

A discovery platform for translational research

Usage Tutorial V4.0

Núria Queralt Rosinach

Integrative Biomedical Informatics Group (IBI)

Research Programme on Biomedical Informatics (GRIB)

Hospital del Mar Research Institute (IMIM)

Pompeu Fabra University (UPF)

Barcelona

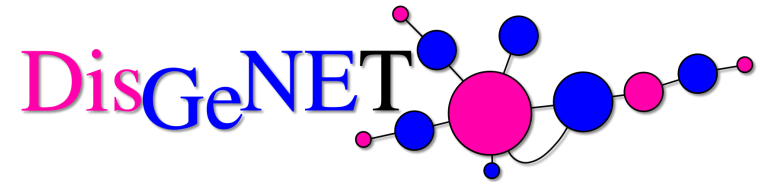


RESEARCH
PROGRAMME
ON BIOMEDICAL
INFORMATICS



Universitat
Pompeu Fabra
Barcelona

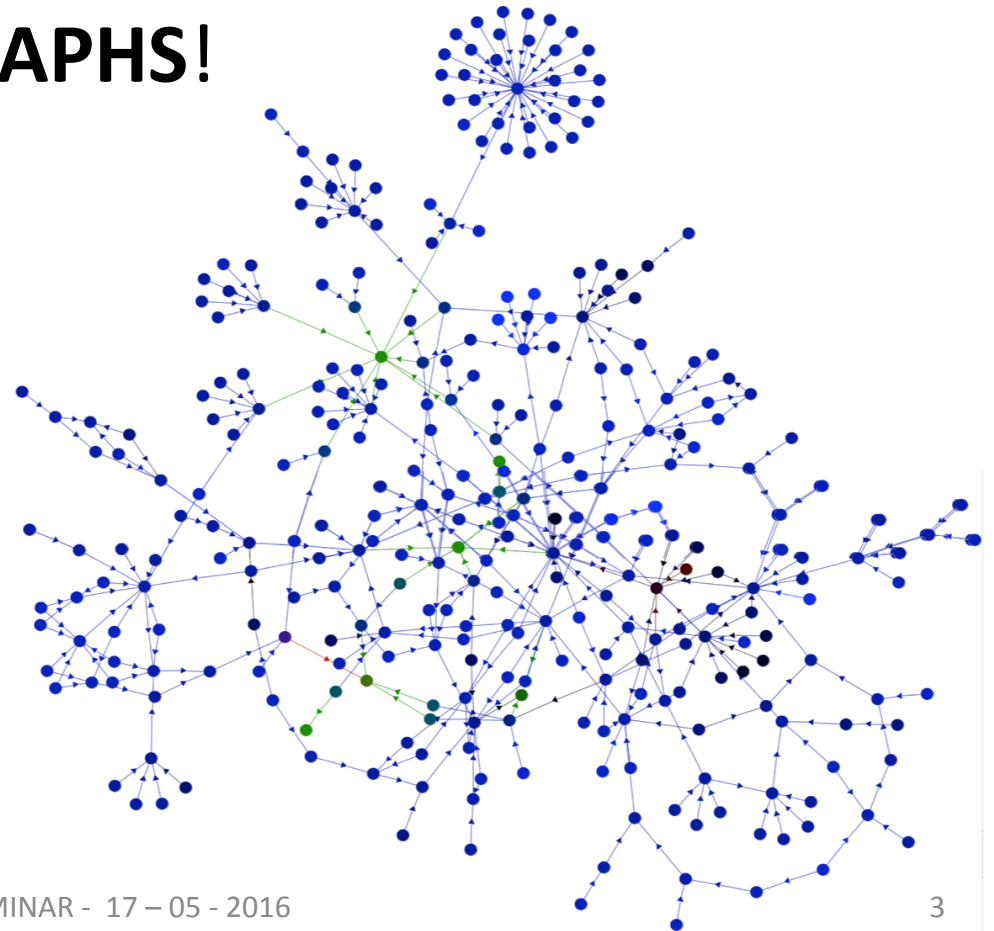
Outline



- Introduction to SPARQL
- DisGeNET Linked Open Data
 - Introduction
 - RDF-LD Description: Data Model, VoID, Interlinking
 - Implementation
 - Accessibility
 - Documentation
 - Use Cases
 - Querying the DisGeNET-RDF
 - Hands-on

