Linked Data for the Life Sciences

Release 2

Michel Dumontier
Associate Professor, Carleton University

on behalf of the Bio2RDF team

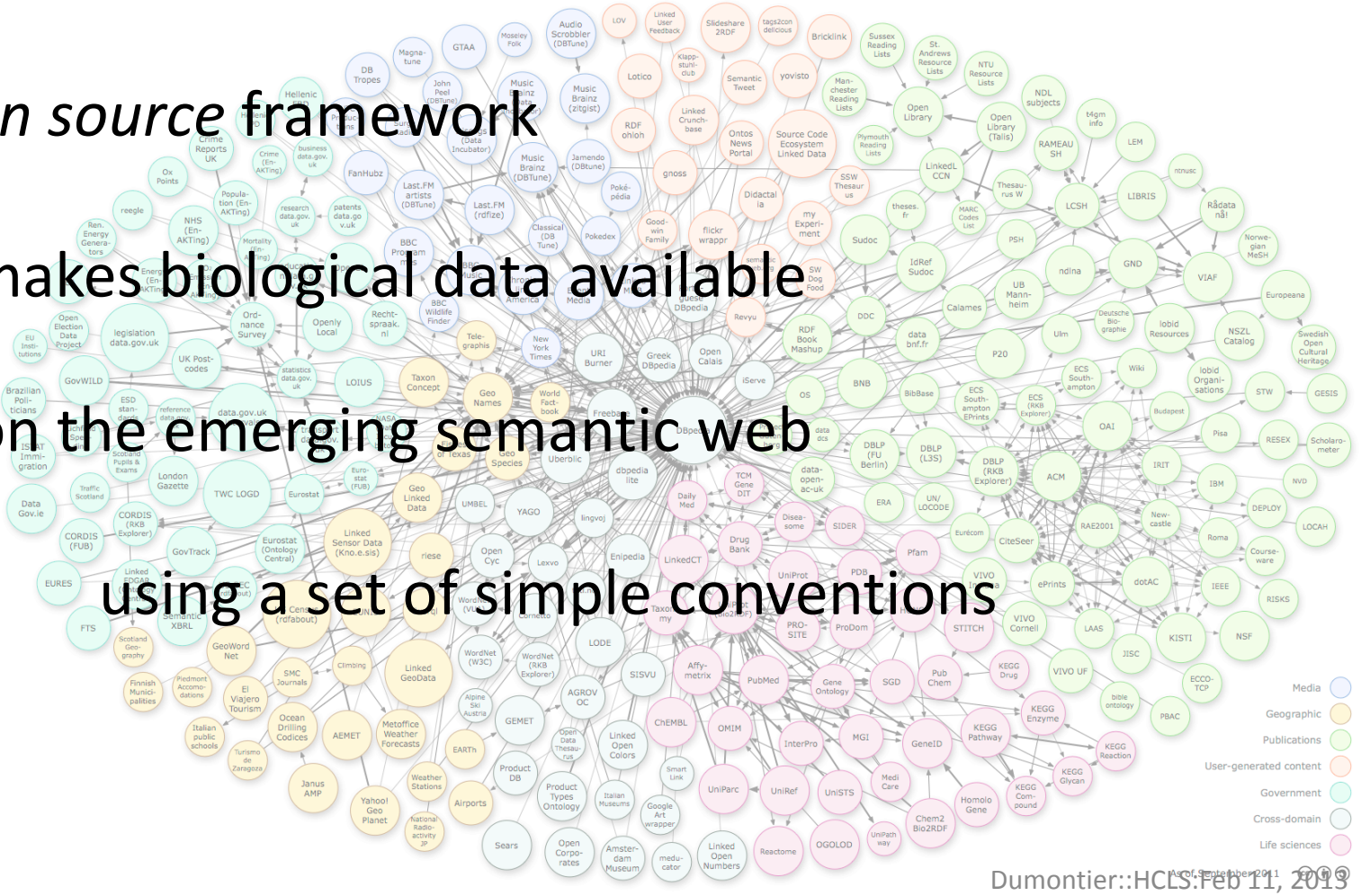
BIO2RDF

is an *open source* framework

that makes biological data available

on the emerging semantic web

using a set of simple conventions



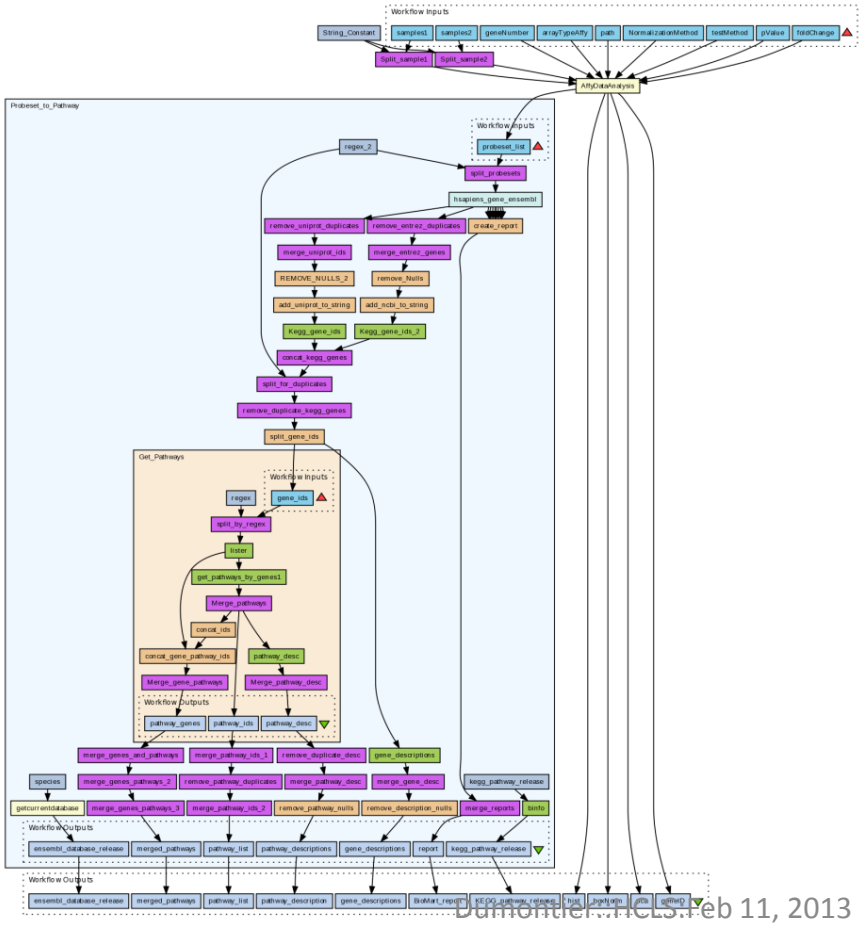
BIO2RDF

reduces the time and effort

involved in data integration

so that *you* can get to

the business
of doing science



Main features of Bio2RDF Release 2

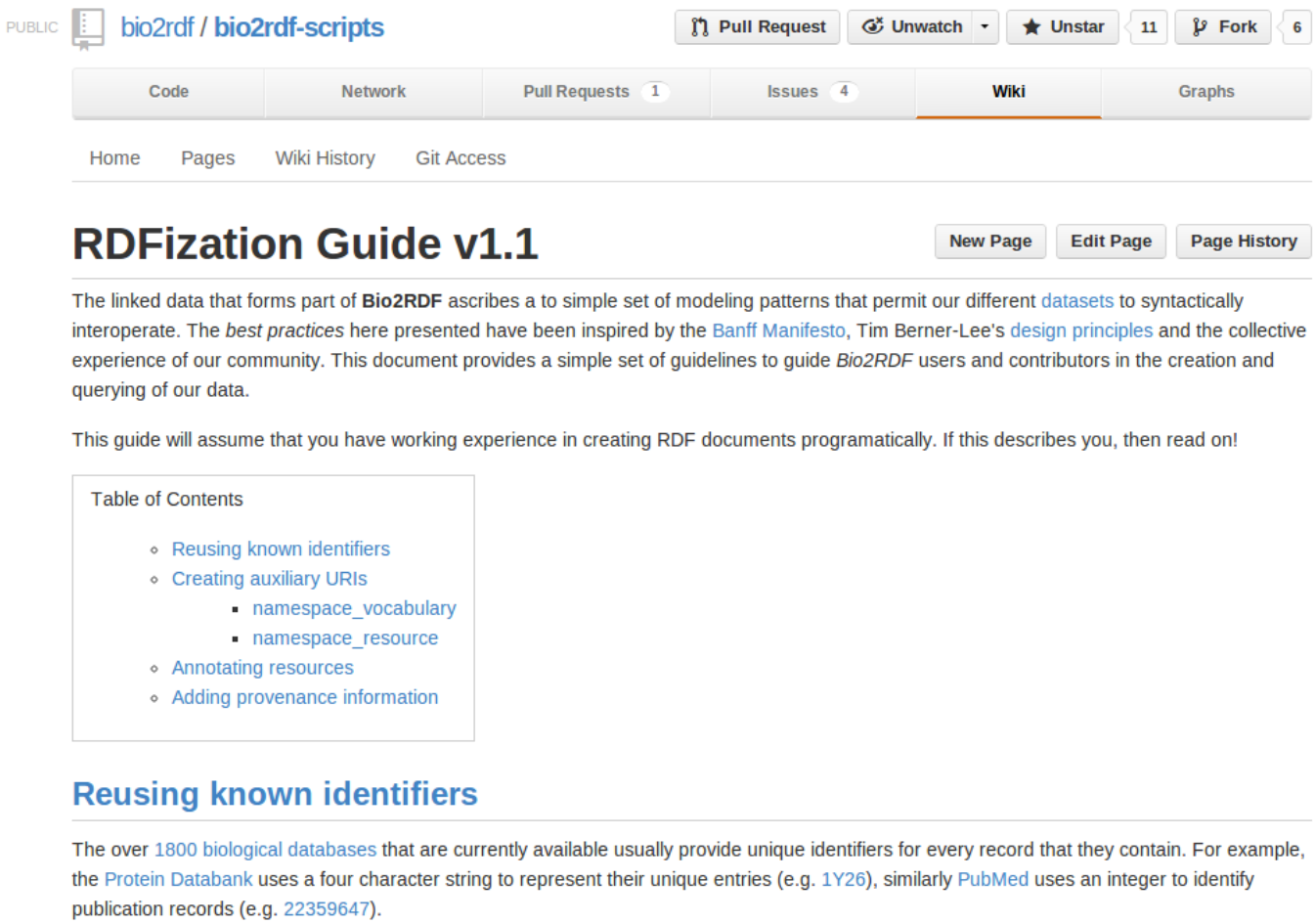
- Bio2RDF conversion scripts, mapping files and web application are **open source** **and** **freely available** at http://github.com/bio2rdf

