The Federal NanoEHS Data Landscape: Machine-learning and Large Language Methods to Improve Data Accessibility, Interoperability and Semantic Queries

Pranav Singh, UCSD
***Mentor*: Dr. Holly Mortensen, US EPA**

Invited Presentation
ICCVAM Public Meeting
July 21, 2025

EPA Disclaimer: The views expressed in this presentation are those of the author(s) and do not necessarily represent the views or policies of the Agency

Overview

- Motivations + Recap of Prev. Work
    - NKB RDF (https://doi.org/10.12688/f1000research.141056.1)
    - OntoSearcher
- Updates
    - CPSC RDF Creation
    - Large Language Model(LLM) assisted querying across NKB, CPSC, and NIOSH graphs
    - Link Prediction for missing links in NKB Graph
- Future Work

Motivations: Integrating New Data + AI with NanoEHS Data

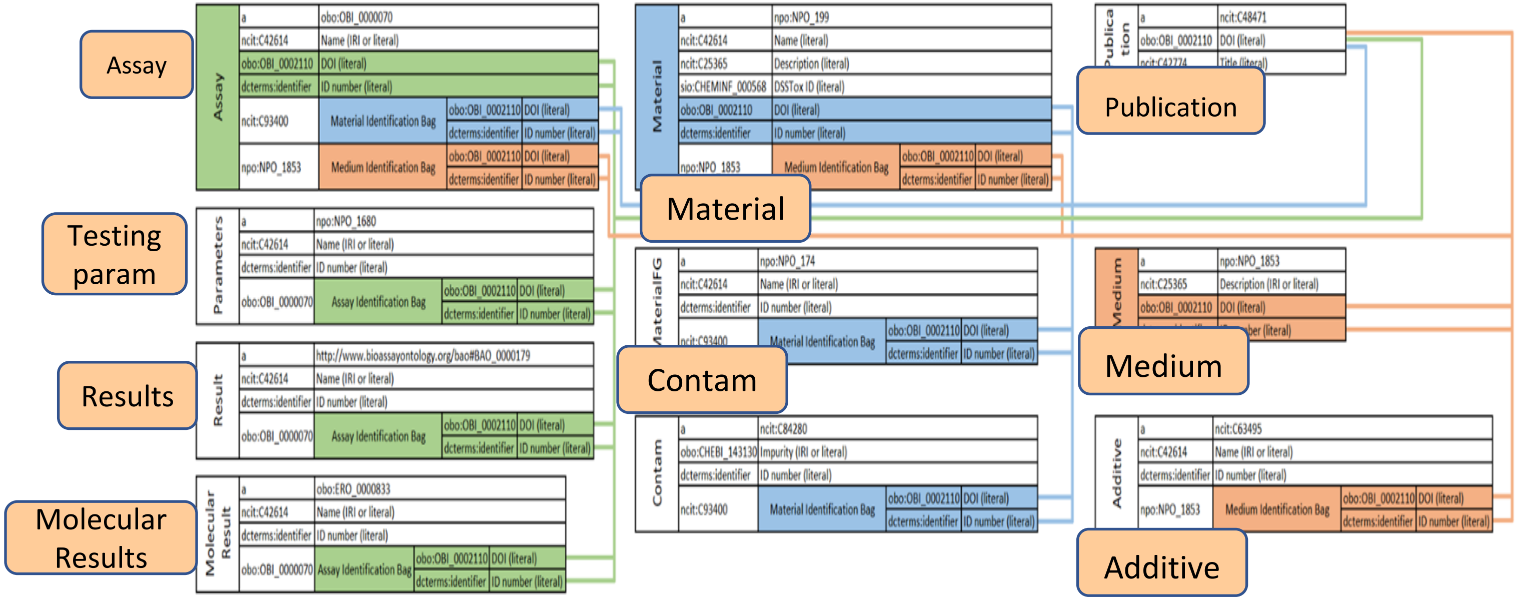
- Continuing to build a network of related federal partner data sources
- Leverage EPA OntoSearcher tool to create new knowledge graphs with similar semantic definitions
- Build integrations with existing artificial intelligence technologies to make data access easier
- Generate possible new directions for R&D based on existing research connections/link as nanoinformatics is still a growing field

Technical Challenges…

- How to Build Off of Previous RDF Creation
- How to Get From Natural Language to SPARQL/Cypher(Depending On Database)
- How to Convert RDF Structure into Feasible Format for Graph-based ML tasks

Key Terms/ Ideas

- Semantic mapping is a way of representing information (concepts or data) as a **graph**
- **NKB Semantic mapping:**



Assay

Assay	a	obo:OBI_0000070		
	ncit:C42614	Name (IRI or literal)		
	obo:OBI_0002110	DOI (literal)		
	dcterms:identifier	ID number (literal)		
	ncit:C93400	Material Identification Bag	obo:OBI_0002110	DOI (literal)
			dcterms:identifier	ID number (literal)
	npo:NPO_1853	Medium Identification Bag	obo:OBI_0002110	DOI (literal)
			dcterms:identifier	ID number (literal)

Testing param

Parameters	a	npo:NPO_1680		
	ncit:C42614	Name (IRI or literal)		
	dcterms:identifier	ID number (literal)		
	obo:OBI_0000070	Assay Identification Bag	obo:OBI_0002110	DOI (literal)
			dcterms:identifier	ID number (literal)

Results

Result	a	http://www.bioassayontology.org/bao#BAO_0000179		
	ncit:C42614	Name (IRI or literal)		
	dcterms:identifier	ID number (literal)		
	obo:OBI_0000070	Assay Identification Bag	obo:OBI_0002110	DOI (literal)
			dcterms:identifier	ID number (literal)

Molecular Results

Molecular Result	a	obo:ERO_0000833		
	ncit:C42614	Name (IRI or literal)		
	dcterms:identifier	ID number (literal)		
	obo:OBI_0000070	Assay Identification Bag	obo:OBI_0002110	DOI (literal)
			dcterms:identifier	ID number (literal)

Material

Material	a	npo:NPO_199		
	ncit:C42614	Name (literal)		
	ncit:C25365	Description (literal)		
	sio:CHEMINF_000568	DSSTox ID (literal)		
	obo:OBI_0002110	DOI (literal)		
	dcterms:identifier	ID number (literal)		
	npo:NPO_1853	Medium Identification Bag	obo:OBI_0002110	DOI (literal)
			dcterms:identifier	ID number (literal)

