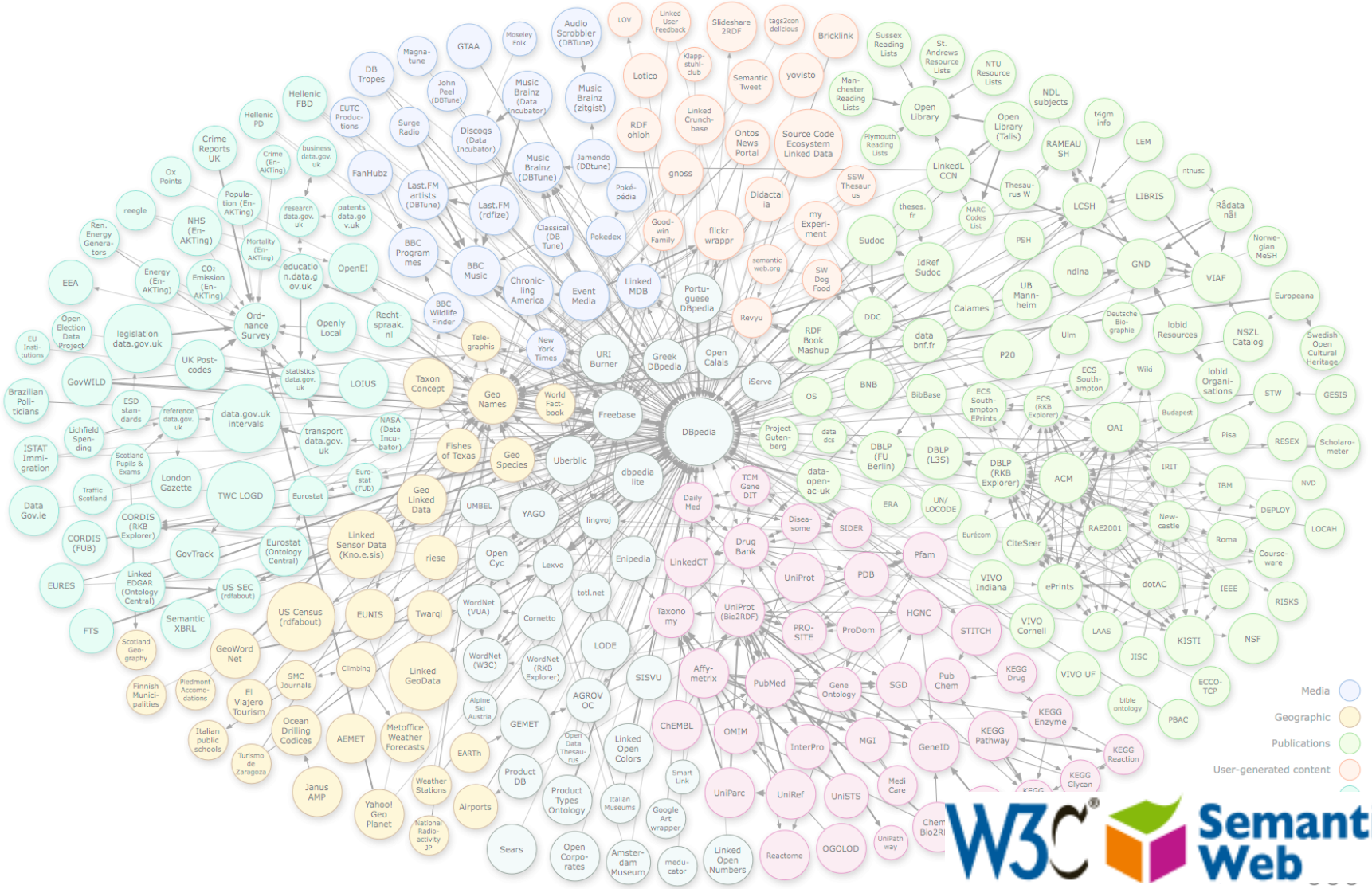
Tactical Formalization of Linked Open Data

Michel Dumontier, Ph.D.