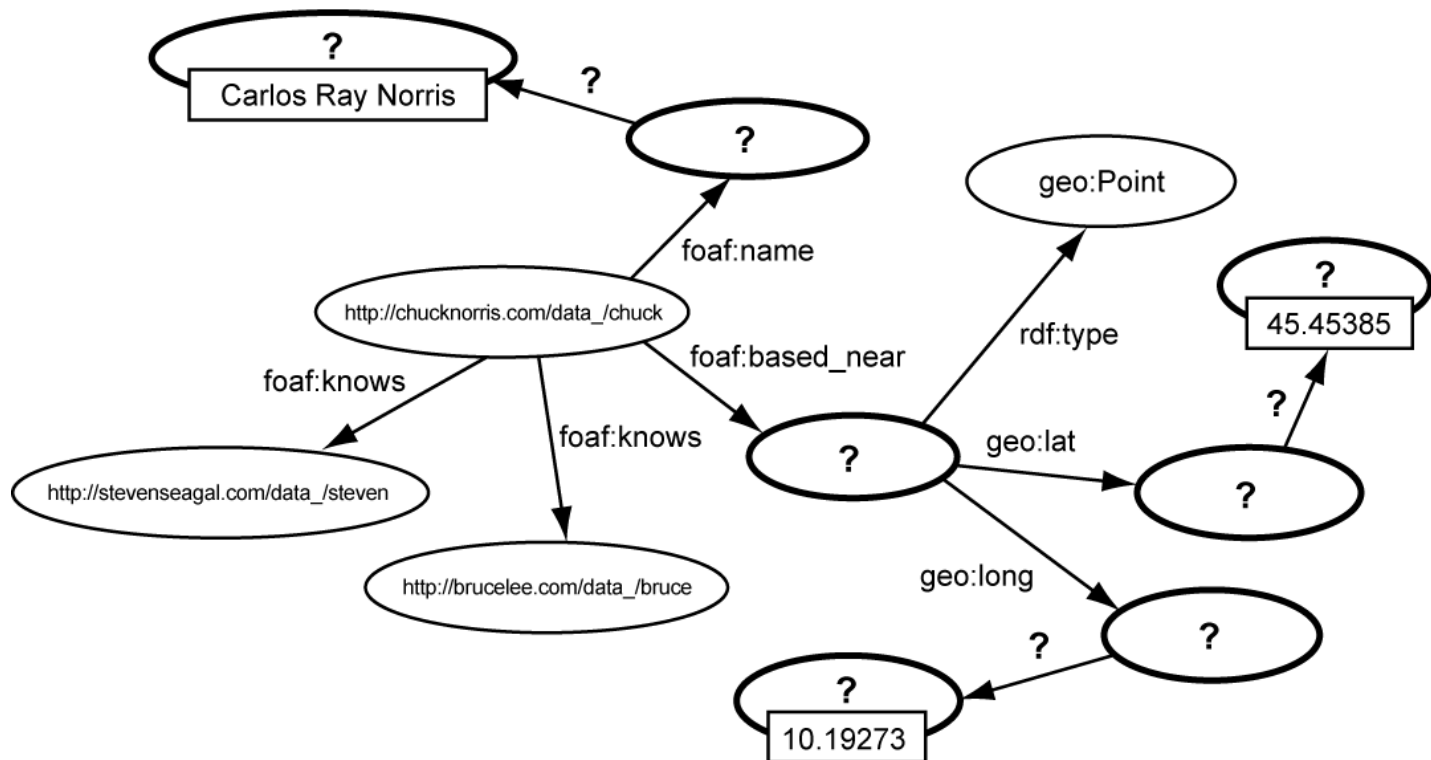
SPARQL

- SPARQL is the protocol/language to query **RDF**
- RDF: thinking in **GRAPHS!**