Contam

MaterialFG	a	npo:NPO_174		
	ncit:C42614	Name (IRI or literal)		
	dcterms:identifier	ID number (literal)		
	ncit:C93400	Material Identification Bag	obo:OBI_0002110	DOI (literal)
			dcterms:identifier	ID number (literal)

Contam	a	ncit:C84280		
	obo:CHEBI_143130	Impurity (IRI or literal)		
	dcterms:identifier	ID number (literal)		
	ncit:C93400	Material Identification Bag	obo:OBI_0002110	DOI (literal)
			dcterms:identifier	ID number (literal)

Publication

Publication	a	ncit:C48471
	obo:OBI_0002110	DOI (literal)
	ncit:C42774	Title (literal)

Medium

Medium	a	npo:NPO_1853
	ncit:C25365	Description (IRI or literal)
	obo:OBI_0002110	DOI (literal)
		ber (literal)

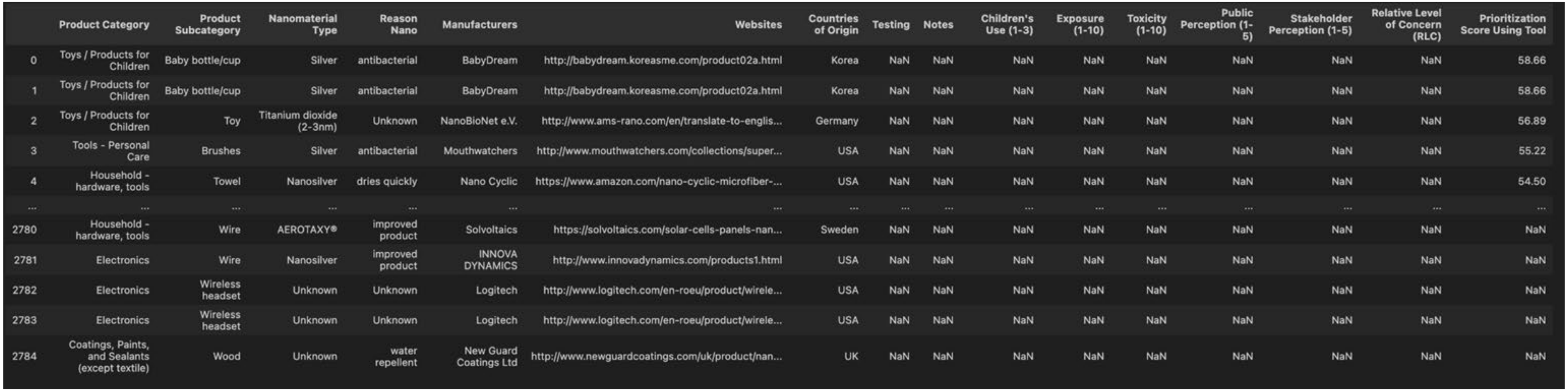
Additive

Additive	a	ncit:C63495		
	ncit:C42614	Name (IRI or literal)		
	dcterms:identifier	ID number (literal)		
	npo:NPO_1853	Medium Identification Bag	obo:OBI_0002110	DOI (literal)
			dcterms:identifier	ID number (literal)

CPSC RDF Creation ⇒ Why?

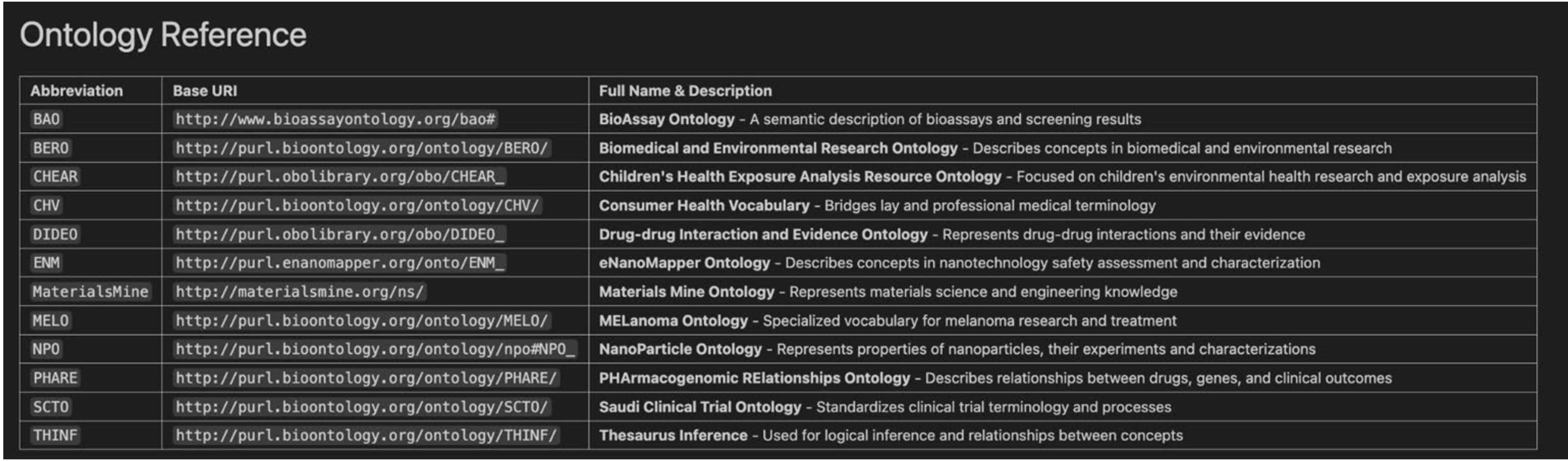
- Resource Description Frameworks(RDFs) are a great way to represent relational data even when predicates(column names) are not the same but are similar in concept
- Using the inherent integrated nature of the RDF, we can now re-use existing vocabulary to query across related graphs and extract related information about nanoparticles/ products in various domains/ scopes