- Bio2RDF enables (syntactic) data integration within and across datasets by using one language (**RDF**), having a **common URI pattern** and a **common resource registry**

- 19 Release 2 datasets have **provenance** and endpoints feature pre-computed **graph summaries** for fast lookup

- Bio2RDF web application enables **entity resolution**, **query federation** across an **expandable distributed network of SPARQL endpoints**

Bio2RDF RDFization guidelines are available at our wiki



PUBLIC bio2rdf / bio2rdf-scripts

Pull Request | Unwatch | Unstar 11 | Fork 6

Code | Network | Pull Requests 1 | Issues 4 | Wiki | Graphs

Home | Pages | Wiki History | Git Access

RDFization Guide v1.1

New Page | Edit Page | Page History

The linked data that forms part of **Bio2RDF** ascribes a to simple set of modeling patterns that permit our different datasets to syntactically interoperate. The *best practices* here presented have been inspired by the Banff Manifesto, Tim Berner-Lee's design principles and the collective experience of our community. This document provides a simple set of guidelines to guide *Bio2RDF* users and contributors in the creation and querying of our data.

This guide will assume that you have working experience in creating RDF documents programatically. If this describes you, then read on!

Table of Contents

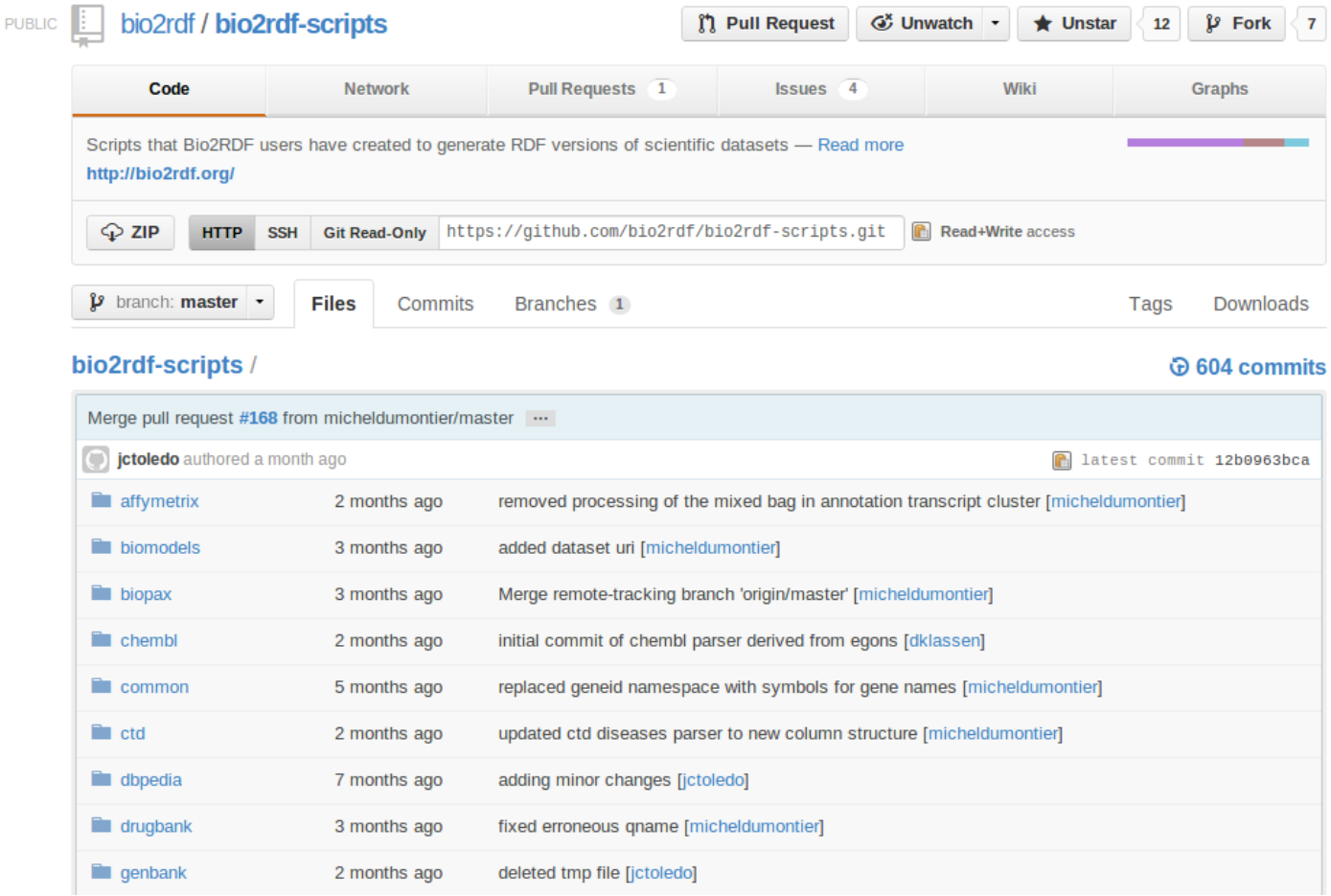
- Reusing known identifiers
- Creating auxiliary URIs
  - namespace_vocabulary
  - namespace_resource
- Annotating resources
- Adding provenance information

Reusing known identifiers

The over 1800 biological databases that are currently available usually provide unique identifiers for every record that they contain. For example, the Protein Databank uses a four character string to represent their unique entries (e.g. 1Y26), similarly PubMed uses an integer to identify publication records (e.g. 22359647).

https://github.com/bio2rdf/bio2rdf-scripts/wiki/RDFization-Guide-v1.1

Bio2RDF converters are open-source and available at GitHub



PUBLIC bio2rdf / bio2rdf-scripts

Pull Request | Unwatch | Unstar 12 | Fork 7

Code | Network | Pull Requests 1 | Issues 4 | Wiki | Graphs

Scripts that Bio2RDF users have created to generate RDF versions of scientific datasets — Read more
http://bio2rdf.org/

ZIP | HTTP | SSH | Git Read-Only | https://github.com/bio2rdf/bio2rdf-scripts.git | Read+Write access

branch: master | Files | Commits | Branches 1 | Tags | Downloads

bio2rdf-scripts / 604 commits

Merge pull request #168 from micheldumontier/master

jctoledo authored a month ago — latest commit 12b0963bca

affymetrix	2 months ago	removed processing of the mixed bag in annotation transcript cluster [micheldumontier]
biomodels	3 months ago	added dataset uri [micheldumontier]
biopax	3 months ago	Merge remote-tracking branch 'origin/master' [micheldumontier]
chembl	2 months ago	initial commit of chembl parser derived from egons [dklassen]
common	5 months ago	replaced geneid namespace with symbols for gene names [micheldumontier]
ctd	2 months ago	updated ctd diseases parser to new column structure [micheldumontier]
dbpedia	7 months ago	adding minor changes [jctoledo]
drugbank	3 months ago	fixed erroneous qname [micheldumontier]
genbank	2 months ago	deleted tmp file [jctoledo]

http://github.com/bio2rdf/bio2rdf-scripts

Bio2RDF data are identified using *simple* http URI patterns

When available, use the provider's identifier in the naming the resource

http://bio2rdf.org/namespace:identifier

e.g.: DrugBank's resource IRI for Leucovorin

http://bio2rdf.org/drugbank:DB00650

Linked Data: You can look it up

Leucovorin [drugbank:DB00650] at Bio2RDF

Search

Find intranamespace links Find global links Links Namespace Find links in namespace

http://bio2rdf.org/drugbank:DB00650

Leucovorin [drugbank:DB00650]

