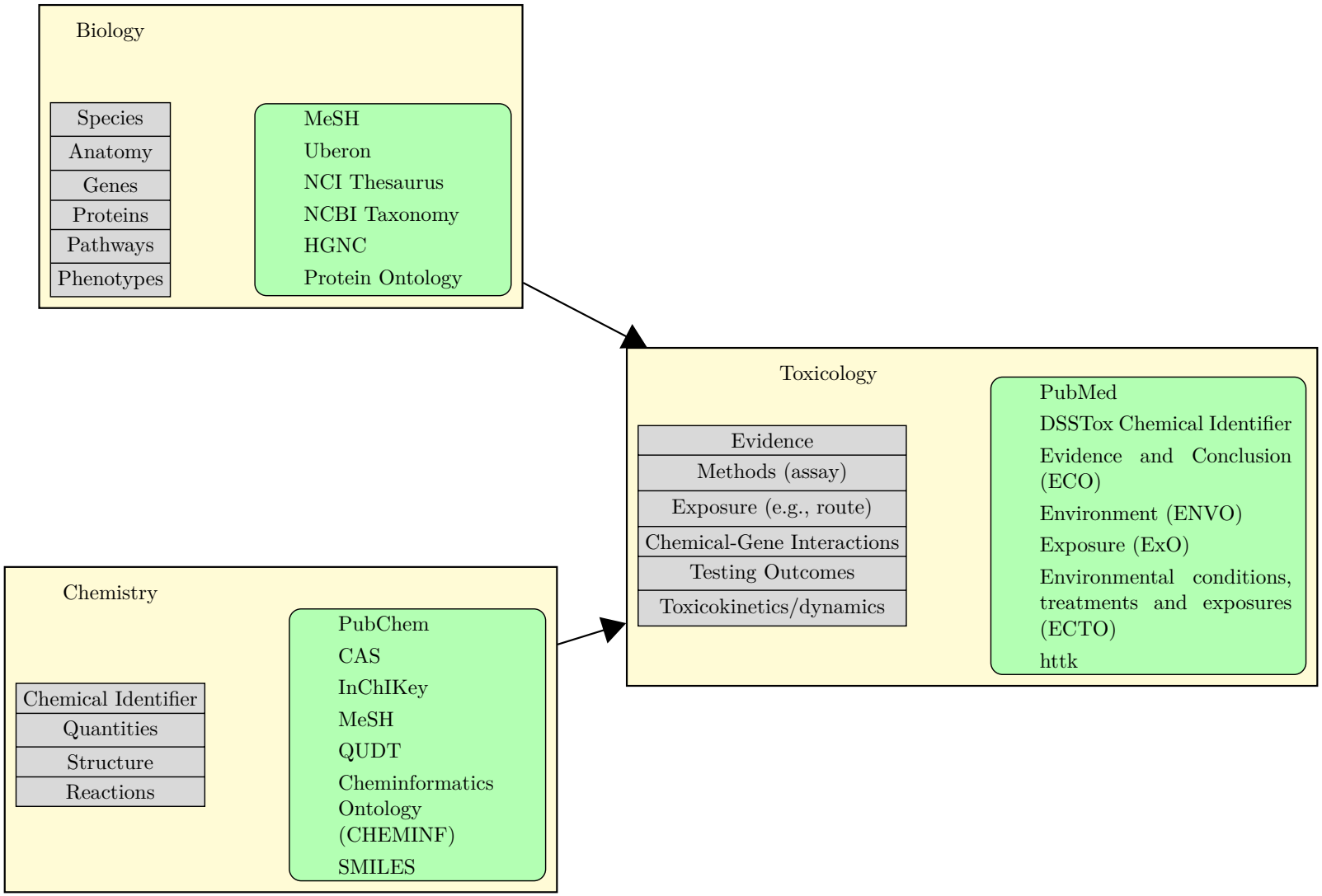
RDF based on triple relationships



Technology

1. Build on existing BioBricks infrastructure
2. Use tabular data as a data source (tabular brick)
3. Map tabular data to graph data (triple brick)
4. Load this triple brick into a triplestore graph database or use directly

Knowledge organization



Early prototype

datasets: ICE, tox21, CTDbase, MeSH, HGNC, CosIng

Early prototype

datasets: ICE, tox21, CTDbase, MeSH, HGNC, CosIng

Looking for

- New datasets
- New use cases
- Improvements to data modeling