Medical and Transmission Vector Vocabulary Alignment with Schema.org

William Smith*, Alan Chappell, and Courtney Corley
Pacific Northwest National Laboratory

ABSTRACT

Available biomedical ontologies and knowledge bases currently lack formal and standards-based interconnections between disease, disease vector, and drug treatment vocabularies. The PNNL Medical Linked Dataset (PNNL-MLD) addresses this gap. This paper describes the PNNL-MLD, which provides a unified vocabulary and dataset of drug, disease, side effect, and vector transmission background information. Currently, the PNNL-MLD combines and curates data from the following research projects: DrugBank, DailyMed, Diseasome, DisGeNet, Wikipedia Infobox, Sider, and PharmGKB. The main outcomes of this effort are a dataset aligned to Schema.org, including a parsing framework, and extensible hooks ready for integration with selected medical ontologies. The PNNL-MLD enables researchers more quickly and easily to query distinct datasets. Future extensions to the PNNL-MLD may include Traditional Chinese Medicine, broader interlinks across genetic structures, a larger thesaurus of synonyms and hypernyms, explicit coding of diseases and drugs across research systems, and incorporating vector-borne transmission vocabularies.

1 INTRODUCTION

Medical vocabularies and ontologies have been developed over the last two decades and represent a large cross-section of Linked Open Datasets. Several research initiatives are now de facto authoritative data stores used by thousands of medical researchers daily including: DrugBank (Law, et al. 2014), PharmGKB (Stanford University 2014), Vectorbase (National Institute of Allergy and Infectious Diseases; National Institutes of Health; Department of Health and Human Services 2014), Uniprot (Consortium 2014), Allen Institute for Brain Science (AIBS) Brain Map (Allen Institute for Brain Science 2014), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, et al. 2014). However, with the collection of these advanced medical vocabularies and descriptive logic rules, a data classification divergence occurred.

Medical research groups rarely attempted to standardize vocabularies and ontologies with other research teams. This created data resources that are not natively interconnected with knowledge bases outside of a specific research objective. Furthermore, specific medical coding may exist on an entity level (OMIM, MeSH, eMedicine, etc), but there is no inherent guarantee across data sources that these codes are available or properly represented in a standard format. Entity matching between datasets is complicated by the fact most medical classes operate on a complex set of synonyms, hypernyms or taxonomical naming schemas that typically are not standardized across research projects and communities.

This effort addresses the tracking of a disease and treatment regimen across vector-borne transmission variables, including geography and species. The variety of issues described renders any available single source of research data unusable to address realistic research questions across the breadth of this domain space. Table 1 represents common diseases and transmission vectors for tracking vector-borne infections that were used as the starting point.

Disease	Transmission Vector
Eastern Equine Encephalitis Virus	Culiseta melanura / Cs. morsitans
Western Equine Encephalitis Virus	Culex / Culiseta
Highlands J Virus	Culiseta melanura
St. Louis Encephalitis Virus	Culex
West Nile Virus	Many
La Crosse Encephalitis	Ochlerotatus triseriatus synonym Aedes triseriatus
Chikungunya	A. albopictus and A. aegypti
Dengue Fever	Genus Aedes, principally A. aegypti

Table 1. Common diseases and associated transition vectors.

2 INITIAL VOCABULARIES AND ONTOLOGIES

One way of making use of the extensive previous work in disease descriptions by different research efforts and enabling associations across these vocabularies is assembling a knowledge base targeting the research area of interest. The more overlapping sets of information present in the resulting knowledge base the better chance a system has of making associations across vocabularies simply because of the availability of information on which to make the associations. For tracking vector-borne infections, disease datasets