Subject	Predicate	Object
http://bio2rdf.org/drugbank:DB00650	http://bio2rdf.org/bio2rdf_resource:urlList	http://bio2rdf.org/html/drugbank:DB00650
	http://bio2rdf.org/drugbank_vocabulary:absorption	Following oral administration, leucovorin is rapidly absorbed. The apparent bioavailability of leucovorin was 97% for 25 mg, 75% for 50 mg, and 37% for 100 mg.
	http://bio2rdf.org/drugbank_vocabulary:affected-organism	Humans and other mammals
	http://bio2rdf.org/drugbank_vocabulary:biotransformation	Hepatic and intestinal mucosal, the main metabolite being the active 5-methyltetrahydrofolate. Leucovorin is readily converted to another reduced folate, 5,10-methylenetetrahydrofolate, which acts to stabilize the binding of fluorodeoxyridylic acid to thymidylate synthase and thereby enhances the inhibition of this enzyme.
	http://bio2rdf.org/drugbank_vocabulary:brand	Calcium citrovorum factor
	http://bio2rdf.org/drugbank_vocabulary:calculated-property	http://bio2rdf.org/drugbank_resource:calculated_property_DB00650_10
		http://bio2rdf.org/drugbank_resource:calculated_property_DB00650_11
		http://bio2rdf.org/drugbank_resource:calculated_property_DB00650_12
		http://bio2rdf.org/drugbank_resource:calculated_property_DB00650_13
		http://bio2rdf.org/drugbank_resource:calculated_property_DB00650_14

Valid Bio2RDF namespaces are listed in a dataset registry

- An initial registry of ~**600** datasets is accessible through an API provided by my PHP-LIB library (available on github). It includes
  - Dataset title, Preferred namespace prefix, Alternative namespace prefixes
- In the summer, we consolidated and curated nearly **2100** entries in a Google spreadsheet, which includes a mostly complete coverage of datasets/collections listed in *Bio2RDF, MIRIAM, BioPortal, UniProt, NCBI, NAR database issue*. New fields were added including:
  - Dataset description, organization, website, HTML template
  - Identifier syntax, license and rights
- Working with identifiers.org team (Nick Juty, Camille Laibe, Nicolas Le Novere) to have a single dataset registry that we can use for both Bio2RDF and identifiers.org
  - **enable automatic cross-links between Bio2RDF and identifiers.org**
  - **syntactic validation of identifier**

vocabulary and resource namespaces are used to describe auxiliary resources

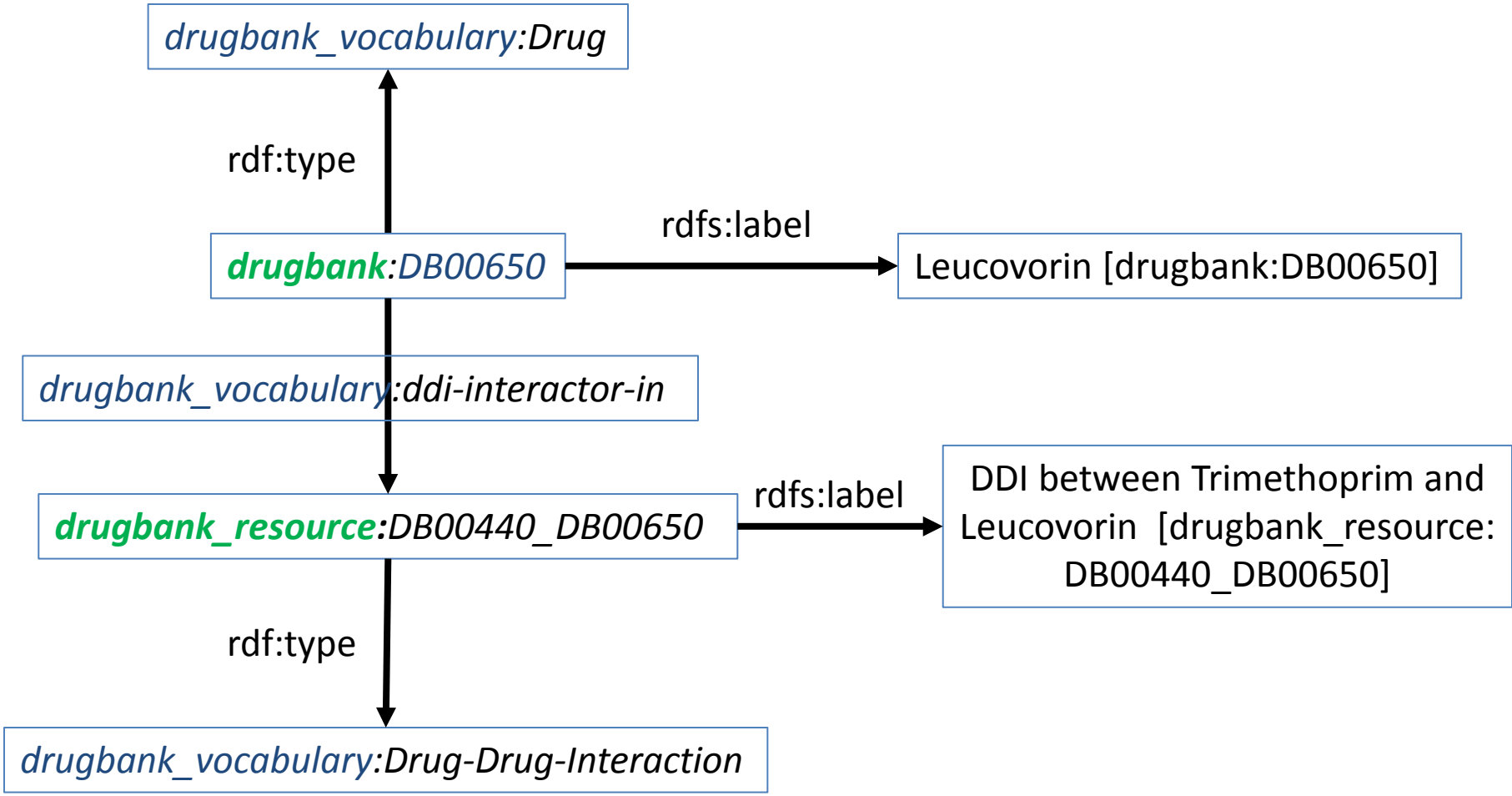
- types and predicates that are generated to support the semantic annotation are in the **vocabulary** namespace

  *http://bio2rdf.org/**drugbank_vocabulary**:Drug (type)*
  *http://bio2rdf.org/**drugbank_vocabulary**:target (predicate)*

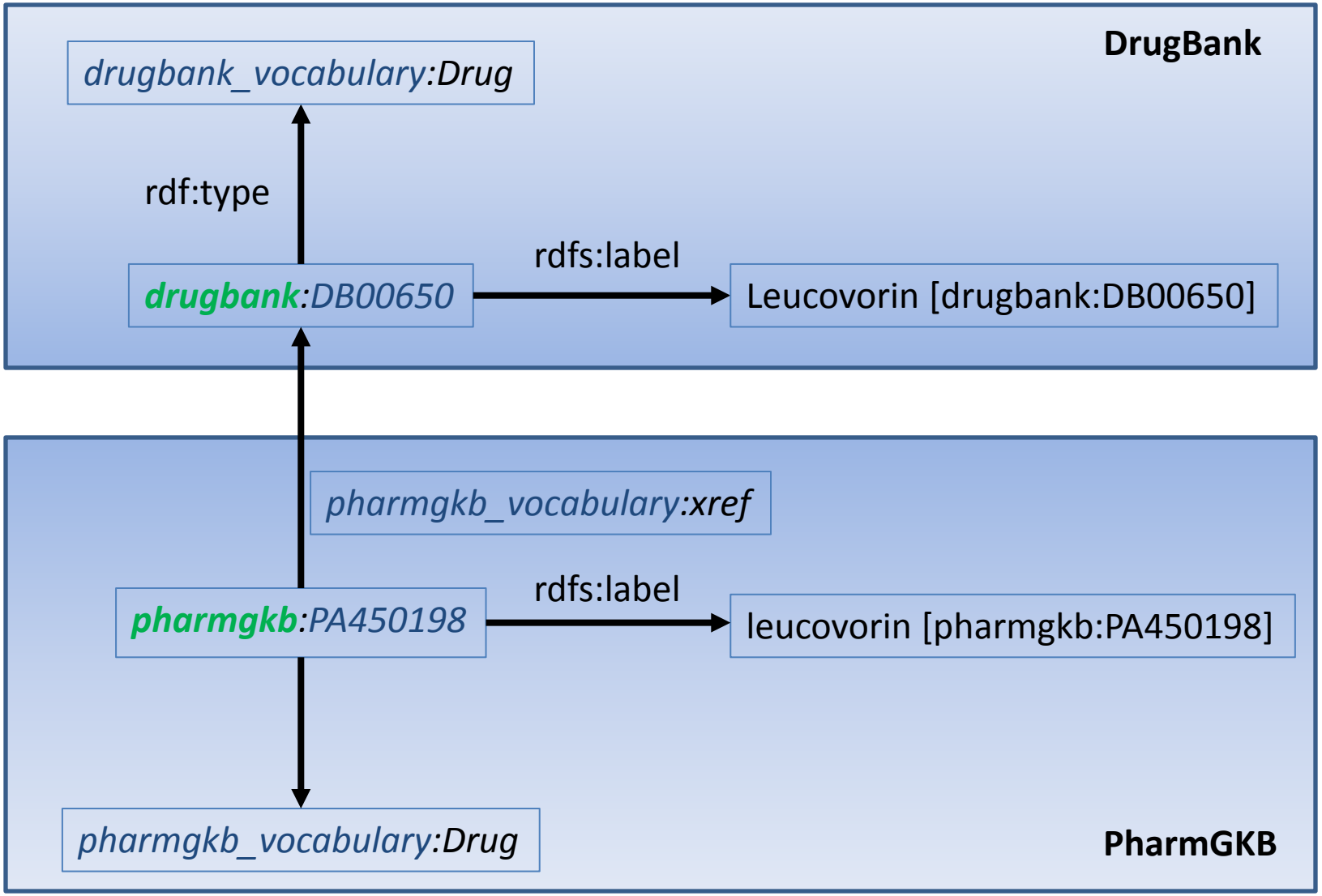
- *n-ary* relations are named in the **resource** namespace