CPSC RDF Creation



	Product Category	Product Subcategory	Nanomaterial Type	Reason Nano	Manufacturers	Websites	Countries of Origin	Testing	Notes	Children's Use (1-3)	Exposure (1-10)	Toxicity (1-10)	Public Perception (1-5)	Stakeholder Perception (1-5)	Relative Level of Concern (RLC)	Prioritization Score Using Tool
0	Toys / Products for Children	Baby bottle/cup	Silver	antibacterial	BabyDream	http://babydream.koreasme.com/product02a.html	Korea	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	58.66
1	Toys / Products for Children	Baby bottle/cup	Silver	antibacterial	BabyDream	http://babydream.koreasme.com/product02a.html	Korea	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	58.66
2	Toys / Products for Children	Toy	Titanium dioxide (2-3nm)	Unknown	NanoBioNet e.V.	http://www.ams-rano.com/en/translate-to-englis...	Germany	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	56.89
3	Tools - Personal Care	Brushes	Silver	antibacterial	Mouthwatchers	http://www.mouthwatchers.com/collections/super...	USA	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	55.22
4	Household - hardware, tools	Towel	Nanosilver	dries quickly	Nano Cyclic	https://www.amazon.com/nano-cyclic-microfiber-...	USA	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	54.50
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2780	Household - hardware, tools	Wire	AEROTAXY®	improved product	Solvoltaics	https://solvoltaics.com/solar-cells-panels-nan...	Sweden	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2781	Electronics	Wire	Nanosilver	improved product	INNOVA DYNAMICS	http://www.innovadynamics.com/products1.html	USA	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2782	Electronics	Wireless headset	Unknown	Unknown	Logitech	http://www.logitech.com/en-roeu/product/wirele...	USA	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2783	Electronics	Wireless headset	Unknown	Unknown	Logitech	http://www.logitech.com/en-roeu/product/wirele...	USA	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2784	Coatings, Paints, and Sealants (except textile)	Wood	Unknown	water repellent	New Guard Coatings Ltd	http://www.newguardcoatings.com/uk/product/nan...	UK	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

CPSC RDF

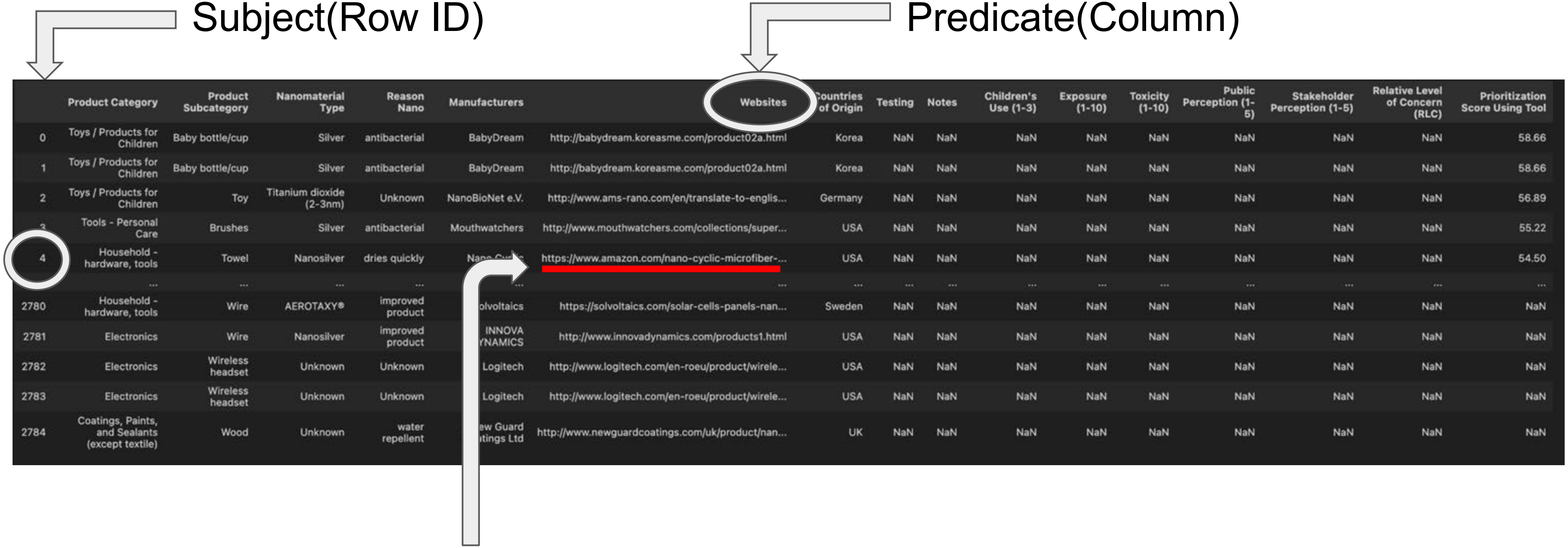


Ontology Reference

Abbreviation	Base URI	Full Name & Description
`BAO`	`http://www.bioassayontology.org/bao#`	**BioAssay Ontology** - A semantic description of bioassays and screening results
`BERO`	`http://purl.bioontology.org/ontology/BERO/`	**Biomedical and Environmental Research Ontology** - Describes concepts in biomedical and environmental research
`CHEAR`	`http://purl.obolibrary.org/obo/CHEAR_`	**Children's Health Exposure Analysis Resource Ontology** - Focused on children's environmental health research and exposure analysis
`CHV`	`http://purl.bioontology.org/ontology/CHV/`	**Consumer Health Vocabulary** - Bridges lay and professional medical terminology
`DIDEO`	`http://purl.obolibrary.org/obo/DIDEO_`	**Drug-drug Interaction and Evidence Ontology** - Represents drug-drug interactions and their evidence
`ENM`	`http://purl.enanomapper.org/onto/ENM_`	**eNanoMapper Ontology** - Describes concepts in nanotechnology safety assessment and characterization
`MaterialsMine`	`http://materialsmine.org/ns/`	**Materials Mine Ontology** - Represents materials science and engineering knowledge
`MELO`	`http://purl.bioontology.org/ontology/MELO/`	**MELanoma Ontology** - Specialized vocabulary for melanoma research and treatment
`NPO`	`http://purl.bioontology.org/ontology/npo#NPO_`	**NanoParticle Ontology** - Represents properties of nanoparticles, their experiments and characterizations
`PHARE`	`http://purl.bioontology.org/ontology/PHARE/`	**PHArmacogenomic RElationships Ontology** - Describes relationships between drugs, genes, and clinical outcomes
`SCTO`	`http://purl.bioontology.org/ontology/SCTO/`	**Saudi Clinical Trial Ontology** - Standardizes clinical trial terminology and processes
`THINF`	`http://purl.bioontology.org/ontology/THINF/`	**Thesaurus Inference** - Used for logical inference and relationships between concepts