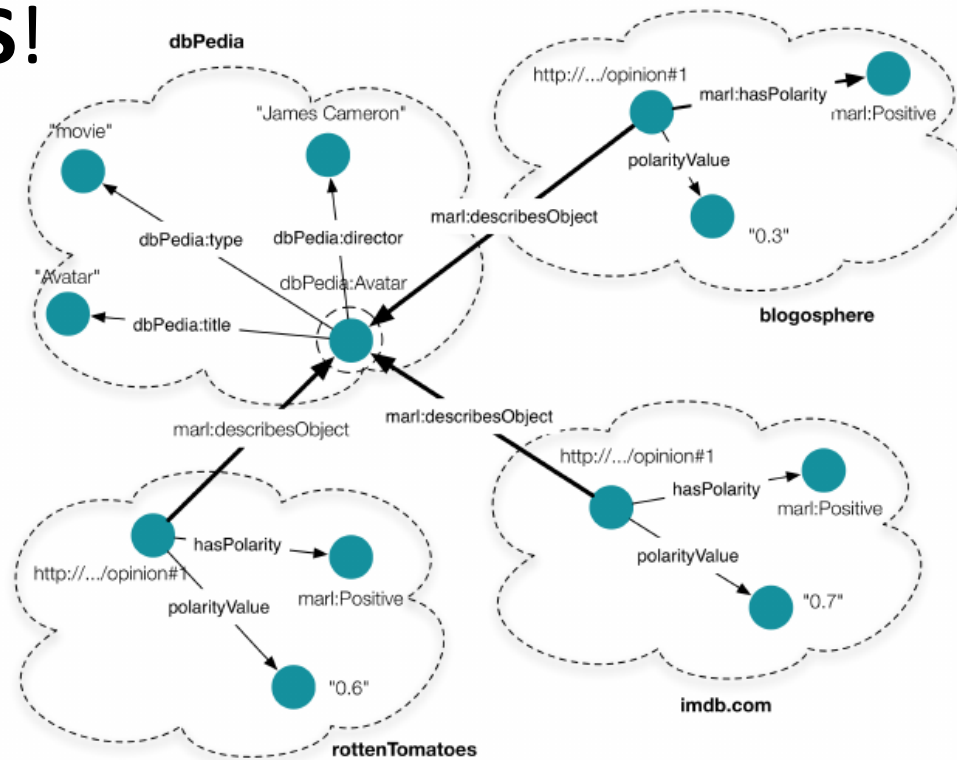
SPARQL

- SPARQL is the protocol/language to query **RDF**
- RDF: thinking in **Queryable GRAPHS!**



SPARQL

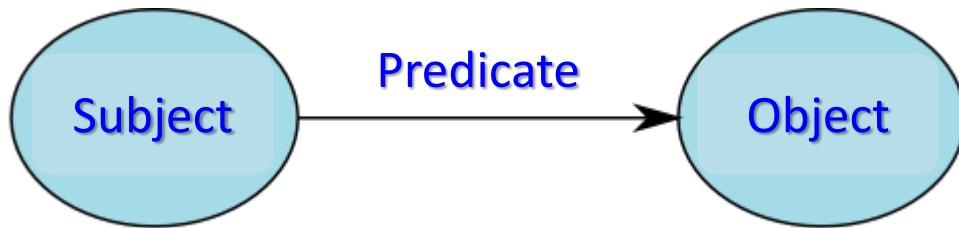
- SPARQL is the protocol/language to query **RDF**
- RDF: thinking in **Queryable** and **Disparate** **GRAPHS!**



**SPARQL 1.1
Federated query**

SPARQL

- Triple: **Subject, Predicate, Object**



SPARQL

- Triple: **<Subject> <Predicate> <Object>**
- **URIs** (can be abbreviated as *prefixed* names)

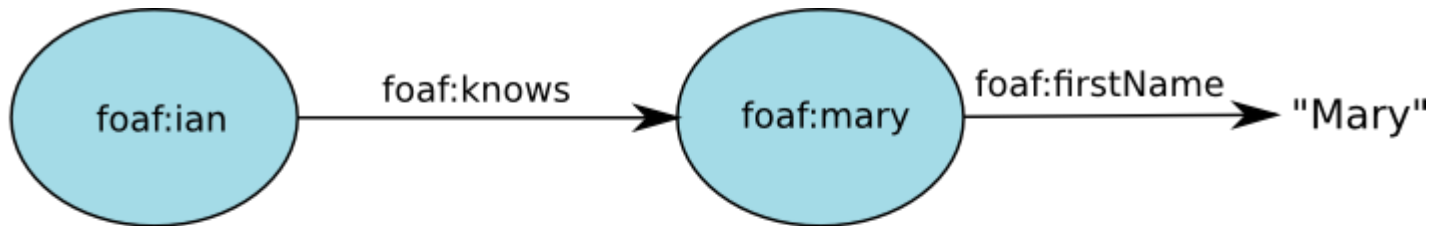
URI= **<http://xmlns.com/foaf/0.1/ian>**