  *http://bio2rdf.org/**drugbank_resource**:DB00440_DB00650*

Every statement expands the network of linked data



Syntactic integration across datasets



You can get what links to it

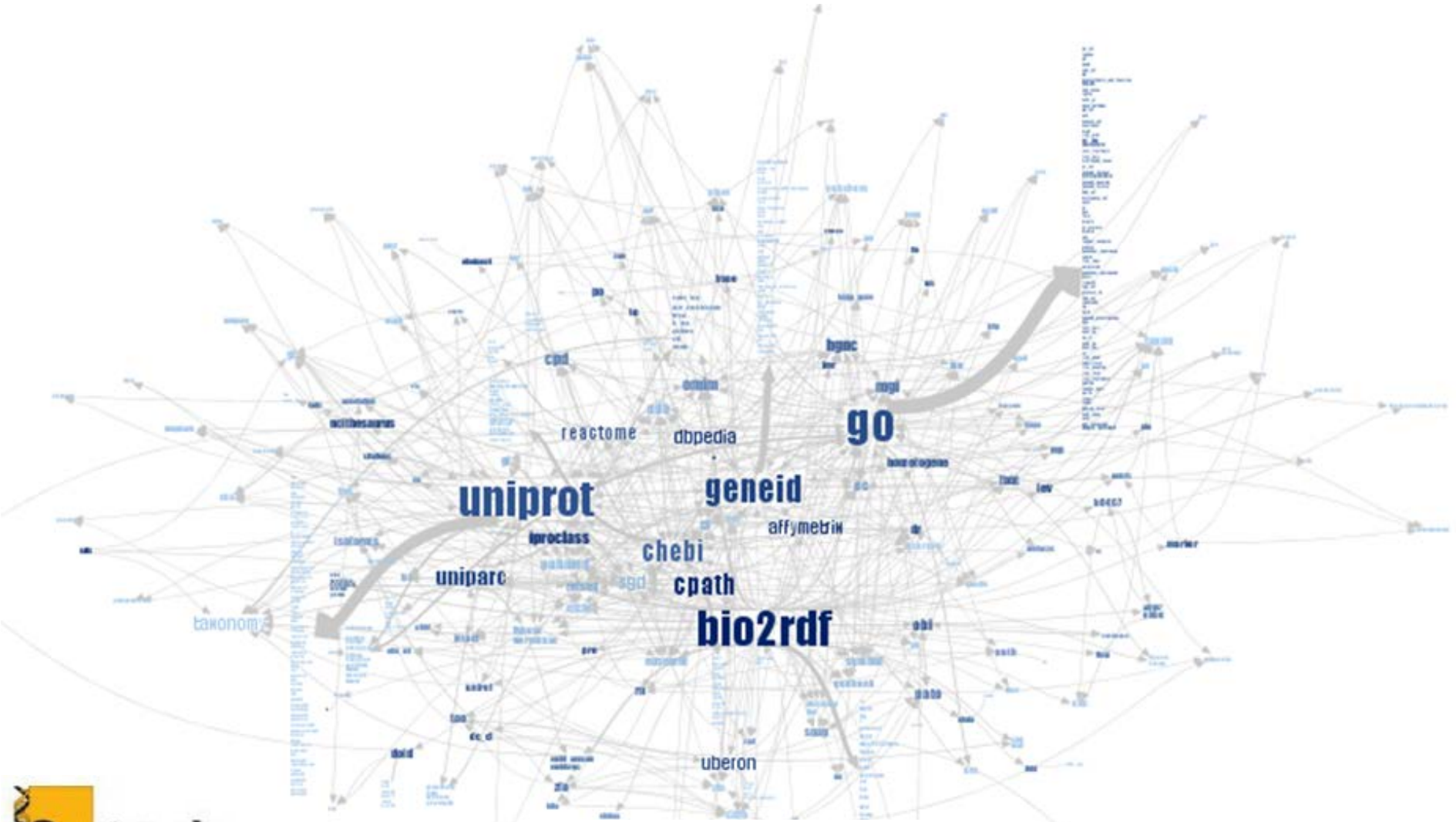
What links to DrugBank's Leucovorin?

http://bio2rdf.org/linksns/drugbank/drugbank:DB00650

Subject	Predicate	Object
http://bio2rdf.org/drugbank_resource:DB00650_359	http://bio2rdf.org/drugbank_vocabulary:drug	http://bio2rdf.org/drugbank:DB00650
	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://bio2rdf.org/drugbank_vocabulary:Drug-Target-Interaction
	http://www.w3.org/2000/01/rdf-schema#label	drug-target interaction Leucovorin and Thymidylate synthase [drugbank_resource:DB00650_359]
http://bio2rdf.org/drugbank_resource:DB00650_drugbank_target:1709	http://bio2rdf.org/drugbank_vocabulary:drug	http://bio2rdf.org/drugbank:DB00650
	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://bio2rdf.org/drugbank_vocabulary:Drug-Transporter-Interaction
	http://www.w3.org/2000/01/rdf-schema#label	drug-transporter interaction Leucovorin and Canalicular multispecific organic anion transporter 2 [drugbank_resource:DB00650_drugbank_target:1709]
http://bio2rdf.org/drugbank_resource:DB00650_drugbank_target:1735	http://bio2rdf.org/drugbank_vocabulary:drug	http://bio2rdf.org/drugbank:DB00650
	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://bio2rdf.org/drugbank_vocabulary:Drug-Transporter-Interaction
	http://www.w3.org/2000/01/rdf-schema#label	drug-transporter interaction Leucovorin and Canalicular multispecific organic anion transporter 1 [drugbank_resource:DB00650_drugbank_target:1735]
http://bio2rdf.org/drugbank_resource:DB00650_drugbank_target:2164	http://bio2rdf.org/drugbank_vocabulary:drug	http://bio2rdf.org/drugbank:DB00650
	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://bio2rdf.org/drugbank_vocabulary:Drug-Transporter-Interaction
	http://www.w3.org/2000/01/rdf-schema#label	drug-transporter interaction Leucovorin and Multidrug resistance-associated protein 4 [drugbank_resource:DB00650_drugbank_target:2164]
http://bio2rdf.org/pharmgkb:PA450198	http://bio2rdf.org/pharmgkb_vocabulary:xref	http://bio2rdf.org/drugbank:DB00650
	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://bio2rdf.org/pharmgkb_vocabulary:Drug
	http://www.w3.org/2000/01/rdf-schema#label	leucovorin [pharmgkb:PA450198]

Powered by Bio2RDF/1.4-SNAPSHOT | View as RDF/XML | View as N3 | View as HTML | View as JSON

The Bio2RDF Network of Linked Data



Centre de Recherche du CHUL (CHUQ)

Université Laval

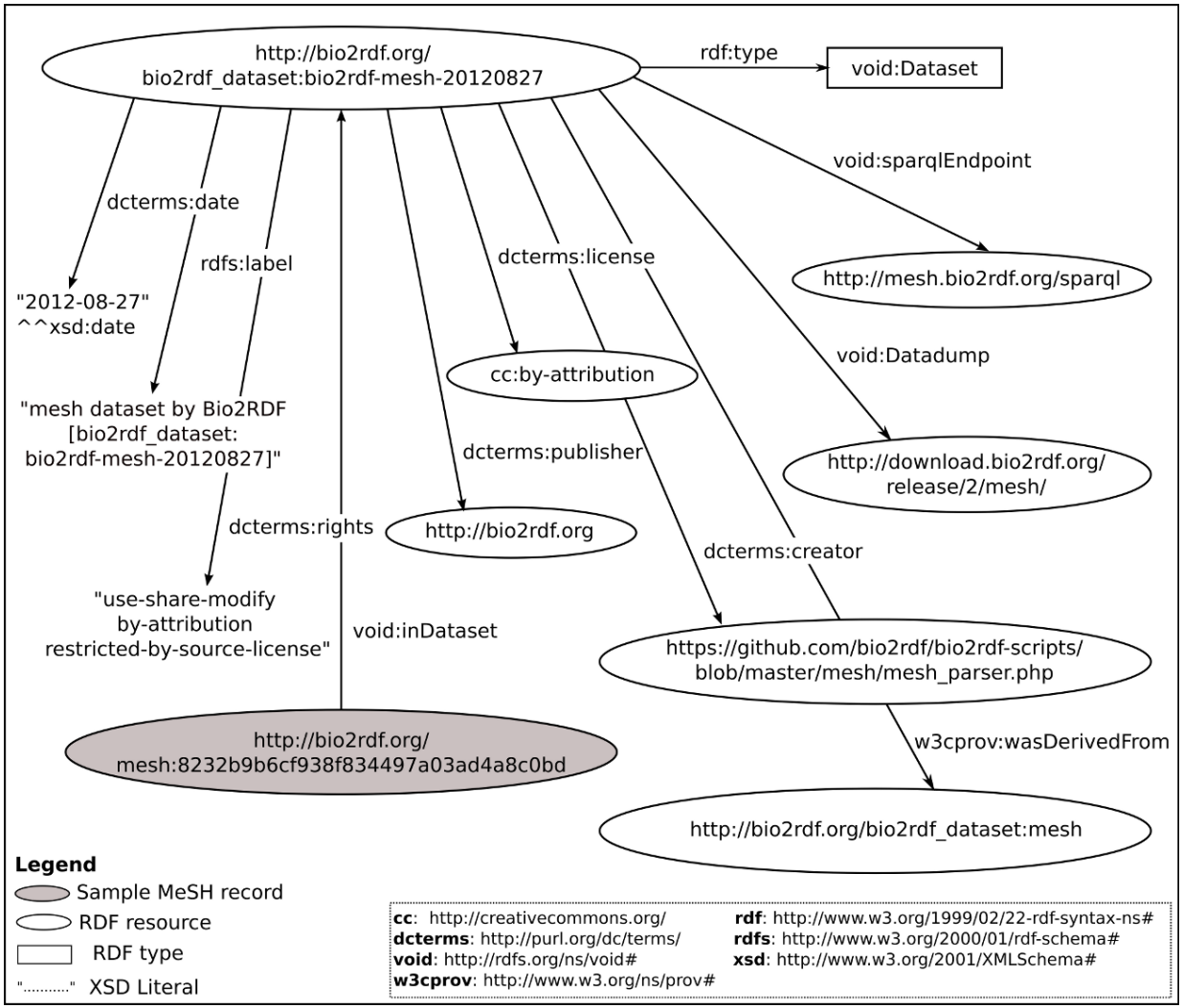
Queensland University of Technology
Microsoft QUT eResearch Centre