Associate Professor of Medicine (Biomedical Informatics)
Stanford University

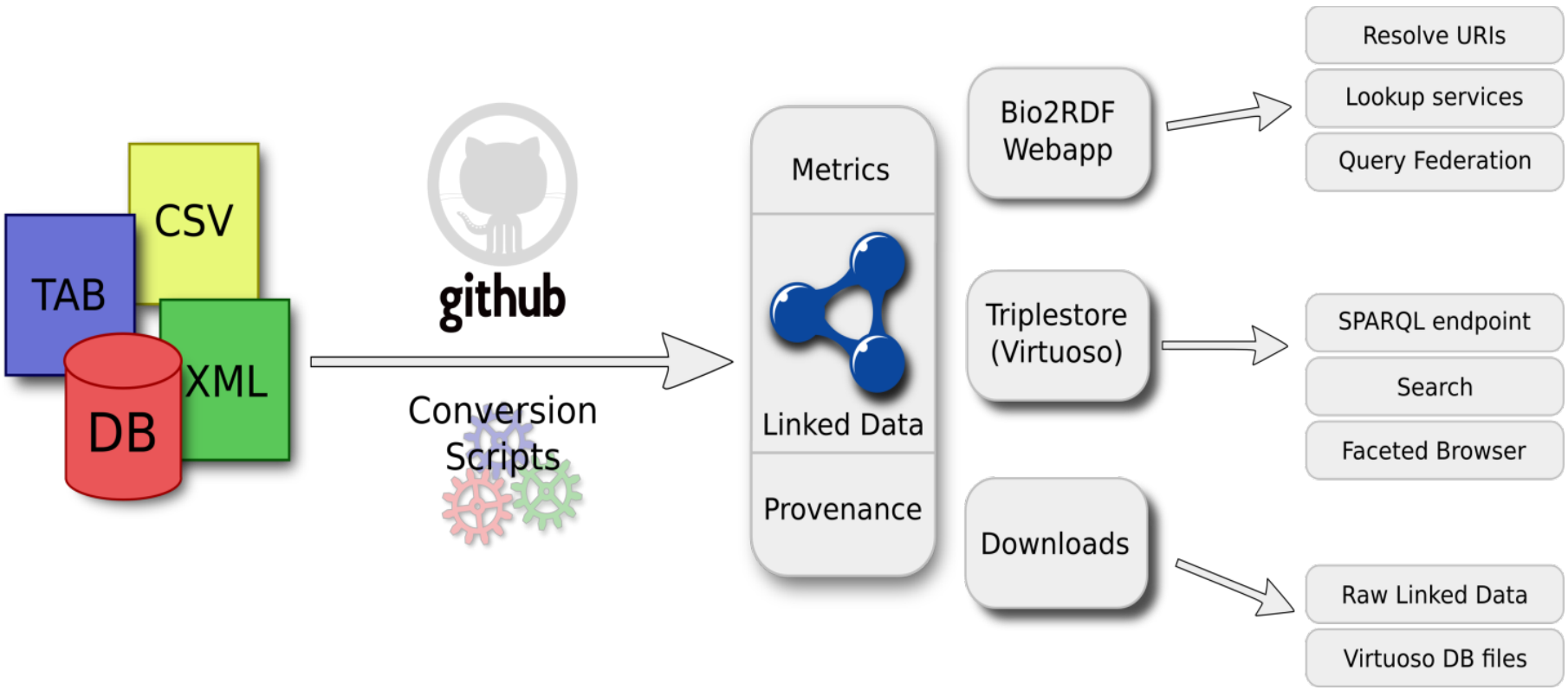
Linked Open Data provides an incredibly dynamic, rapidly growing set of interlinked resources