* To whom correspondence should be addressed: william.smith@pnnl.gov

Dataset	Schema Entities	Schema Predicates	Schema Objects	Unaligned Entities	Unaligned Predicates	Unaligned Objects
Diseasome	4,213	7	31,538	3,938	13	43,836
PharmGKB	3,442	2	43,030	0	3	10,326
DisGeNet	13,172	1	13,172	0	3	39,516
Wikipedia Infobox	2,273	2	5,747	0	3	5,179
DailyMed	5,019	3	11,729	9,294	25	151,243
DrugBank	4,772	10	155,410	19,686	89	29,230
Sider	2,661	10	51,244	9	89	32,370

Table 2: Entity, predicate, and object counts after Schema.org alignment.

are a primary focus. Therefore, the team initially collected authoritative resources with a large amount of disease entities and extensive properties attached to each entity. The chosen datasets and entity count estimates include: Diseasome (Goh, et al. 2007), PharmGKB, DisGeNet (DisGeNet 2014), and Wikipedia Infobox (Wikimedia Foundation 2014). Table 2 depicts the data sets incorporated and the scale of the associated relevant vocabularies. These datasets provided different levels of expression across diseases, an example being PharmGKB having a small number of diseases with many properties versus DisGeNet having several times more entities expressed with a single name property and medical code.

Drug datasets, while initially not appearing to be part of the use case of tracking vector-borne infections, are useful as a direct path for aligning diseases across naming conventions. The selected drug datasets and estimated entity counts include: DailyMed (United States National Library of Medicine 2014) and DrugBank. In practice, drug datasets contain an extensive listing of medical codes, collected from prior research, across databases often missing from disease datasets. While these codes can be imprecise, they provide a starting point for entity interlinks and additional data enrichment through NLP and Linked Data techniques. When we focus on the disease medical codes affected by a specific treatment, the medical codes in the drug datasets enable us to programmatically create *owl:sameAs* relations across diseases in the disease data sets that are missing explicit matching medical codes or proper names. As a result, when drugs listing extensive medical codes are used as a reference point, diseases often can be more fully described, as missing medical codes are combined across datasets for more complete Linked Open Data.

Side effects were also included in the initial PNNL-MLD. This additional information enables detecting symptoms and matching the symptom to a disease or drug combination. The Sider (Kuhn, et al. 2010) dataset was selected as the lone source due to limited availability, but Sider contained dozens of different connections per entity across drugs further helping to align the combined dataset.

3 TARGET VOCABULARY: SCHEMA.ORG

In order to facilitate easier query description through a consistent vocabulary, the project chose one primary vocabulary to encompass the collected data. Selection of this vocabulary is driven by two primary considerations: 1) adequate expressiveness for the queries, and 2) not overly prescriptive such that it creates conflicts with the individual dataset semantics. The selection of this primary vocabulary is important, as it is an opportunity to promote wider use of the assembled dataset through adoption of an impactful or widely used vocabulary.

Schema.org (Google Inc; Microsoft Inc; Yahoo Inc 2014) was released in June 2011, and has become the search industry preferred standard for publishing search engine readable data. After the release of schema.org a RDFS (W3C RDF Working Group 2004) mapping was created and hosted on http://schema.rdfs.org, and this mapping is now a standard for Linked Data research utilizing Schema.org. Finally, at the end of June 2011, Schema.org released an official OWL (W3C OWL Working Group 2012) version of the Schema.org ontology bridging the gap between vocabulary and description logic.

Schema.org provides a base ontology class for medical entities available as a subclass of *Thing* entitled *MedicalEntity*. The subclasses of the *MedicalEntity* class were selected to represent the disease, drug, and side effect entities available within the PNNL-MLD. Table 3 lists the selected sub-classes.

Schema.org Class	Entity
MedicalCondition	Disease
MedicalCause	Disease Cause
MedicalSignOrSymptom	Disease Symptom
MedicalTherapy, Drug	Drug
MedicalCode	Entity Code
MedicalEntity	Side Effect

Table 3. Schema.org classes selected to represent use case entities.