Carleton University

Powered by Virtuoso

LinkingOpenData

Every Bio2RDF dataset now contains provenance metadata



Features

- Entity-Dataset
- Date Created
- License & Rights
- Source
- Creator
- Publisher
- Availability
- SPARQL endpoint
- Data dump

Vocabularies

VoID
Dublin Core
W3C Provenance
Bio2RDF vocabulary

For every Bio2RDF dataset we pre-compute 9 descriptive metrics

- total number of triples
- number of unique subjects
- number of unique predicates
- number of unique objects
- number of unique types
- unique predicate-object links and their frequencies
- unique predicate-literal links and their frequencies
- *unique subject type-predicate-object type links and their frequencies*
- unique subject type-predicate-literal links and their frequencies

Accessing Bio2RDF dataset metrics

- Each Bio2RDF endpoint contains a named graph that holds the pre-computed metrics
  - *http://bio2rdf.org/bio2rdf-[namespace]-statistics*
- Metrics can be queried using SPARQL, *e.g.*:

```
SELECT *
FROM <http://bio2rdf.org/bio2rdf-drugbank-statistics>
WHERE {
        ?dataset a <http://bio2rdf.org/dataset_vocabulary:Endpoint> .
        ?dataset <http://bio2rdf.org/dataset_vocabulary:has_triple_count> ?tc  .
        ?dataset <http://bio2rdf.org/dataset_vocabulary:has_unique_subject_count> ?sc .
        ?dataset <http://bio2rdf.org/dataset_vocabulary:has_unique_predicate_count> ?pc .
        ?dataset <http://bio2rdf.org/dataset_vocabulary:has_unique_object_count> ?oc .
        …
}
```

https://github.com/bio2rdf/bio2rdf-scripts/wiki/Bio2RDF-dataset-metrics

Graph summaries can also assist in query formulation

Subject Type	Subject Count	Predicate	Object Type	Object Count
Pharmaceutical	11512	form	Unit	56
Drug-Transporter-Interaction	1440	drug	Drug	534
Drug-Transporter-Interaction	1440	transporter	Target	88
Drug	1266	dosage	Dosage	230
Patent	1255	country	Country	2
Drug	1127	product	Pharmaceutical	11512
Drug	1074	ddi-interactor-in	Drug-Drug-Interaction	10891
Drug	532	patent	Patent	1255
Drug	277	mixture	Mixture	3317
Dosage	230	route	Route	42
Drug-Target-Interaction	84	target	Target	43

```
PREFIX drugbank_vocabulary: <http://bio2rdf.org/drugbank_vocabulary:>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?ddi ?d1name ?d2name
WHERE {
        ?ddi a drugbank_vocabulary:Drug-Drug-Interaction .
        ?d1 drugbank_vocabulary:ddi-interactor-in ?ddi .
        ?d1 rdfs:label ?d1name .
        ?d2 drugbank_vocabulary:ddi-interactor-in ?ddi .
        ?d2 rdfs:label ?d2name.
    FILTER (?d1 != ?d2)
}
```

You can use the SPARQLed query assistant with updated endpoints

http://sindicetech.com/sindice-suite/sparqled/

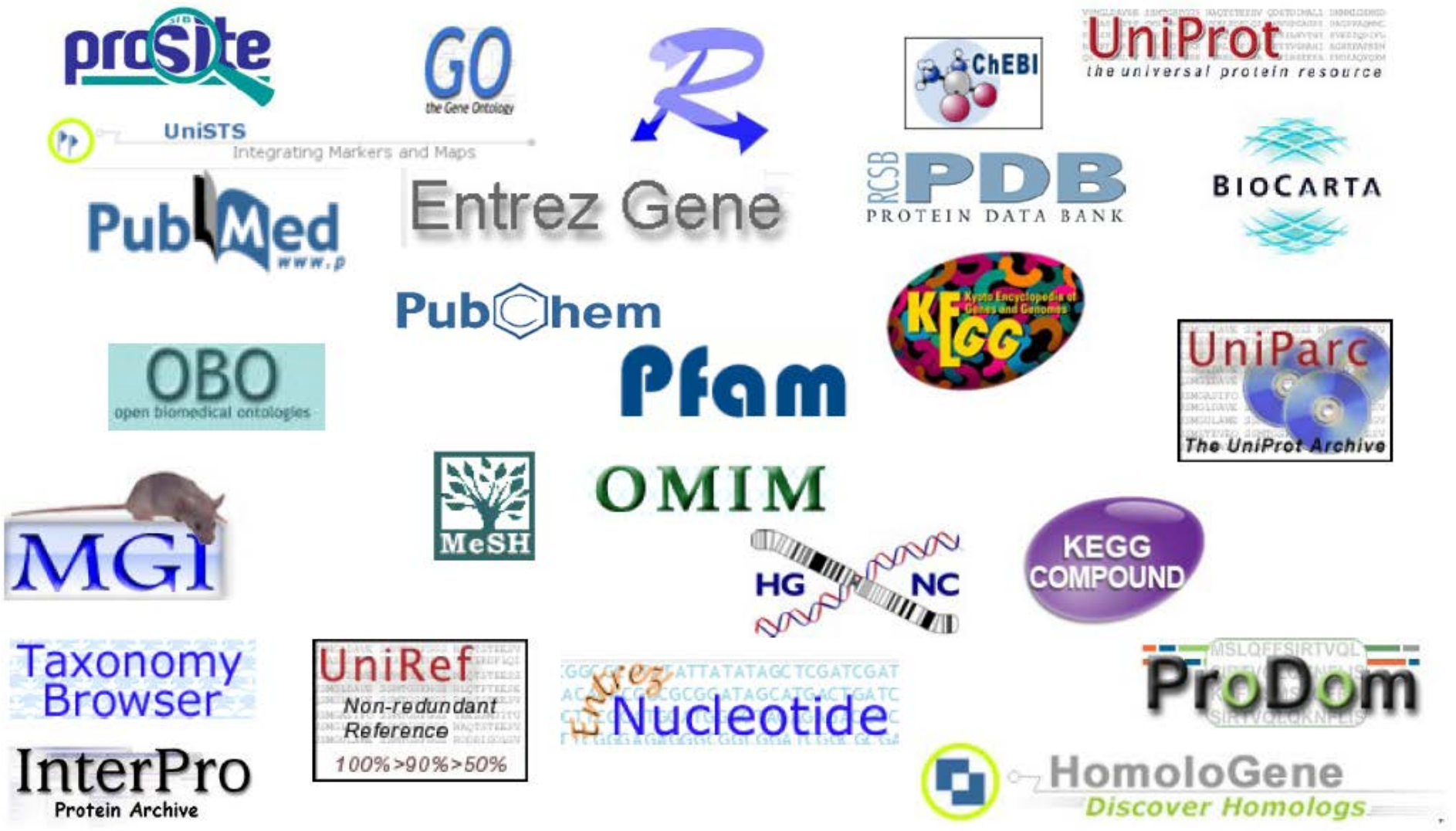
```
1  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2
3  SELECT ?drugname ?genename WHERE {
4     ?s a <http://bio2rdf.org/pharmgkb_vocabulary:Drug-Gene-Association> .
5     ?s <
6  }
7  LIMIT 1
8
```