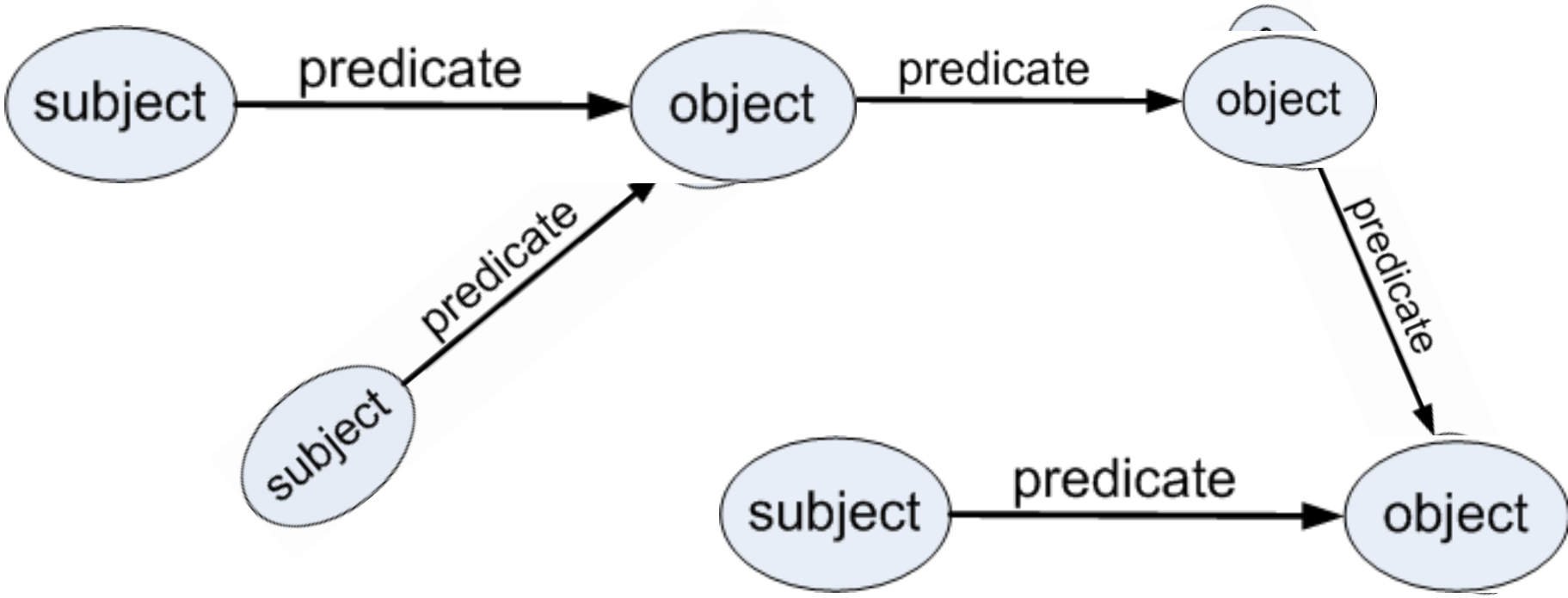
BIO2RDF

Linked Data for the Life Sciences

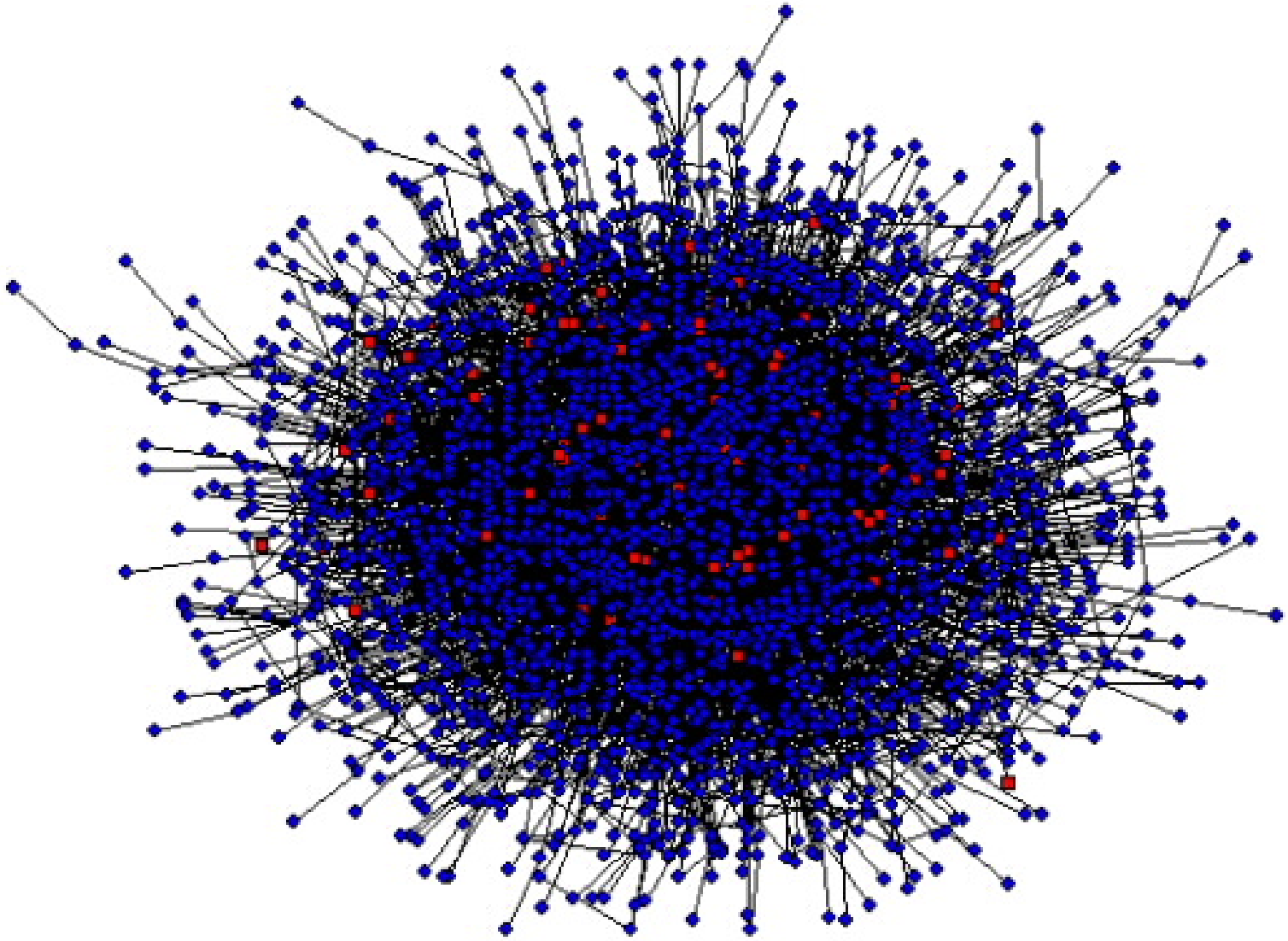
Bio2RDF converts bio-data in RDF format and ensures URI integrity by conferring with its registry of datasets



the simplicity of the triple makes it easy to proliferate



but the lack of coordination makes Linked Open Data is quite chaotic and unwieldy