prefix= **foaf:ian**



SPARQL

- Triple: **Subject, Predicate, Object**
- **URIs** (can be abbreviated as *prefixed* names)
- Objects: can be **literals**: strings, integers, booleans



```
foaf:ian foaf:knows foaf:mary.  
foaf:mary foaf:firstName "Mary".
```


SPARQL Query Structure

prefix declarations

PREFIX foaf:<http://xmlns.com/foaf/0.1/>

dataset definition

FROM <DATASET GRAPH>

result clause

SELECT /CONSTRUCT/ASK/DESCRIBE ..OUTPUT..

query pattern

WHERE { graph pattern }

query modifiers

ORDER BY ...

SPARQL Query Example

- **Data:** <http://rdf.disgenet.org/download/v4.0.0/DisGeNET-RDF-Example.ttl>
<<http://identifiers.org/pubmed/7919432>>
a ncit:C47902 ;
rdfs:comment "Scientific Article [pubmed:7919432] from MEDLINE/PubMed, which is a database of the U.S. National Library of Medicine, that gives evidence for at least one gene-disease association in DisGeNET. Articles are identified by the PubMed ID."@en ;
rdfs:label "7919432 [pubmed:7919432]"@en ;
dcterms:identifier "pubmed:7919432"^^xsd:string ;
dcterms:issued "1994"^^xsd:gYear ;
dcterms:title "7919432"@en ;
void:inDataset <<http://rdf.disgenet.org/v4.0.0/void/pubmed>> ;
skos:exactMatch <http://rdf.ncbi.nlm.nih.gov/pubchem/reference/PMID7919432> .
- Query
- Result

SPARQL Query Example

- Data
- Query:

PREFIX rdf: <<http://www.w3.org/1999/02/22-rdf-syntax-ns#>>

PREFIX dcterms: <<http://purl.org/dc/terms/>>

PREFIX ncit: <<http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>>

```
SELECT DISTINCT ?pmid ?year
```

```
WHERE {
```

```
  ?pmid rdf:type ncit:C47902 .
```

```
  ?pmid dcterms:issued ?year
```

```
}
```

```
LIMIT 10
```

- Result

SPARQL Query Example

- Data
- Query
- Result:

<http://rdf.disgenet.org/lodestar/sparql>


The word "FIGHT!" is written in a bold, stylized, italicized font. The letters are orange with a black outline and a white highlight, giving it a 3D, metallic appearance. The exclamation point is also stylized and integrated into the overall design.



[Home](#) [About](#) [Search](#) [Browser](#) [Downloads](#) [Cytoscape](#) [RDF](#) [Help](#)

DisGeNET Linked Open Data

DisGeNET as Linked Open Data

- **RDF and trusty nanopublications**
 - URIs: RDF providers or 
 - SIO
 - Use of standards (**11 ontologies** in NCBO)



• Metadata description ( HCLS)