<http://bio2rdf.org/pharmgkb_vocabulary:association_type>
<http://www.w3.org/2000/01/rdf-schema#label>
<http://bio2rdf.org/pharmgkb_vocabulary:drug>
<http://bio2rdf.org/pharmgkb_vocabulary:gene>
<http://rdfs.org/ns/void#nDataset>
<http://bio2rdf.org/pharmgkb_vocabulary:article>
<http://bio2rdf.org/pharmgkb_vocabulary:pd_relationship>
<http://bio2rdf.org/pharmgkb_vocabulary:pk_relationship>

```
1  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2
3  SELECT ?drugname ?genename WHERE {
4     ?s a <http://bio2rdf.org/pharmgkb_vocabulary:Drug-Gene-Association> .
5     ?s <http://bio2rdf.org/pharmgkb_vocabulary:drug> ?drug .
6     ?s <http://bio2rdf.org/pharmgkb_vocabulary:gene> ?gene .
7     ?drug <http://www.w3.org/2000/01/rdf-schema#label> ?drugname .
8     ?gene <http://www.w3.org/2000/01/rdf-schema#label> ?genename .
9  }
10 LIMIT 10
```

graph: http://sindicetech.com/analytics

Bio2RDF covers the major biological databases



Bio2RDF Release 2 – New and Updated Datasets

Dataset	Namespace	# of triples
Affymetrix	affymetrix	44469611
Biomodels*	biomodels	589753
Comparative Toxicogenomics Database	ctd	141845167
DrugBank	drugbank	1121468
NCBI Gene	ncbigene	394026267
Gene Ontology Annotations	goa	80028873
HUGO Gene Nomenclature Committee	hgnc	836060
Homologene	homologene	1281881
InterPro*	interpro	999031
iProClass	iproclass	211365460
iRefIndex	irefindex	31042135
Medical Subject Headings	mesh	4172230
NCBO BioPortal*	bioportal	15384622
National Drug Code Directory*	ndc	17814216
Online Mendelian Inheritance in Man	omim	1848729
Pharmacogenomics Knowledge Base	pharmgkb	37949275
SABIO-RK*	sabiork	2618288
Saccharomyces Genome Database	sgd	5551009
NCBI Taxonomy	taxon	17814216
Total	**19**	**1010758291**

Status of Bio2RDF Release 1 datasets

Dataset	Status
Atlas	Maintained – will not be updated
BIND	Deprecated – in iRefIndex
BioCarta	Deprecated – in Pathway Commons
BioCyc	Deprecated – in Pathway Commons
EC	Deprecated – in Gene Ontology/UniProt
GenBank	Maintained – will be updated
HHPID	Maintained – will not be updated
INOH	Deprecated – in Pathway Commons
KEGG	Maintained – will not be updated
MGI	Maintained – will be updated
Pubmed	Maintained – will be updated
PID	Deprecated – in Pathway Commons
Reactome	Deprecated – in Pathway Commons
RefSeq	Maintained – will be updated

A PHP-based library acts a point of integration

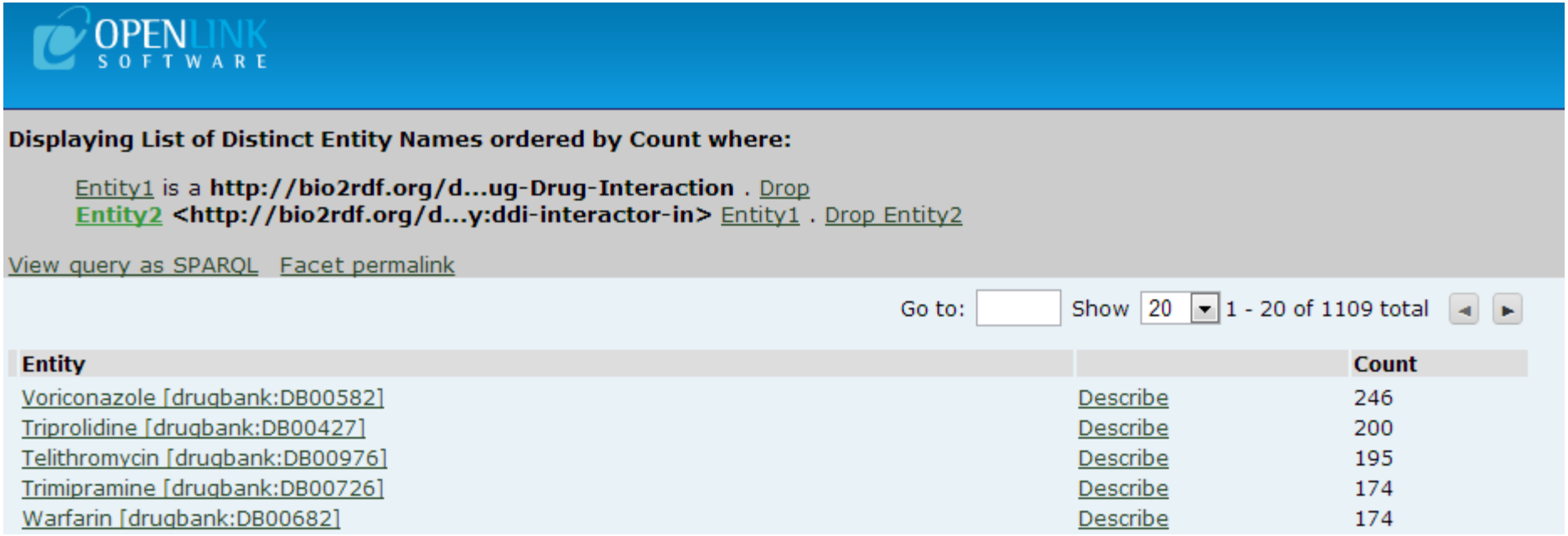
- Provides a set of APIs
  - to produce RDF and OWL statements
  - to generate valid Bio2RDF URIs by checking against a dataset registry
  - to generate dataset provenance

Oh Bio2RDF, how can I access you?

Let me count the ways:

1. Downloads (data, stats, virtuoso db)
2. Web interface (lookup + services)
3. SPARQL endpoint
4. SPARQLed editor
5. Virtuoso Faceted Browser

Use virtuoso's built in faceted browser to construct increasingly complex queries with little effort



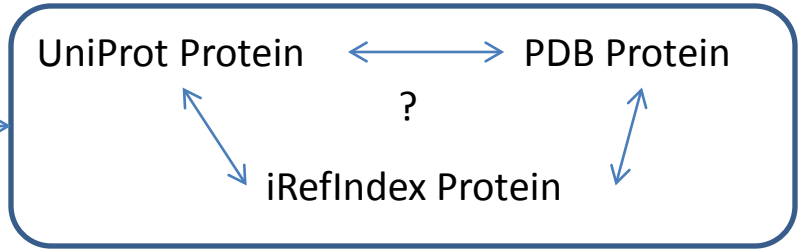
Third Party Software

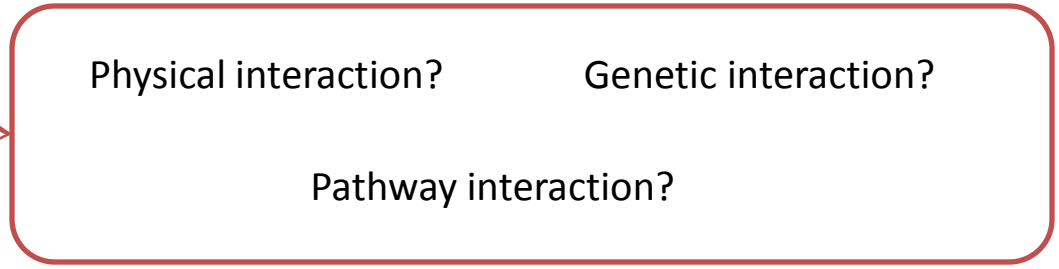
- IO Informatics – Sentient Knowledge Explorer
- Metaome – Distillbio
- Fluid Ops – Information Workbench

Heterogeneous biological data on the semantic web is difficult to query

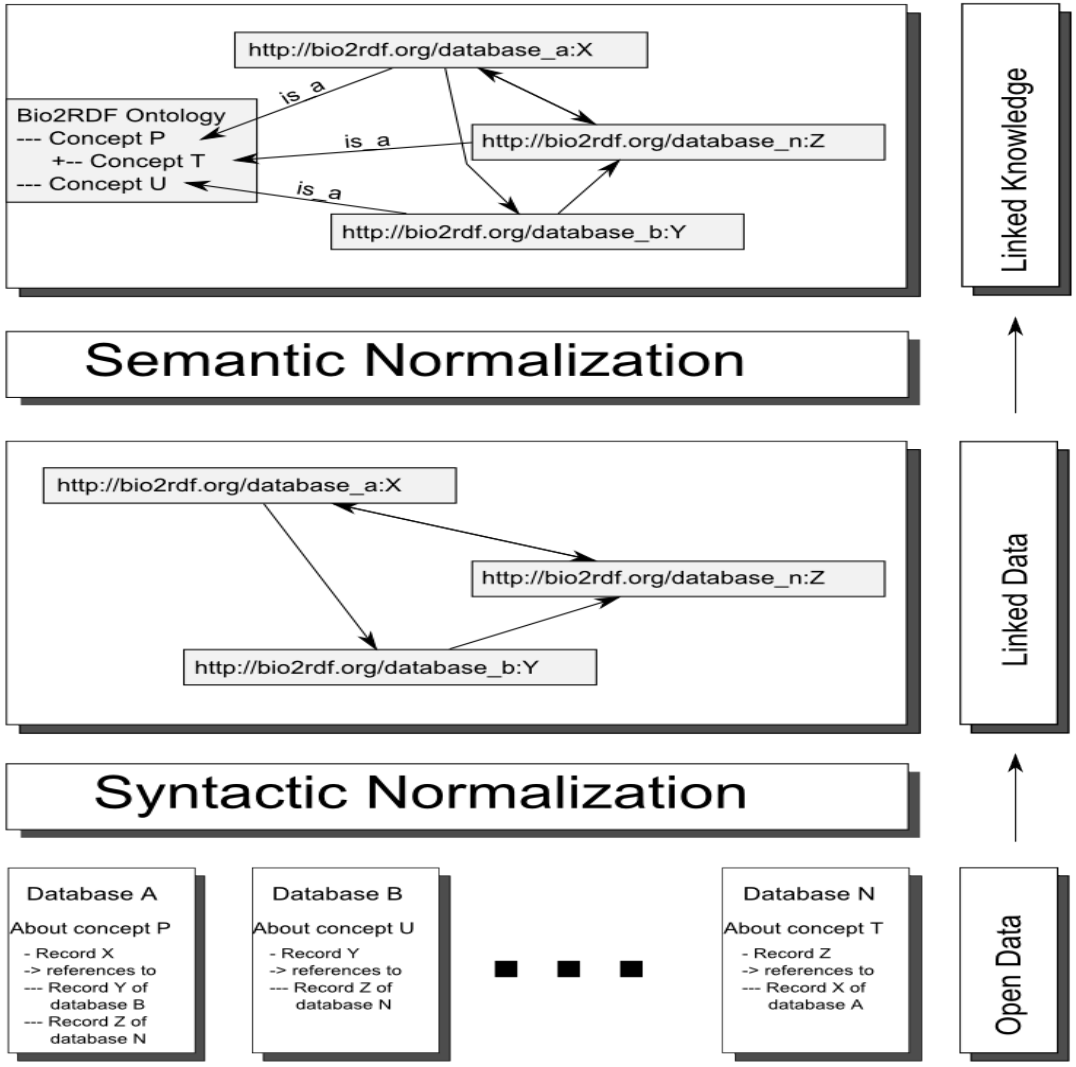
Question: Find all proteins that interact with beta amyloid (uniprot:P05067)

```
SELECT * WHERE {
   ?protein a bio2rdf:Protein .
   ?protein bio2rdf:interacts_with uniprot:P05067 .
}
```