- Subject -> Predicate -> Object (SPO) triple
- We often write it as Subject has Predicate of value Object

CPSC RDF Creation



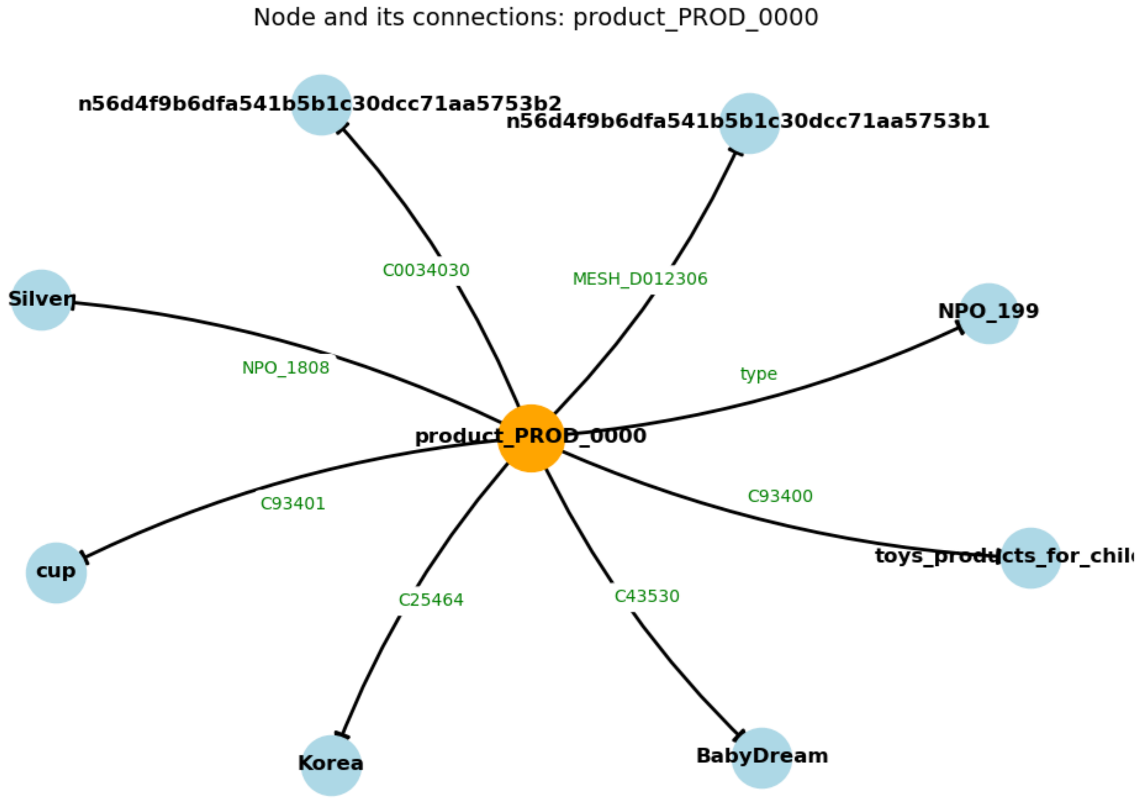
Subject(Row ID)

Predicate(Column)

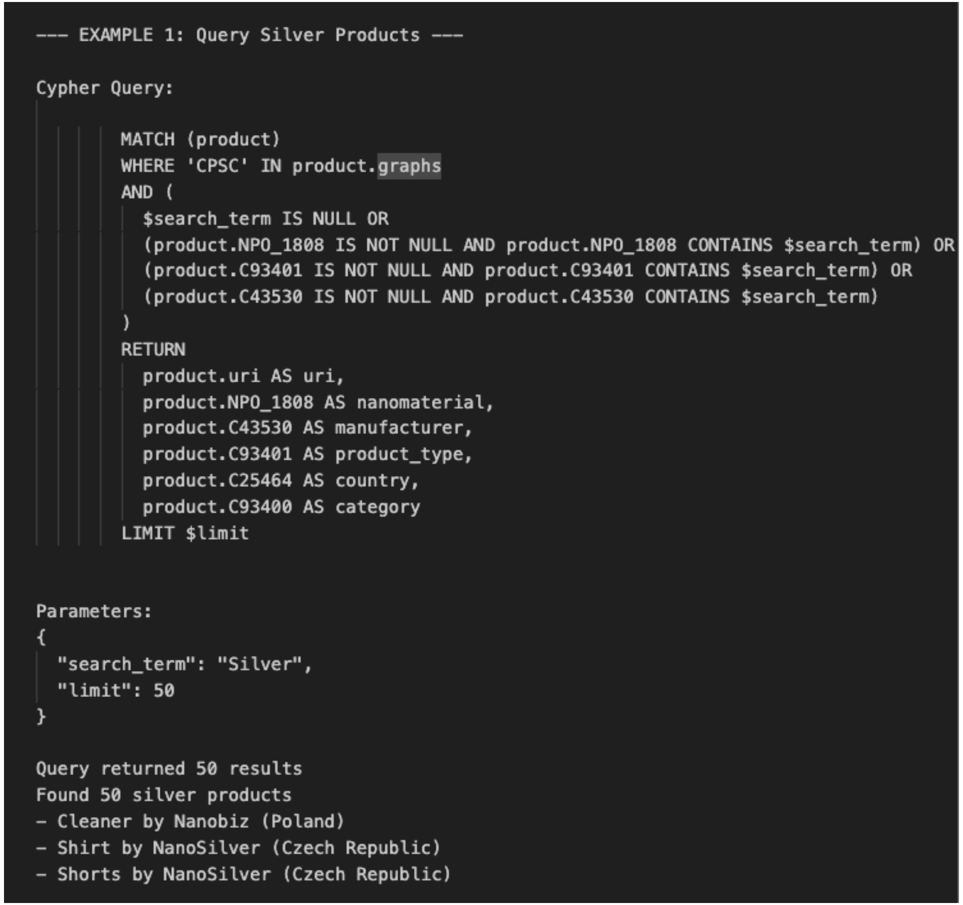
Object(Actual Value)