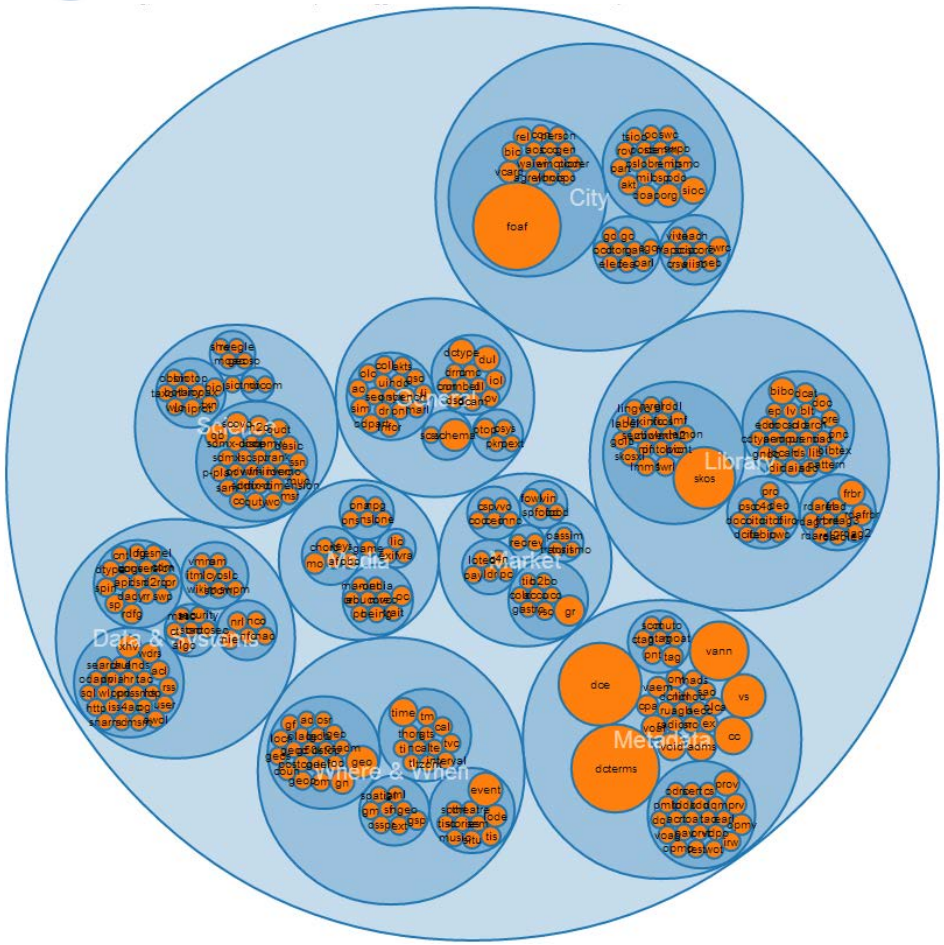
Massive Proliferation of Ontologies / Vocabularies

BioPortal

Statistics

Ontologies	378
Classes	5,964,932
Resources Indexed	39
Indexed Records	5,126,145
Direct Annotations	1,883,854,337
Direct Plus Expanded Annotations	24,828,631,205

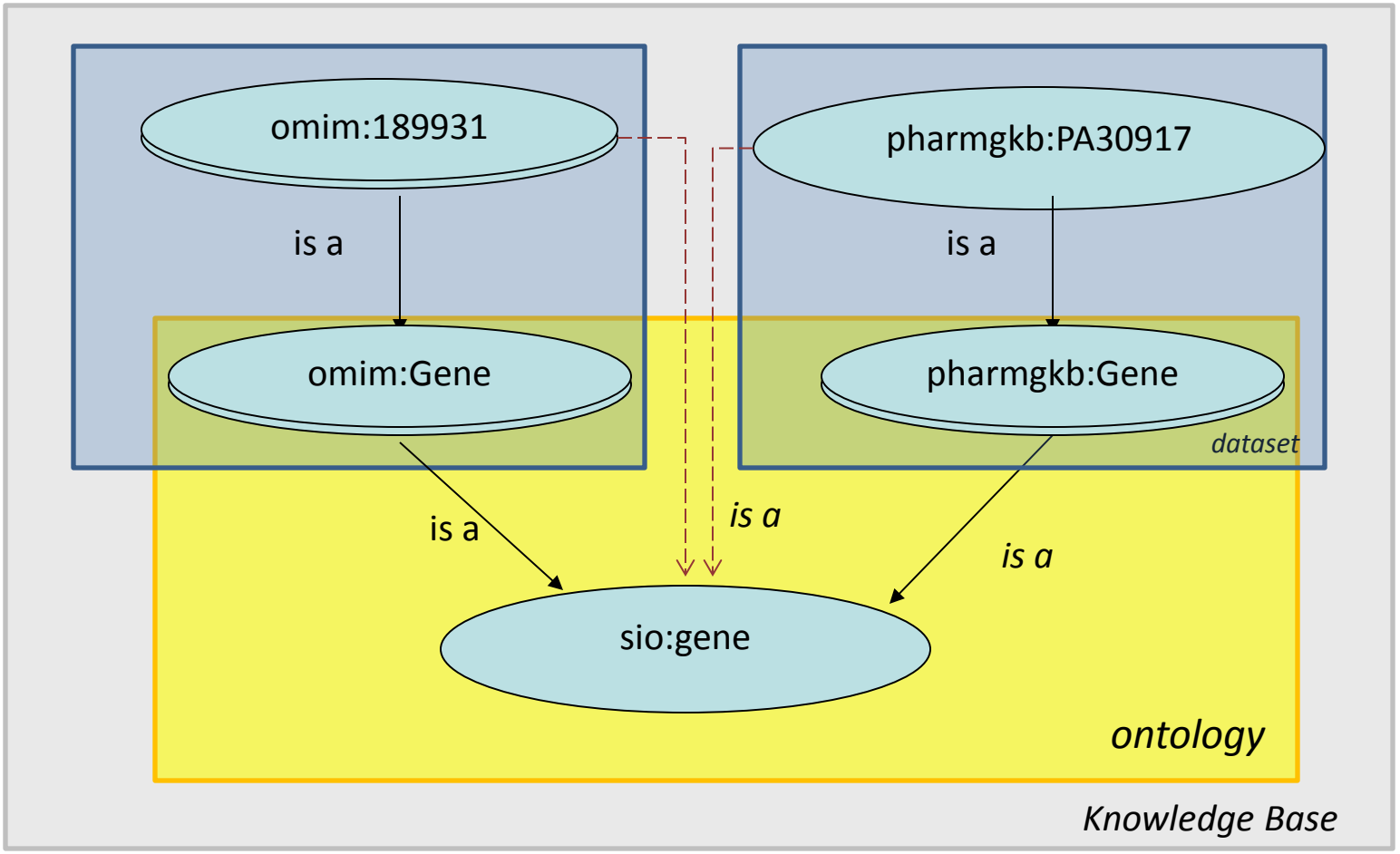
Linked Open Vocabularies (LOV)