4 VOCABULARY ALIGNMENT

Simply adding a primary vocabulary to the datasets is not adequate to simplify querying. The source datasets must be aligned with the primary vocabulary so that queries will return results that span and integrate all the available information. The central goal in this alignment is to provide a mapping of the source vocabularies to the new primary vocabulary that preserves the semantics of the source but bridges the divergence between the different knowledge representations.

4.1 Base dataset alignment

The project selected the URI: http://beowulf.pnnl.gov/2014/ to serve as the RDF (W3C RDF Working Group 2004) prefix base for all aligned data. We used this new base URI to simplify software development later in the alignment process. Furthermore, all properties were immediately aligned by import dataset, prefix associations demonstrated by the following:

beo-<dataset-name>:propertyName.

By first associating property and class values with an original prefix denoting dataset we could now track properties that were not explicitly aligned to Schema.org. The *rdfs:label* and *owl:sameAs* properties were left unmodified throughout the entire process, and **schema:alternateName** is used to track synonyms of *rdfs:label*.

4.2 Disease dataset alignment

Four large datasets of varying entity counts and properties were the first targets after the base import of the PNNL-MLD. The first substitution took place by converting all unique entity IRIs to a common format:

beo-disease:<disease-id>

We then added the Schema.org declaration of class:

a schema:MedicalCondition

Primary preventions were added to diseases as drug IRIs were detected:

schema:primaryPrevention beo-drug:<original-drug-id>, …

Finally, we use Table 4 to ensure we can match back to online medical resources and unify datasets:

Schema.org Class	Entity
MedicalCode	IRI
MedicalPage	URI
code	Unknown Code Type

Table 4. Alignment of Schema.org classes to medical resources.

4.3 Drug dataset alignment

Two datasets comprised drug metadata and provided interlinks to side effect metadata. These entities were referenced in both disease and side effect datasets as potential treatments (disease) and causes of (side effect). The first substitution took place by converting all unique entity IRIs to a common format:

beo-drug:<disease-id>

We then added the Schema.org declaration of class:

a schema:Drug

Drug, a subclass of *MedicalTherapy*, was selected due to the semantics of the original data. Drugs have the same medical coding standards as Table 4, but the attributes linking the drugs are more abstract including two descriptions of the drug:

schema:potentialAction
schma:description

To link the drug entity to a disease we replace:

beo-drugbank:possibleDiseaseTarget

with:

schema:possibleTreatment

Finally, drugs can interact with each other creating adverse reactions. The DrugBank dataset provides the interconnections for this possibility. We aligned these reactions by creating the entity type:

a beo-drugbank:drug_interactions

And ensuring the new entity has at least two of the following relations:

schema:interactingDrug beo-drug:id

4.4 Side Effect dataset alignment

The single Sider dataset provides the final links to drug entities with each side effect's unique IRI converted to:

beo-interaction:<effect-id>

Then adding the Schema.org declaration of class:

a schema:medicalEntity

Completing the ontology requires one last step linking drugs to side effects with the drug entity property:

schema:seriousAdverseOutcome beo-interaction:<id>

5 QUERY A DISEASE

Using Dengue Fever as an example disease we can now use **schema:MedicalCondition** to query across all of the disease datasets. The SPARQL (W3C SPARQL Working Group 2013) query below locates the available information in the combined dataset about any medical condition with "dengue" in its name and collects the comments that describe the source of that information.

```
@prefix schema: <http://schema.org/>
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
```



```
SELECT ?label ?comment
WHERE {
    ?item a schema:MedicalCondition .
    ?item rdfs:label ?label .
    FILTER (regex(?label, 'dengue', 'i')) .
OPTIONAL { ?item rdfs:comment ?comment }}
```

Running this query on the PNNL-MLD returns Table 5.

?label	?comment
"Dengue shock syndrome"	"Imported from DISGENET"
"Dengue"	"Imported from PharmGKB "
"Dengue Hemorrhagic Fever"	"Imported from PharmGKB"
"Dengue"	"Imported from DISGENET"
"Dengue Hemorrhagic Fever"	"Imported from DISGENET "
"Dengue fever, protection against"	<diseasome>
"Dengue_fever,_protection_against"	<diseasome>

Table 5. Result of SPARQL query on Dengue Fever.