	Product Category	Product Subcategory	Nanomaterial Type	Reason Nano	Manufacturers	Websites	Countries of Origin	Testing	Notes	Children's Use (1-3)	Exposure (1-10)	Toxicity (1-10)	Public Perception (1-5)	Stakeholder Perception (1-5)	Relative Level of Concern (RLC)	Prioritization Score Using Tool
0	Toys / Products for Children	Baby bottle/cup	Silver	antibacterial	BabyDream	http://babydream.koreasme.com/product02a.html	Korea	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	58.66
1	Toys / Products for Children	Baby bottle/cup	Silver	antibacterial	BabyDream	http://babydream.koreasme.com/product02a.html	Korea	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	58.66
2	Toys / Products for Children	Toy	Titanium dioxide (2-3nm)	Unknown	NanoBioNet e.V.	http://www.ams-rano.com/en/translate-to-englis...	Germany	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	56.89
3	Tools - Personal Care	Brushes	Silver	antibacterial	Mouthwatchers	http://www.mouthwatchers.com/collections/super...	USA	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	55.22
4	Household - hardware, tools	Towel	Nanosilver	dries quickly	Nano [illegible]	https://www.amazon.com/nano-cyclic-microfiber-...	USA	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	54.50
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2780	Household - hardware, tools	Wire	AEROTAXY®	improved product	olvoltaics	https://solvoltaics.com/solar-cells-panels-nan...	Sweden	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2781	Electronics	Wire	Nanosilver	improved product	INNOVA YNAMICS	http://www.innovadynamics.com/products1.html	USA	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2782	Electronics	Wireless headset	Unknown	Unknown	Logitech	http://www.logitech.com/en-roeu/product/wirele...	USA	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2783	Electronics	Wireless headset	Unknown	Unknown	Logitech	http://www.logitech.com/en-roeu/product/wirele...	USA	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2784	Coatings, Paints, and Sealants (except textile)	Wood	Unknown	water repellent	ew Guard atings Ltd	http://www.newguardcoatings.com/uk/product/nan...	UK	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