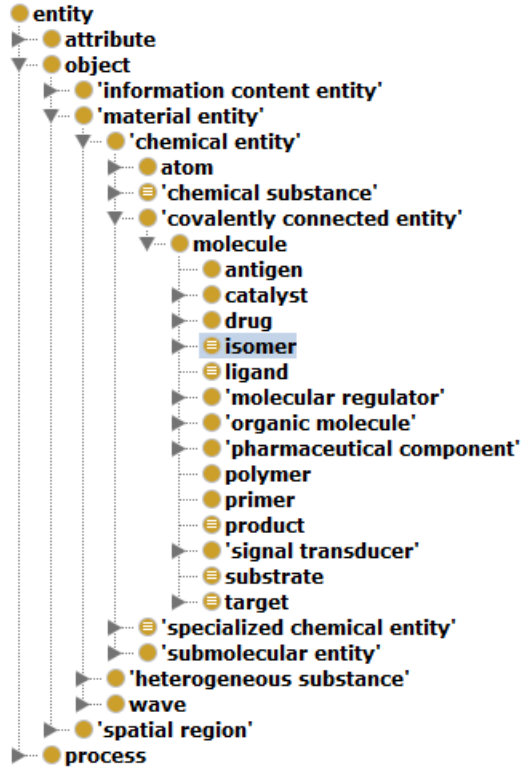
Despite all the data, it's still hard to find answers to questions

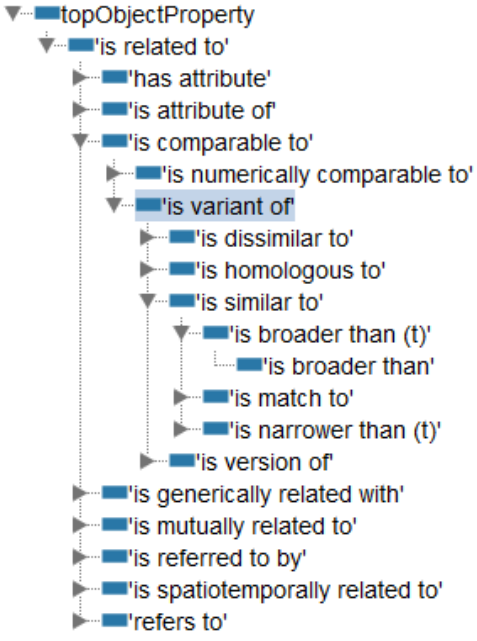
Because there are many ways to represent the same data
and each dataset represents it differently

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.





SRIQ(D)
10700+ axioms
1300+ classes
201 object properties (inc. inverses)
1 datatype property

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'

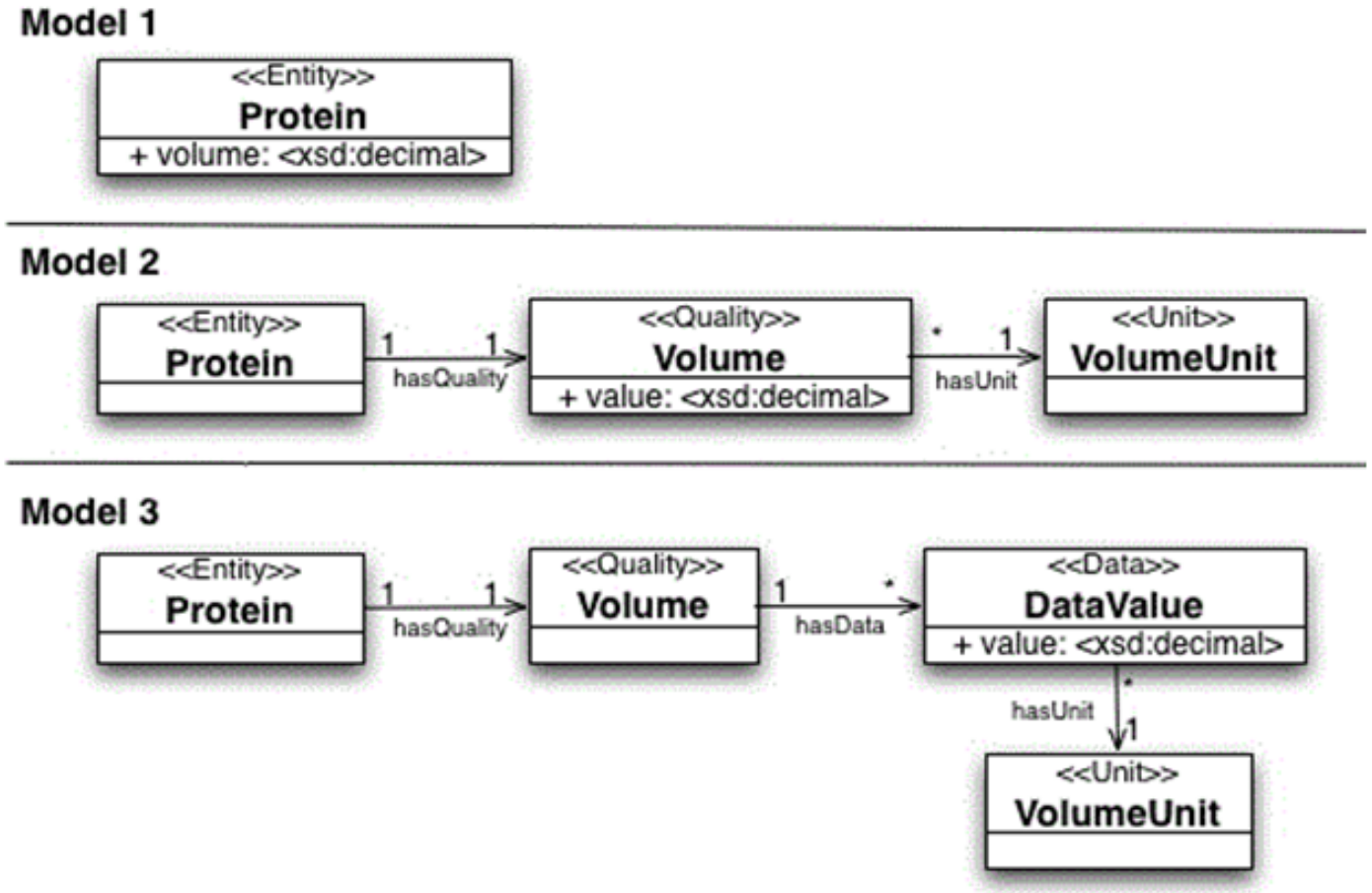
'processual entity'
or 'material entity'
or region