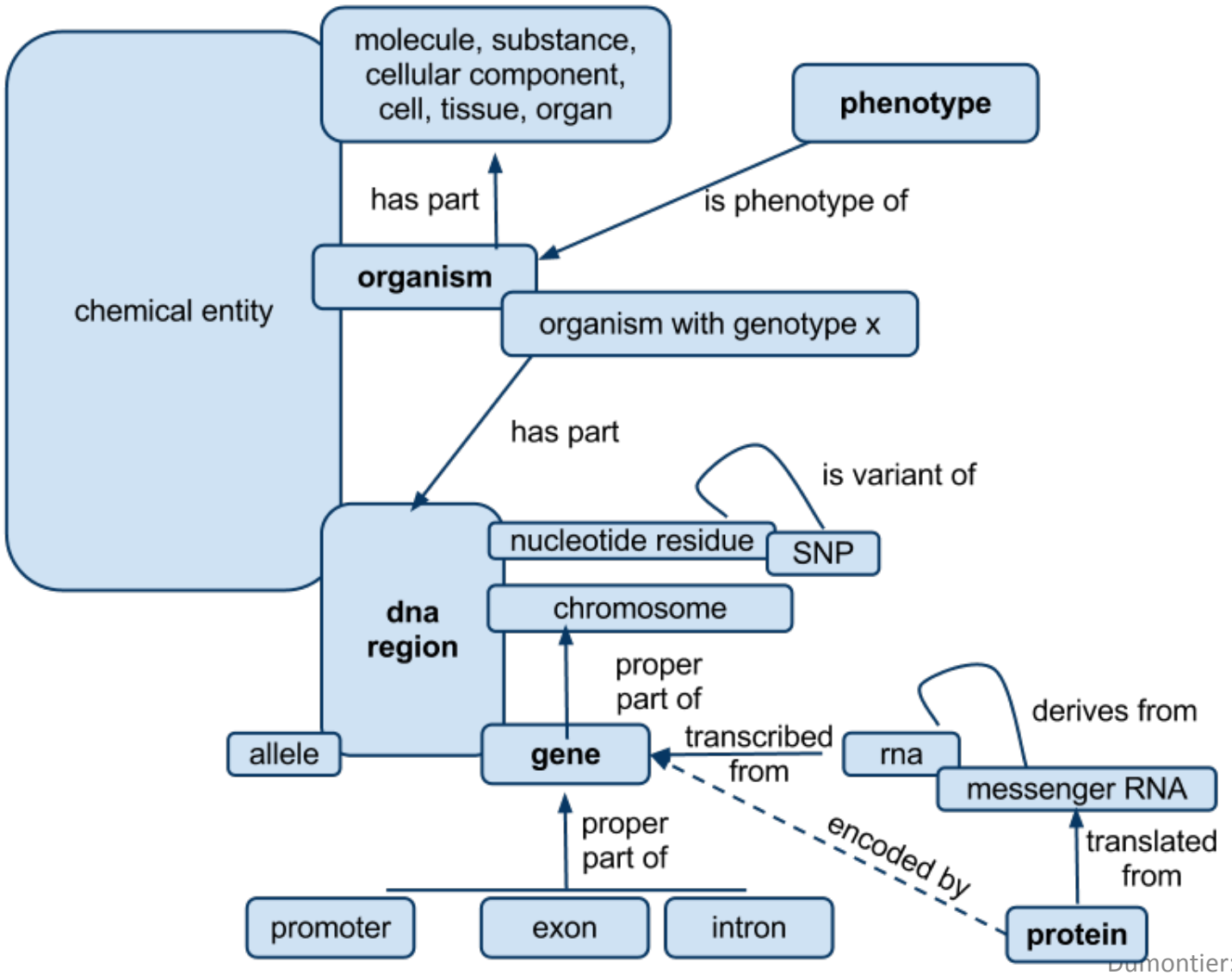
RDF-based Linked Data is a great first step, but it's not enough.

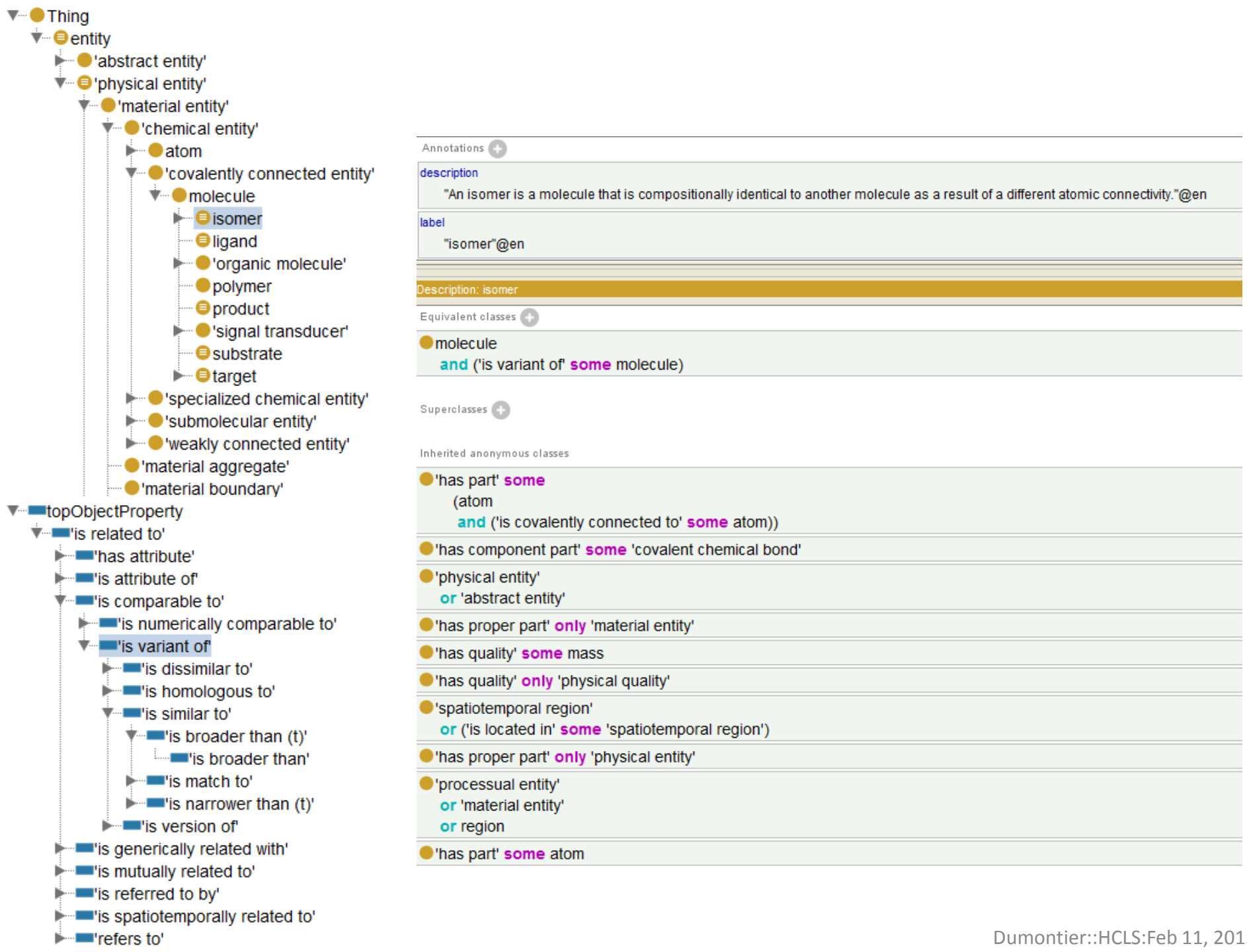


From linked data to linked knowledge through syntactic and semantic normalization.