CPSC RDF Creation



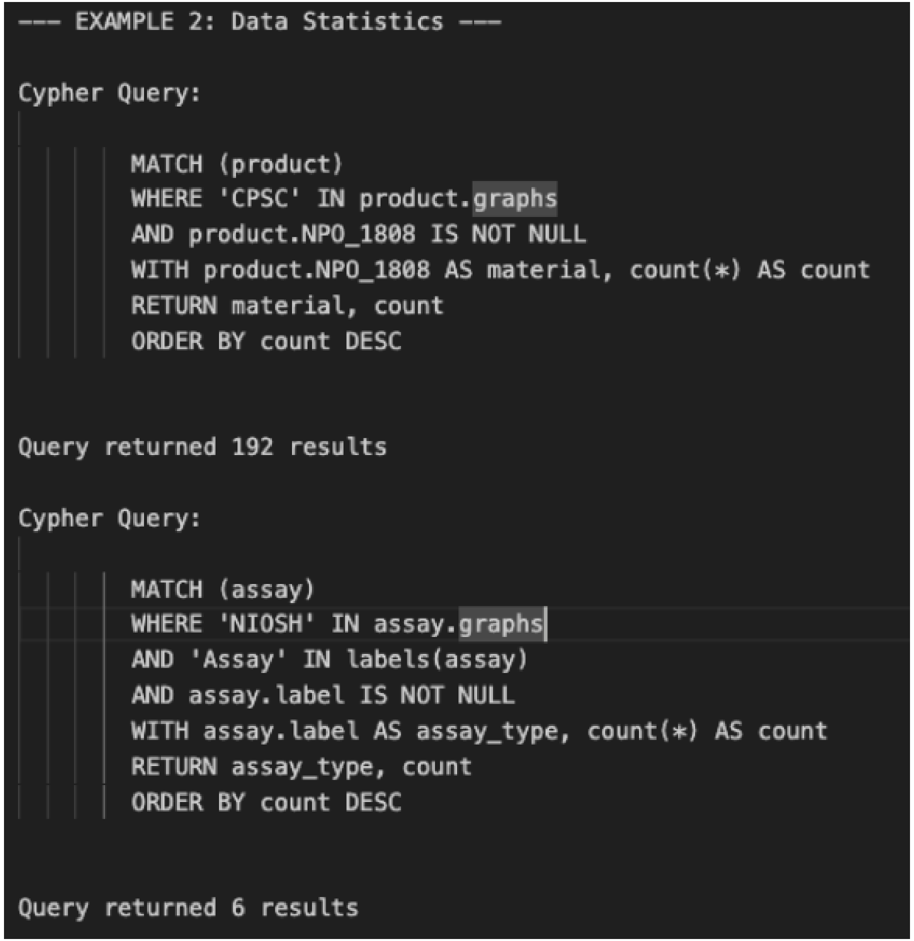
Cypher Queries across CPSC Graph



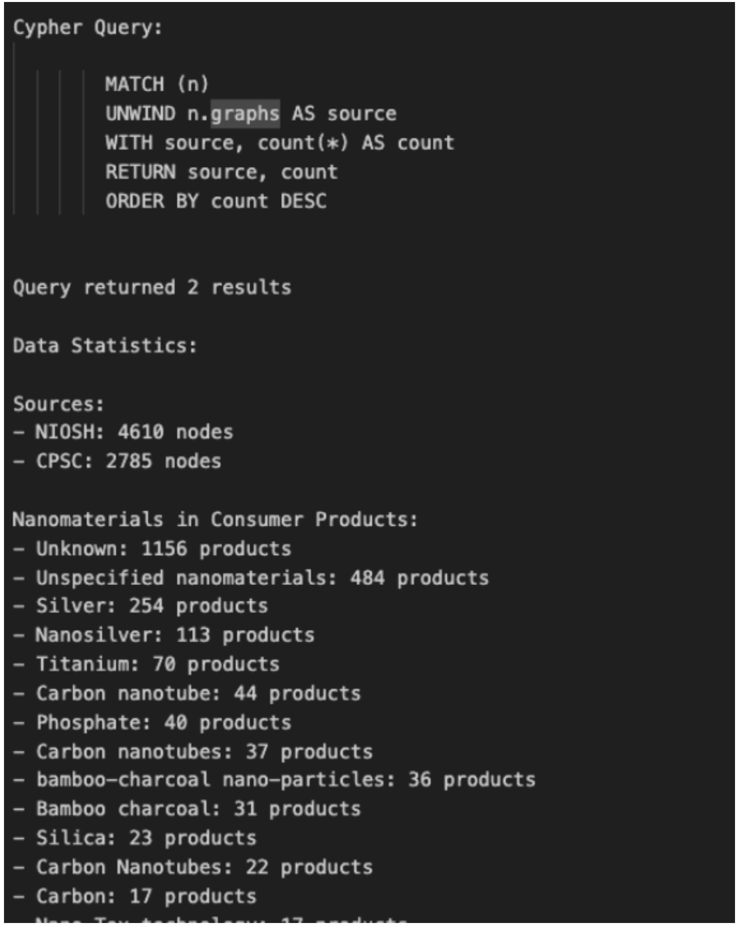
```
--- EXAMPLE 1: Query Silver Products ---

Cypher Query:

        MATCH (product)
        WHERE 'CPSC' IN product.graphs
        AND (
          $search_term IS NULL OR
          (product.NPO_1808 IS NOT NULL AND product.NPO_1808 CONTAINS $search_term) OR
          (product.C93401 IS NOT NULL AND product.C93401 CONTAINS $search_term) OR
          (product.C43530 IS NOT NULL AND product.C43530 CONTAINS $search_term)
        )
        RETURN
          product.uri AS uri,
          product.NPO_1808 AS nanomaterial,
          product.C43530 AS manufacturer,
          product.C93401 AS product_type,
          product.C25464 AS country,
          product.C93400 AS category
        LIMIT $limit


Parameters:
{
  "search_term": "Silver",
  "limit": 50
}

Query returned 50 results
Found 50 silver products
- Cleaner by Nanobiz (Poland)
- Shirt by NanoSilver (Czech Republic)
- Shorts by NanoSilver (Czech Republic)
```

- Why Cypher instead of SPARQL
    - Neo4j supports RDF graphs, easier to set up than SPARQL endpoint
    - Cypher is more widespread than SPARQL and there is more up-to-date resources/documentation

Queries Across CPSC and NIOSH Graphs



```
--- EXAMPLE 2: Data Statistics ---

Cypher Query:

        MATCH (product)
        WHERE 'CPSC' IN product.graphs
        AND product.NPO_1808 IS NOT NULL
        WITH product.NPO_1808 AS material, count(*) AS count
        RETURN material, count
        ORDER BY count DESC


Query returned 192 results

Cypher Query:

        MATCH (assay)
        WHERE 'NIOSH' IN assay.graphs
        AND 'Assay' IN labels(assay)
        AND assay.label IS NOT NULL
        WITH assay.label AS assay_type, count(*) AS count
        RETURN assay_type, count
        ORDER BY count DESC


Query returned 6 results
```



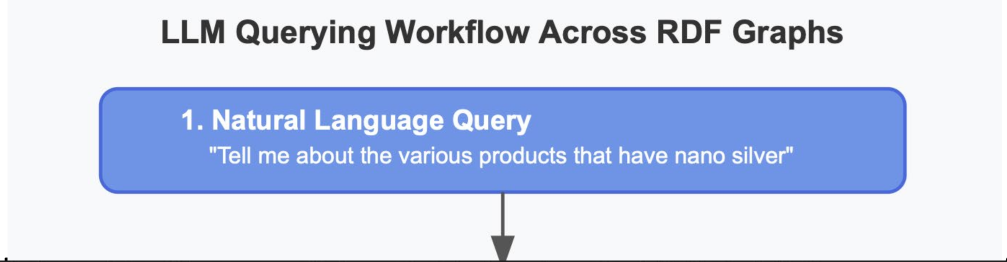
```
Cypher Query:

        MATCH (n)
        UNWIND n.graphs AS source
        WITH source, count(*) AS count
        RETURN source, count
        ORDER BY count DESC


Query returned 2 results

Data Statistics:

Sources:
- NIOSH: 4610 nodes
- CPSC: 2785 nodes

Nanomaterials in Consumer Products:
- Unknown: 1156 products
- Unspecified nanomaterials: 484 products
- Silver: 254 products
- Nanosilver: 113 products
- Titanium: 70 products
- Carbon nanotube: 44 products
- Phosphate: 40 products
- Carbon nanotubes: 37 products
- bamboo-charcoal nano-particles: 36 products
- Bamboo charcoal: 31 products
- Silica: 23 products
- Carbon Nanotubes: 22 products
- Carbon: 17 products
```

But what is the issue?

Learning how to write queries takes time!

Can we use AI to make query writing/data access easier?

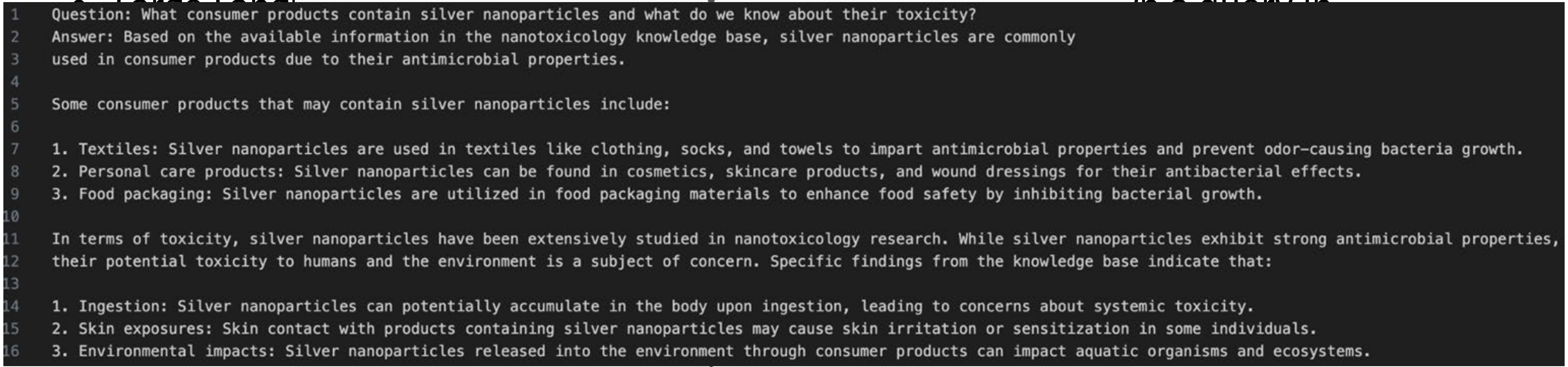
YES!