'has part' some atom

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

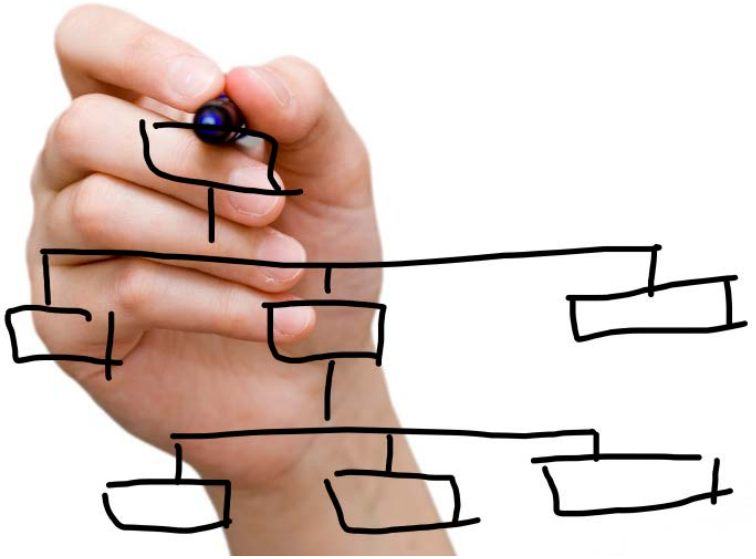
```
SELECT ?chem, ?prot, ?proc
FROM <http://bio2rdf.org/ctd>
WHERE {
      ?chemical a sio:chemical-entity.
      ?chemical rdfs:label ?chem.
      ?chemical sio:is-participant-in ?process.
      ?process rdfs:label ?proc.
     FILTER regex (?process, "http://bio2rdf.org/go:")
      SERVICE <http://sgd.bio2rdf.org/sparql>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

multiple formalizations of the same kind of data has emerged, each with their own merit



Three ways to model the relationship between a protein and the volume it occupies.

Multi-Stakeholder Efforts to Standardize Representations are Reasonable, *Long Term Strategies* for Data Integration



tactical formalization

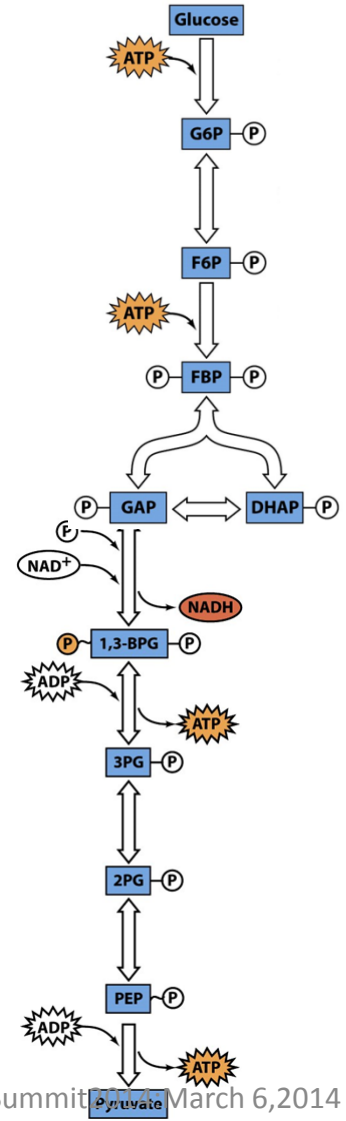
Take what you need
and represent it in a way that directly serves your objective

discovery of drug and disease pathway associations

Biological Pathways Define A Biological Objective

def: A biological pathway is constituted by a set of molecular components that undertake some biological transformation to achieve a stated objective

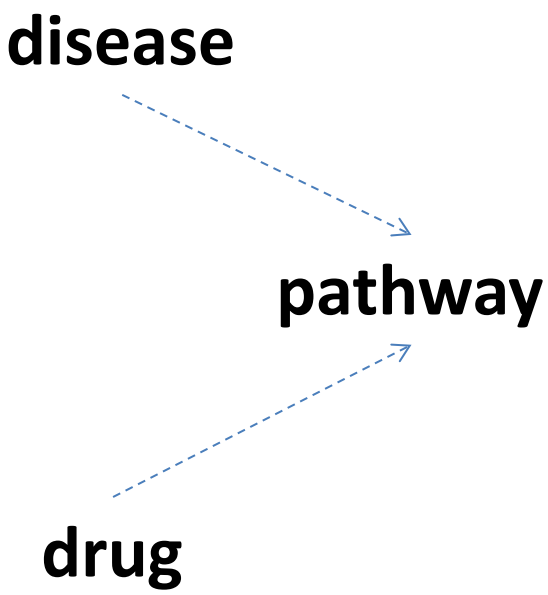
glycolysis : a pathway that converts glucose to pyruvate



aberrant and pharmacological pathways

Q1. Can we identify pathways that are associated with a particular disease or class of diseases?

Q2. Can we identify pathways are associated with a particular drug or class of drugs?



Identification of drug and disease enriched pathways

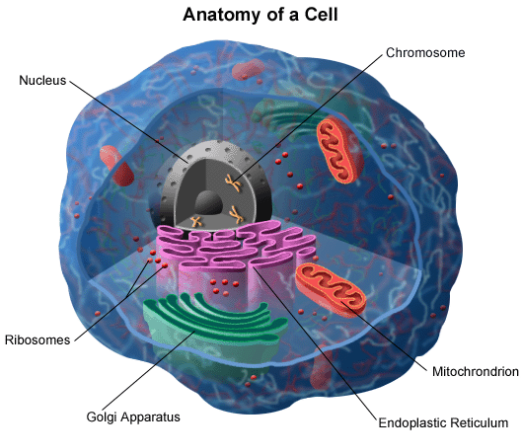
- Approach
  - Integrate 3 datasets
    - **DrugBank, PharmGKB and CTD**
  - Integrate 7 terminologies
    - **MeSH, ATC, ChEBI, UMLS, SNOMED, ICD, DO**
  - Formalize
  - Identify significant associations using enrichment analysis over the fully inferred knowledge base

Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics. Bioinformatics. 2012.

Have you heard of OWL?