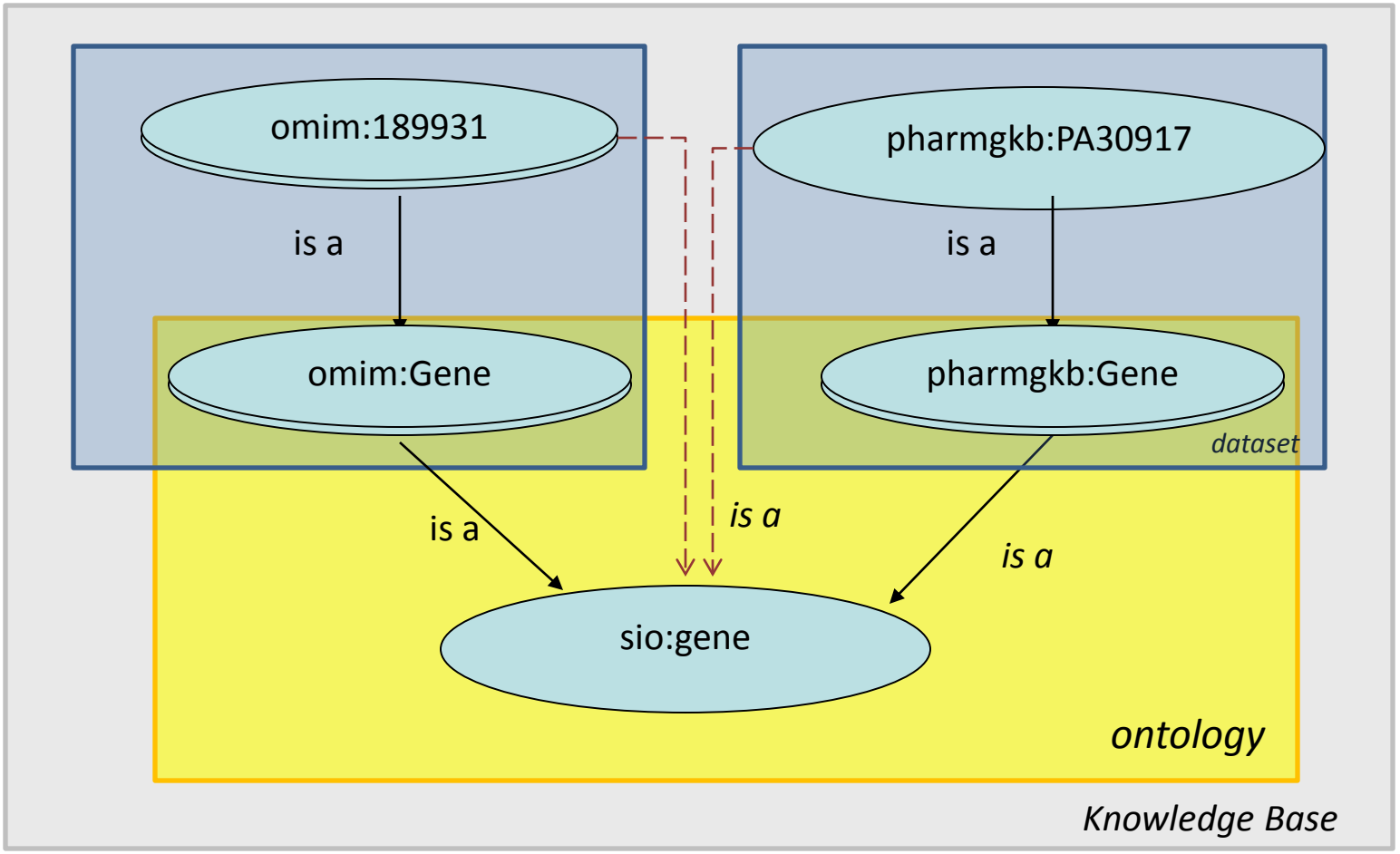
ontology as a strategy to formally represent and integrate knowledge

SIO provides an OWL ontology for the representation of diverse biomedical knowledge




Thing
entity
'abstract entity'
'physical entity'
'material entity'
'chemical entity'
atom
'covalently connected entity'
molecule
isomer
ligand
'organic molecule'
polymer
product
'signal transducer'
substrate
target
'specialized chemical entity'
'submolecular entity'
'weakly connected entity'
'material aggregate'
'material boundary'
topObjectProperty
'is related to'
'has attribute'
'is attribute of'
'is comparable to'
'is numerically comparable to'
'is variant of'
'is dissimilar to'
'is homologous to'
'is similar to'
'is broader than (t)'
'is broader than'
'is match to'
'is narrower than (t)'
'is version of'
'is generically related with'
'is mutually related to'
'is referred to by'
'is spatiotemporally related to'
'refers to'
Annotations
description
"An isomer is a molecule that is compositionally identical to another molecule as a result of a different atomic connectivity."@en
label
"isomer"@en
Description: isomer
Equivalent classes
molecule
and ('is variant of' some molecule)
Superclasses
Inherited anonymous classes
'has part' some
(atom
and ('is covalently connected to' some atom))
'has component part' some 'covalent chemical bond'
'physical entity'
or 'abstract entity'
'has proper part' only 'material entity'
'has quality' some mass
'has quality' only 'physical quality'
'spatiotemporal region'
or ('is located in' some 'spatiotemporal region')
'has proper part' only 'physical entity'
'processual entity'
or 'material entity'
or region
'has part' some atom

Semantic data integration, consistency checking and query answering over Bio2RDF with the Semanticscience Integrated Ontology (SIO)



Querying Bio2RDF Linked Open Data with a Global Schema. Alison Callahan, José Cruz-Toledo and Michel Dumontier. Bio-ontologies 2012.

Bio2RDF types include processes, material entities and informational entities

CTD: Chemical, Disease, Chemical-Disease Interaction, Chemical-Gene Interaction

NCBIGene: Gene, Protein, Model Organism, Publication

HGNC: Accession Number, Gene, Gene Symbol

iRefIndex: Protein Complex, Protein Interaction

MGI: Gene Marker, Gene Symbol

PharmGKB: Gene-Disease Associations, Disease, Drug, Gene

SGD: Enzyme, Pathway, Protein, RNA, Reaction,

Location, Experiment

https://github.com/bio2rdf/bio2rdf-mapping

Bio2RDF and SIO powered SPARQL 1.1 federated query: Find chemicals in CTD and proteins in SGD that participate in the same GO process

```
SELECT ?chem, ?prot, ?proc
FROM <http://bio2rdf.org/ctd>
WHERE {
      ?chemical a sio:chemical-entity.
      ?chemical rdfs:label ?chem.
      ?chemical sio:is-participant-in ?process.
      ?process rdfs:label ?proc.
     FILTER regex (?process, "http://bio2rdf.org/go:")
      SERVICE <http://sgd.bio2rdf.org/sparql> {
                ?protein a sio:protein .
                ?protein sio:is-participant-in ?process.
                ?protein rdfs:label ?prot .
      }
}
```

Bio2RDF Release 2 – A summary

- Updated data conversion source code to use PHP API (all available through GitHub)
- Simple Bio2RDF IRI design patterns that facilitate syntactic consistency and interoperability backed by simple registry
- Dataset provenance and metrics
- We welcome contributions from the community
    - **Join our mailing list at bio2rdf@googlegroups.com**

Future Directions

- Aiming for twice yearly release schedule
- Update large scale datasets
  - RefSeq, Genbank, PubMed, PDB
- Incorporate EBI & W3C Linking Open Drug Data (LODD) effort
  - **SIDER** (beta), TCM (RDF available), CHEMBL (in dev), **OMIM** (released), DailyMed (RDF available), **LinkedCT** (beta)
- Extended dataset coverage by tapping into existing endpoints (uniprot, bioportal, ebi-rdf?)
- OpenBioCloud w/DERI + SindiceTech
- Showcase with other third party tools

Acknowledgements

Bio2RDF

Peter Ansell, Francois Belleau, **Allison Callahan**, Jacques Corbeil, **Jose Cruz-Toledo**, Alex De Leon, Steve Etlinger, James Hogan, Nichealla Keath, Jean Morissette, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault and Paul Roe

OpenBioCloud

Dana Klassen and Giovanni Tumarello

SADI: *Christopher Baker*, Melanie Courtot, Jose Cruz-Toledo, Steve Etlinger, Nichealla Keath, Artjom Klein, Luke McCarthy, Silvane Paixao, Ben Vandervalk, Natalia Villanueva-Rosales, *Mark Wilkinson*

W3C HCLS: *J Luciano, B Andersson, C Batchelor, O Bodenreider, T Clark, C Denney, C Domarew, T Gambet, L Harland, A Jentzsch, V Kashyap, P Kos, J Kozlovsky, T Lebo, SM Marshall, JP McCusker, DL McGuinness, C Ogbuji, E Pichler, R Powers, E Prud hommeaux, M Samwald, L Schriml, PJ Tonellato, PL Whetzel, J Zhao, S Stephens, C Denney, J Luciano, J McGurk, Lynn Schriml, and Peter J. Tonellato.*

dumontierlab.com

michel_dumontier@carleton.ca

Website: http://dumontierlab.com

Presentations: http://slideshare.com/micheldumontier

MITACS
NSERC CRSNG
Canada Foundation for Innovation
Fondation canadienne pour l'innovation
canarie
Carleton UNIVERSITY
Health Canada
Ontario