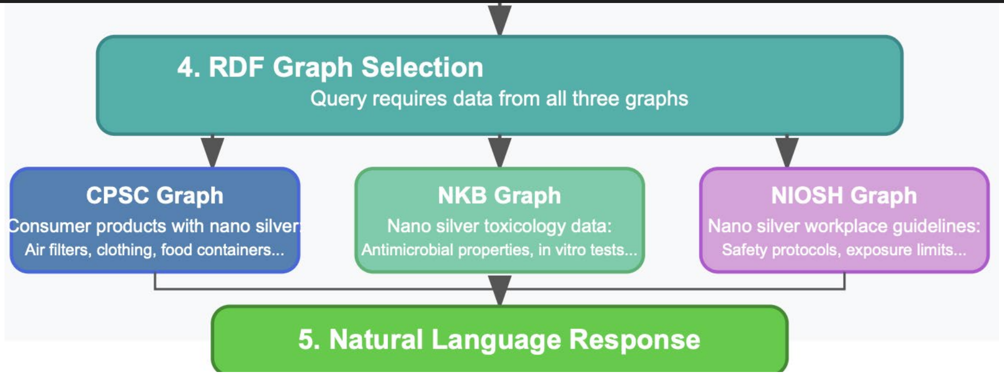
LLM Querying Workflow Across RDF Graphs

1. Natural Language Query
"Tell me about the various products that have nano silver"

- Large Lang
n a query in



```
Question: What consumer products contain silver nanoparticles and what do we know about their toxicity?
Answer: Based on the available information in the nanotoxicology knowledge base, silver nanoparticles are commonly
used in consumer products due to their antimicrobial properties.

Some consumer products that may contain silver nanoparticles include:

1. Textiles: Silver nanoparticles are used in textiles like clothing, socks, and towels to impart antimicrobial properties and prevent odor-causing bacteria growth.
2. Personal care products: Silver nanoparticles can be found in cosmetics, skincare products, and wound dressings for their antibacterial effects.
3. Food packaging: Silver nanoparticles are utilized in food packaging materials to enhance food safety by inhibiting bacterial growth.

In terms of toxicity, silver nanoparticles have been extensively studied in nanotoxicology research. While silver nanoparticles exhibit strong antimicrobial properties,
their potential toxicity to humans and the environment is a subject of concern. Specific findings from the knowledge base indicate that:

1. Ingestion: Silver nanoparticles can potentially accumulate in the body upon ingestion, leading to concerns about systemic toxicity.
2. Skin exposures: Skin contact with products containing silver nanoparticles may cause skin irritation or sensitization in some individuals.
3. Environmental impacts: Silver nanoparticles released into the environment through consumer products can impact aquatic organisms and ecosystems.
```



4. RDF Graph Selection
Query requires data from all three graphs

CPSC Graph
Consumer products with nano silver:
Air filters, clothing, food containers...

NKB Graph
Nano silver toxicology data:
Antimicrobial properties, in vitro tests...

NIOSH Graph
Nano silver workplace guidelines:
Safety protocols, exposure limits...

5. Natural Language Response

So now we can access data easier!

But there is a catch!

We can only get extracted outputs depending on how semantically rich our graphs actually are!
(which leads us to the next problem)

How can we manage missingness/inconsistencies in data?

Link Prediction From Existing Graphs

- Non-Established nomenclature as well as missingness present in data often leads to an incomplete picture of nanotoxicology data

- As a result, scientists in collaboration with one another might have differing opinions on what the missing data values would be

- What if there was a way to learn the graph structure and predict missing links between entities to enrich our data?

What is Link Prediction?