RDF triples are underspecified bits of knowledge. OWL can help you nail down what was really meant

Natural language statements:
The nucleus is a key part of the cell.



RDF triple:
<nucleus> <part-of> <cell>

OWL

- Nucleus and Cell are classes
- part-of is a relation between 2 instances
- Formalization: every instance of Nucleus ‘is part of’ at least one instance of Cell

OWL axiom:
Nucleus subClassOf part-of some Cell

assigning meaning to triples: domain expertise + logics required!

Convert RDF triples into OWL axioms.

Triple in RDF:
<C1 R C2>

- C1 and C2 are classes, R a relation between 2 classes
- intended meaning:
  - C1 SubClassOf: C2
  - C1 SubClassOf: R some C2
  - C1 SubClassOf: R only C2
  - C2 SubClassOf: R some C1
  - C1 SubClassOf: S some C2
  - C1 SubClassOf: R some (S only C2)
  - C1 DisjointFrom: C2
  - C1 and C2 SubClassOf: owl:Nothing
  - R some C1 DisjointFrom: R some C2
  - C1 EquivalentClasses C2
  - ...
- in general: P(C1, C2), where P is an OWL axiom (template)

Challenge: Formalizing data requires one to commit to a particular meaning – to make an ontological commitment

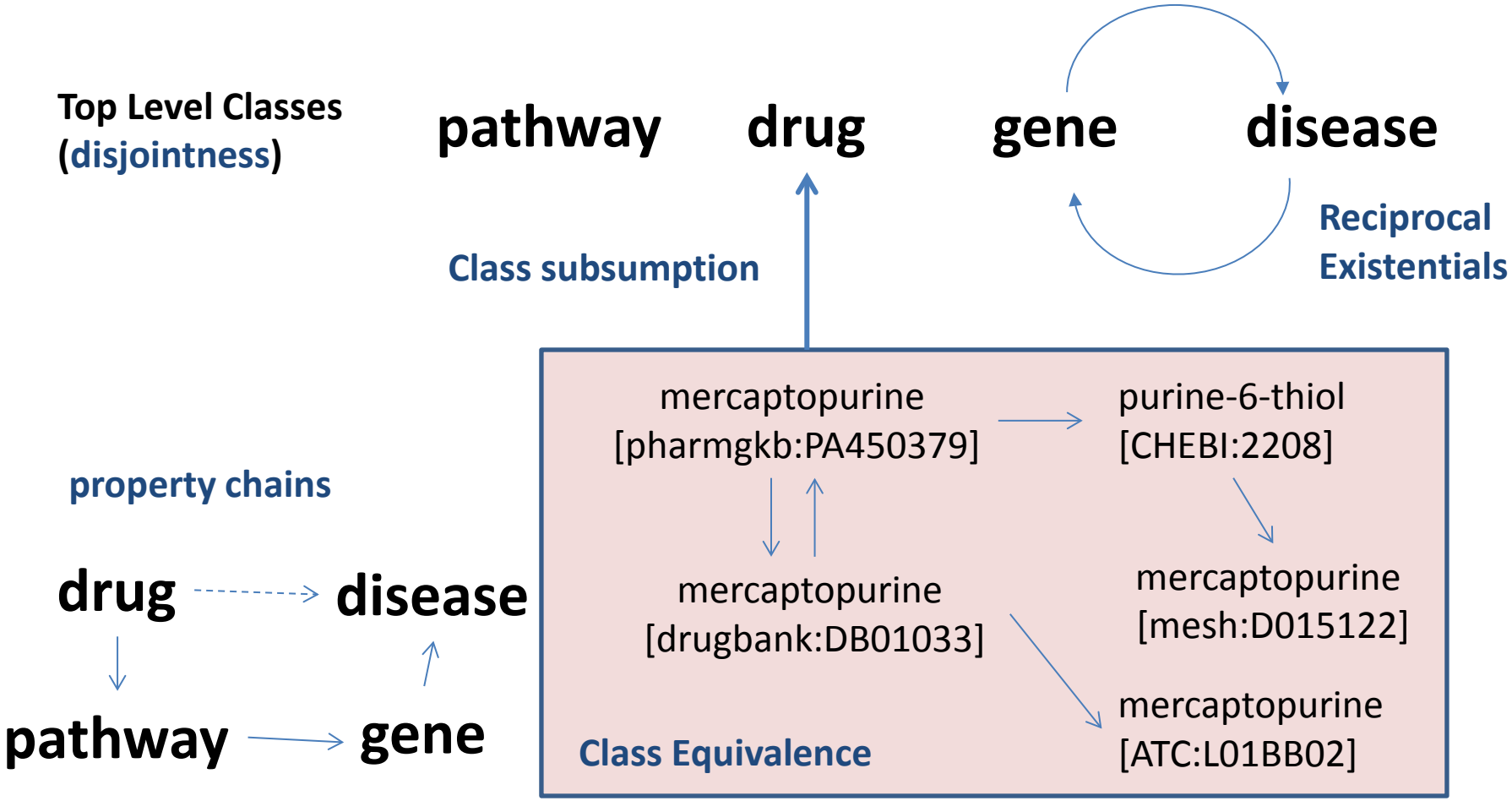
Axiom Patterns for Triples

<nucleus> <part-of> <cell>

?X part-of ?Y

translated to axiom pattern
?X subClassOf: part-of some ?Y

-> Nucleus subClassOf: part-of some Cell



Formalized as an OWL-EL ontology

650,000+ classes, 3.2M subClassOf axioms, 75,000 equivalentClass axioms

Benefits: Enhanced Query Capability

- Use any mapped terminology to query a target resource.
- Use knowledge in target ontologies to formulate more precise questions
  - ask for drugs that are associated with diseases of the joint: 'Chikungunya' (do:0050012) is defined as a viral infectious disease located in the 'joint' (fma:7490) and caused by a 'Chikungunya virus' (taxon:37124).
- Learn relationships that are inferred by automated reasoning.
  - alcohol (ChEBI:30879) is associated with alcoholism (PA443309) since alcoholism is directly associated with ethanol (CHEBI:16236)
