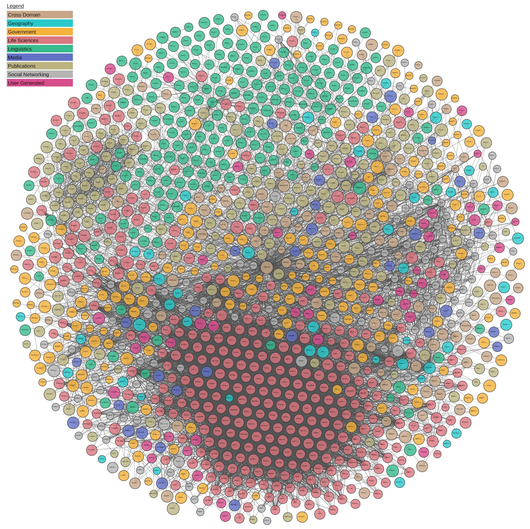
PRESENTER: **Juan M. Banda**
EMAIL: **jbanda@gsu.edu**
TWITTER: **@drjmbanda**

- The usage of controlled biomedical vocabularies is the cornerstone that enables seamless interoperability when using a common data model across multiple sites.
- The Observational Health Data Science and Informatics (OHDSI) initiative combines over 100 controlled vocabularies into its own.
- However, the OHDSI vocabulary is limited in the sense that it combines multiple terminologies and does not provide a direct way to link them outside of their own self-contained scope.
- This issue makes the tasks of enriching feature sets by using external resources extremely difficult.
- **In order to address these shortcomings, we have created a linked data version of the OHDSI vocabulary, connecting it with already established linked resources like bioportal, bio2rdf, etc. with the ultimate purpose of enabling interoperability of resources previously foreign to the OHDSI universe.**



OHDSI2RDF: Fully connecting the Observational Health Data Science and Informatics (OHDSI) initiative with the world of linked open data

During our time at the Biomedical Link Data Hackathon 5 in Kashiwa, Japan we developed the first attempt to create an RDF version of the OHDSI vocabulary with linkages to UMLS and BioPortal.

In order to link the OHDSI vocabulary with UMLS, leveraged Ananke, a resource built for the mapping of UMLS Concept Unique Identifiers (CUIs) into OHDSI concept_id's, which are the unique identifiers assigned to all concepts in the vocabulary. This allows us to use BioPortals URI's for the CUIs and make the necessary connections when using their SPARQL endpoints for federated queries.

The RDF conversion results in a total of 24 million triples and takes around 15 minutes. Our resource links a total of 861,732 OHDSI concept_id's from SNOMED, 286,256 concept_id's from RxNORM, 109,706 concept_id's from ICD10, and 22,029 concept_id's from ICD9, all linked directly to bioportal. We also include 1,321,986 mappings to UMLS via Ananke.

As authors of the Automated PHenotype Routine for Observational Definition, Identification, Training and Evaluation (APHRODITE) R package, our goals are as follows:

(1) Be able to expand and enrich our feature sets for phenotyping. With one of the main feature spaces of APHRODITE being clinical narratives, these are annotated using the OHDSI vocabulary. Having a linked version of it will allow us to expand any particular feature domain with other linked resources to SNOMEDCT, RxNORM, etc.

(2) Besides producing a machine learning model for the target phenotype, one of the outputs of APHRODITE is a list of relevant features that add interpretability to any model. This list of features covers the most important domains in the OHDSI CDM and vocabulary. We want to be able to produce this list as a linked resource that will allow researchers to enhance their understanding by being able to semantically link them to other resources like the Human Phenotype Ontology among others.

Take a picture to **download** the **full paper**

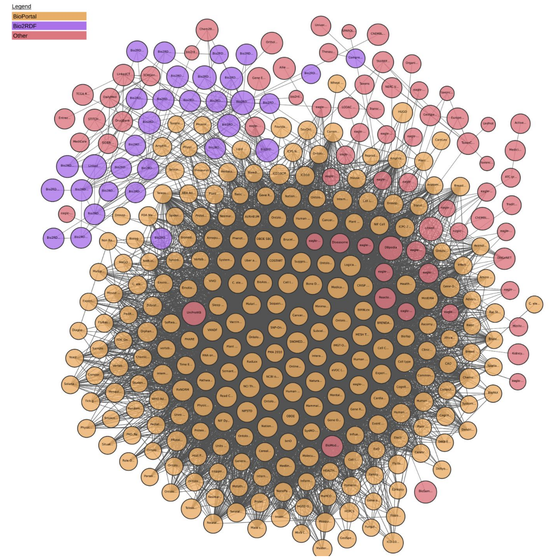
https://doi.org/10.5808/GI.2019.17.2.e13

Want to contribute?

We are still missing additional linkages and new mappings. Currently we provide 2 versions of OHDSI2RDF: 1) RDF generation code, 2) Fully formed tripples using Ananke v 1.0 and the OHDSI vocabulary v5.0 11-FEB-19.

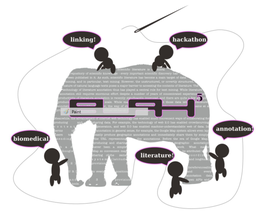
Find this project here:

https://github.com/thepanacealab/OHDSI2RDF



Juan M Banda, PhD
Georgia State University, Atlanta, GA, USA

Biomedical Linked Annotation Hackathon 5
12 - 15 February 2019, Kashiwa, Japan



PANACEA LAB

Georgia State University

OHDSI