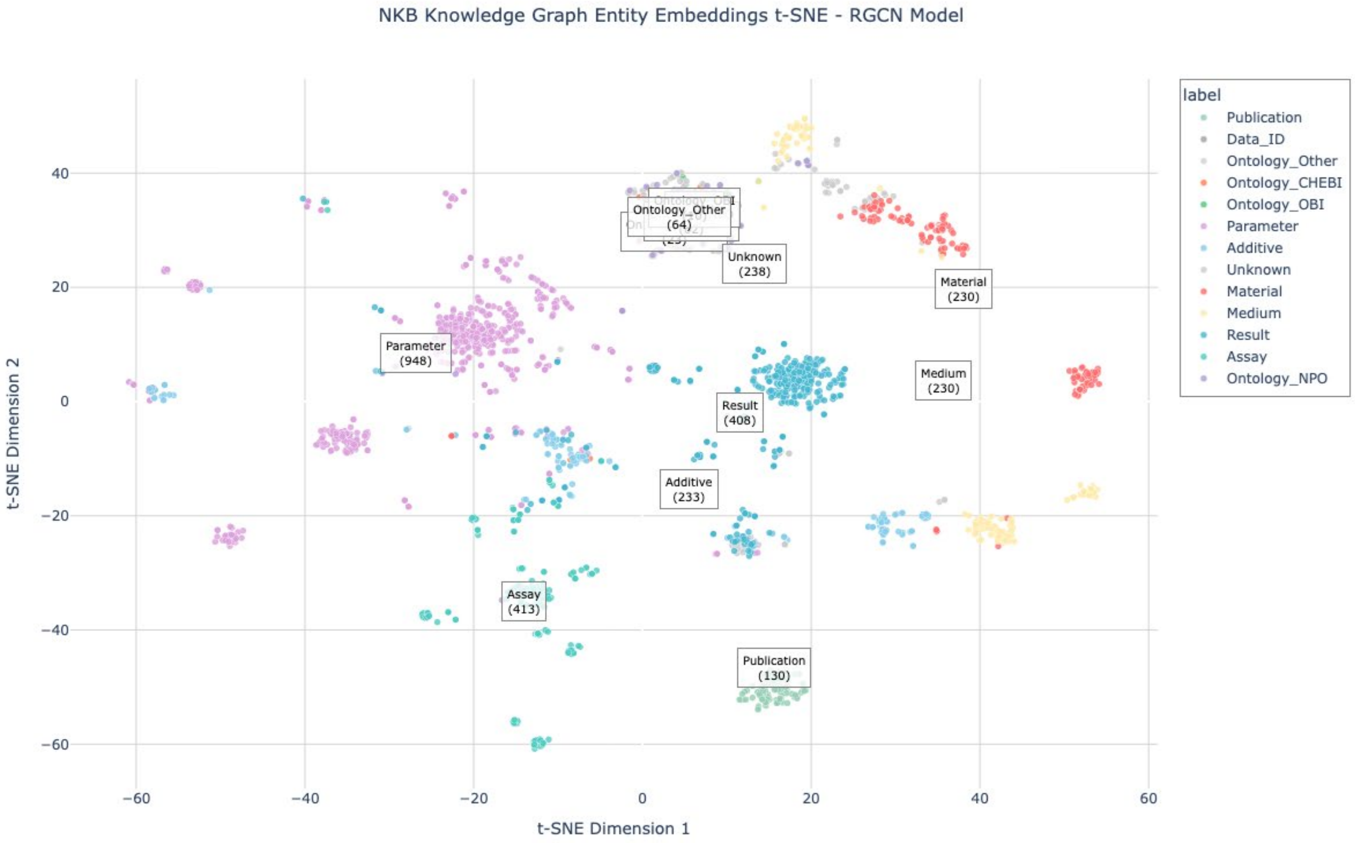
- How can we use existing graph structure and links to predict potential links between nodes?
- Procedure
  - Map Knowledge Graph into embedding space → Calculate probability of new links based on existing graph structure → Score Predictions →Evaluate Performance
  - There are 2 main types of embedding techniques: Shallow Embeddings & Deep Embeddings
    - Shallow Embeddings use a direct lookup for each entity/relation
    - Deep Embeddings use neural networks to generate embeddings based on node features and neighborhood structure

Link Prediction in the NKB RDF

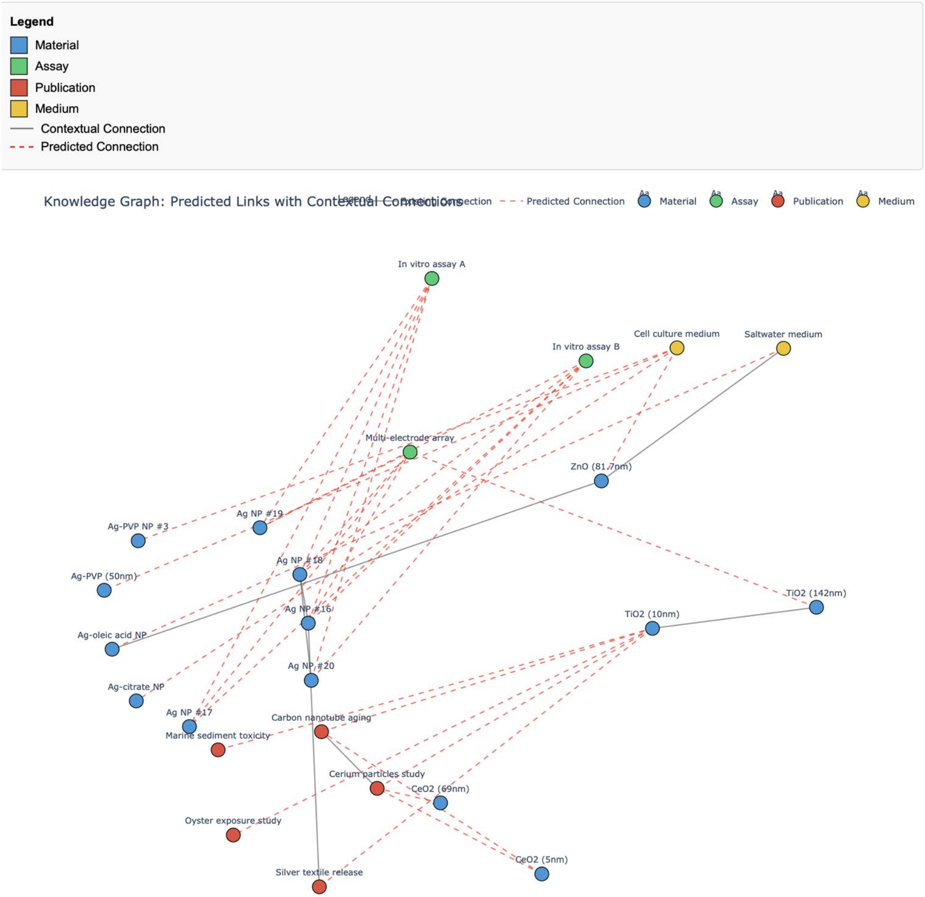
- Deep Embeddings are more computationally taxing but are able to generate novel predictions which can help us indicate potential new directions of research!
- Shallow Embeddings are good for efficient prediction but are only able to predict relations it has already seen, leading to overconfidence in potentially faulty predictions

Embeddings for the NKB Graph

- We can apply these two approaches to identify subgraph/neighborhood structure in our larger graphs rather than trying to visualize everything all at once
- To visualize these embeddings, we will use something called t-SNE(t-distributed stochastic neighbor embedding) for dimensionality reduction


NKB Knowledge Graph Entity Embeddings t-SNE - RGCN Model

Deep Embeddings & Link Prediction with NKB Graph



Takeaways

- Graph databases and storage present merits that we can leverage for data aggregation
- We can make querying across these databases easier so that scientists can actually focus on the science
- We can using the new graph structure that we have created for analysis and prediction tasks which would not have been possible with just a relational database

Acknowledgements

Dr. Holly Mortensen, US EPA/ORD

Dr. Rhema Bjorkland, NNI

Dr. Ryan Welch, NNI

Dr. Quinn Spadola, NSF

Dr. Branden Brough, Berkeley Labs

Thank you all for listening!