  - 'parasitic infectious disease' (do:0001398) retrieves 129 *disease associated* drugs, 15 more than are directly associated.

Knowledge Discovery through Data Integration *and* Enrichment Analysis

- **OntoFunc**: Tool to discover significant associations between sets of objects and ontology categories. enrichment of attribute among a selected set of input items as compared to a reference set. hypergeometric or the binomial distribution, Fisher's exact test, or a chi-square test.

- We found 22,653 disease-pathway associations, where for each pathway we find genes that are linked to disease.
  - **Mood disorder** (do:3324) associated with **Zidovudine Pathway** (pharmgkb:PA165859361). Zidovudine is used to treat HIV/AIDS. Side effects include fatigue, headache, myalgia, malaise and anorexia

- We found 13,826 pathway-chemical associations
  - **Clopidogrel** (chebi:37941) associated with **Endothelin signaling pathway** (pharmgkb:PA164728163). Clopidogrel inhibits platelet aggregation and prolongs bleeding time. Endothelins are proteins that constrict blood vessels and raise blood pressure.

Tactical Formalization + Automated Reasoning Offers Compelling Value *for Certain Problems*

We need to be smart about the goal, and how best to achieve it. Tactical formalization is another tool in the toolbox.

We've formalized data as OWL ontologies:

- To identify mistakes in human curated knowledge
- To identify conflicting meaning in terms
- To identify mistakes in the representation of RDF data
  - incorrect use of relations
  - incorrect assertion of identity (owl:sameAs)
- To verify, fix and exploit Linked Data through expressive OWL reasoning
- To generate/infer new triples to write back into RDF and use for efficient retrieval

Many other applications can be envisioned.

Acknowledgements

Bio2RDF Release 2:

Allison Callahan, Jose Cruz-Toledo, Peter Ansell

Aberrant Pathways: Robert Hoehndorf, Georgios Gkoutos

dumontierlab.com

michel.dumontier@stanford.edu

Website: http://dumontierlab.com

Presentations: http://slideshare.com/micheldumontier