The results in Table 5 expose a current limitation of the system due to regex matching of the label property. Because the query can now reach across several different datasets with conflicting naming schemes an additional normalization process is needed during the data import to normalize labels for all of the entities linked with *owl:sameAs.*

The results in Table 5 show that one simple query now identifies data from three different sources. This begins to show the value of the combined dataset. However, to explore the full impact of the alignment a more complex query is needed that requires the integration of information from multiple sources. Expanding on our previous query we can search across all originally returned "dengue" conditions and append the drug links and treatments added with Schema.org .

```
@prefix schema: <http://schema.org/>
@prefix rdfs: <http://www.w3.org/2000/01/rdf-
schema#>

SELECT DISTINCT ?drugLabel ?diseaseTarget
WHERE {
GRAPH ?G{
    ?item a schema:MedicalCondition .
    ?item rdfs:label ?label .
    FILTER (regex(?label, 'dengue', 'i')) .}
GRAPH ?G1{
    ?item schema:primaryPrevention ?drug .
    ?drug rdfs:label ?drugLabel .}
GRAPH ?G2{
    ?drug schema:possibleTreatment ?target .
    ?target rdfs:label ?diseaseTarget }}
```

This query returns Table 6.

?drugLabel	?diseaseTarget
"Alpha-D-Mannose"	"Dengue_fever,_protection_against"
"Fucose"	"Dengue_fever,_protection_against"

Table 6. Result of SPARQL query for drugs treating Dengue Fever showing value of aligned vocabulary.

Another limitation of the current PNNL-MLD is exposed reviewing the results of Table 6. When creating interlinks across diseases, only the Diseasome entities were referenced in the corresponding drug datasets as possible targets for treatment. To correct this oversight we also need to include *owl:sameAs* associations within our queries, or select a logical reasoner capable of associating and returning all related entities upon a single link between a disease and drug.

Most importantly, Table 6 depicts the value of the combined and aligned PNNL-MLD dataset. Queries like the one given here that require information linking diseases to treatments or symptoms or side effects are now greatly simplified and can focus on a single vocabulary. Schema.org provided classes and properties appropriate for drafting queries that can provide views of the data not visible using only a single source of data.

No technical limitation exists that would restrict a user from loading all of the datasets into separate graphs of an available triplestore and querying the different vocabularies across graphs. However, when we align these datasets into the PNNL-MLD we achieve four major benefits:

1. Queries are now simplified. Early drafts for querying across all of the graphs required queries that were dozens of lines in length, and portions of the queries varied drastically in format and language.
2. A standardized vocabulary, that is industry recognized, is now in place for application development.
3. All of the graphs, when aligned into the PNNL-MLD, are now equally extensible. Adding new vocabularies and ontologies to the original data would require special updates to each dataset, and require updates to each specific portion of a query using that dataset.
4. As shown in Table 2, when a dataset is converted using RDF, and not generated from a different file type (unaligned entities = 0), the Schema.org entities now have a much higher ratio of Schema.org predicates to object triple mappings. By flattening the ontology a simplified query now has access to a much greater range of values and entities.

Additionally, because all modification and additions made while aligning are programmatically defined rather than human expert mediated, new version of the PNNL-MLD can be easily created as source datasets produce new versions.

6 CURRENT LIMITATIONS

The complete PNNL-MLD is now capable of being queried through SPARQL using only Schema.org associations. However, there are still shortcomings in searching for drugs and diseases by name, including the corresponding regex filters. To resolve this conflict a primary label for a group of entities related by *owl:sameAs* should be selected upon entity interlinking with the previous labels turned into **schema:alternateName** properties. Queries should then be composed to either search for a primary name and/or alternate synonym. To remove duplicates imported from different datasets a reasoner capable of merging *owl:sameAs* relations should be used when querying the complete PNNL-MLD.

Medical coding was not at first considered a feature of the application and early versions of the PNNL-MLD did not prioritize accurately creating the properties in Table 3. As it became more apparent diseases and drugs were not consistently labeled across datasets, and outside database entities generally were consistent across datasets, more focus was added to ensure medical codes were applied to drug and disease entities. However, this process was never finalized through Linked Data authentication to ensure the medical codes supplied were accurate for the attached entity.

6.1 Future work

To address current limitations we need to focus on best practices utilizing linked data (Heath and Bizer 2011), and expanding vector transmission geo-properties.

1. Authenticate medical coding. Confirm the entity is correctly aligned to outside sources.
2. Add Gazetteer to provide formal geographic naming entities while also mapping a list of local colloquialisms for geographic regions.
3. Add Vectorbase. (National Institute of Allergy and Infectious Diseases; National Institutes of Health; Department of Health and Human Services 2014)

7 CONCLUSIONS

The broader implications of aligning datasets under a common vocabulary, and making them available using Linked Open Data best practices, is to standardize and expand the original research objectives. When we augment the unique vocabulary and ontology mappings of individual research programs with the broader Schema.org vocabulary, we create data interlinks that enable conceptualization of new questions that bridge the earlier work without requiring replicating research with a broader focus. This combination of separate datasets with common data points aligned to nonexclusive properties and ontology rules simplifies queries, and creates a new superset built for application development and public discovery.

ACKNOWLEDGEMENTS

This work was funded by a contract with the Defense Threat Reduction Agency (DTRA), Joint Science and Technology Office for Chemical and Biological Defense under project number CB10082. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle under Contract DE-AC05-76RL01830.

REFERENCES

Allen Institute for Brain Science. *Allen Human Brain Atlas.* 2014. http://human.brain-map.org/ (accessed 2014).

Ashburner, Michael. *BioPortal.* 2014. http://bioportal.bioontology.org/ontologies/GAZ.

Consortium, The UniProt. "UniProt: a hub for protein information ." *Oxford Journals* 43, no. D1 (2014).

DisGeNet. 10 2014. http://www.disgenet.org/web/DisGeNET/v2.1/dbinfo.

Goh, Kwang-Il, Michael Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. "The Human Disease Network." *Proc Natl Acad Sci USA*, 4 2007.

Google Inc; Microsoft Inc; Yahoo Inc. 2014. http://schema.org/.

Heath, Tom, and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space.* 1. Berlin: Morgan & Claypool, 2011.

Kanehisa, M, S Goto, Y Sato, M Kawashima, M Furumichi, and M Tanabe. "Data, information, knowledge and principle: back to metabolism in KEGG." *Nucleic Acids Res*, Jan 2014.

Kuhn, M, M Campillos, I Letunic, LJ Jensen, and P Bork. "A side effect resource to capture phenotypic effects of drugs." *Epub* (NCBI), 1 2010.

Law, V, et al. "DrugBank 4.0: Shedding new light on drug metabolism." *PubMed*, no. 24203711 (2014).

National Institute of Allergy and Infectious Diseases; National Institutes of Health; Department of Health and Human Services. 2014. https://www.vectorbase.org.

Stanford University. 2014. https://www.pharmgkb.org/.

United States National Library of Medicine. 10 1, 2014. http://dailymed.nlm.nih.gov/.

W3C OWL Working Group. 2012. http://www.w3.org/TR/2012/REC-owl2-overview-20121211/.

W3C RDF Working Group. 2004. http://www.w3.org/TR/2004/REC-rdf-mt-20040210/.

W3C SPARQL Working Group. "SPARQL 1.1 Query Language." *W3C Recommender*. March 2013. http://www.w3.org/TR/sparql11-query/ (accessed October 2014).

Wikimedia Foundation. 2014. http://www.wikidata.org