• Interlinking



• Access

• Download Data Dump

• SPARQL Endpoint

• Faceted Browser

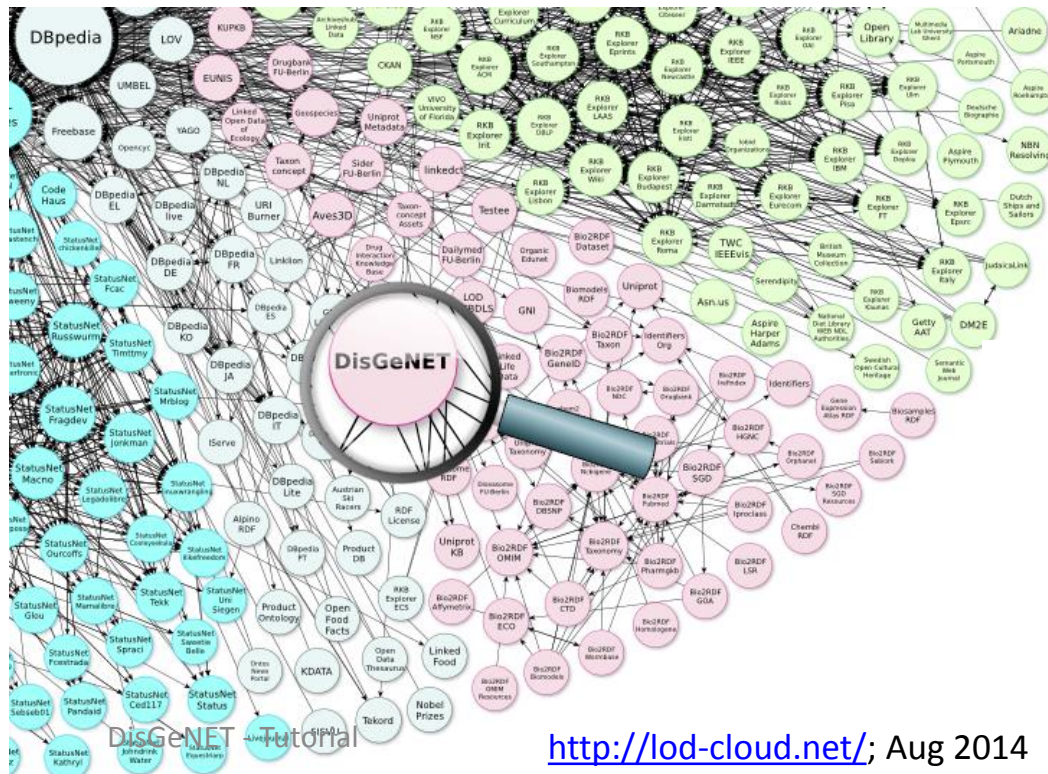


• Nanopublication Network

• Open license

• Datahub

• Software



<http://lod-cloud.net/>; Aug 2014

DisGeNET-RDF

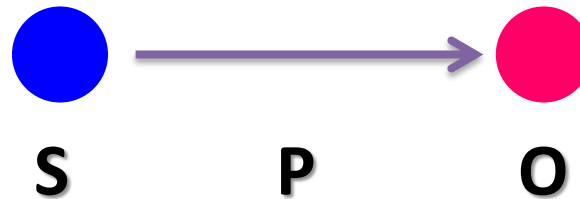


Data Model

- How to describe an **association**?

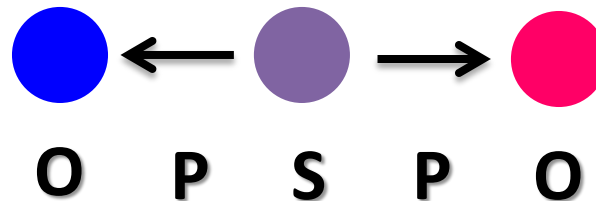
a) As a **property**

Gene associated **Disease**



b) As a **class**

Gene Association **Disease**



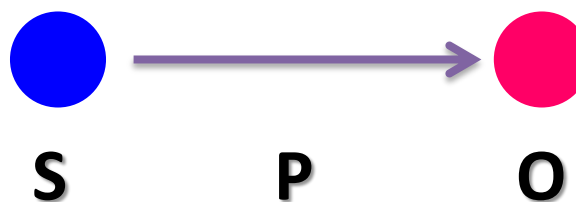


Data Model

- How to describe an **association**?

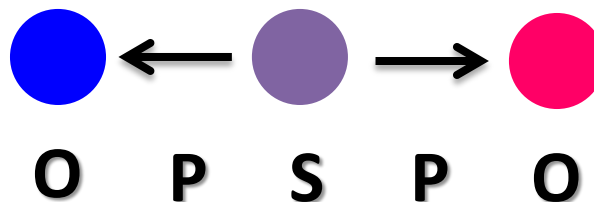
Gene associated Disease

a) As a **property**



b) As a **class**

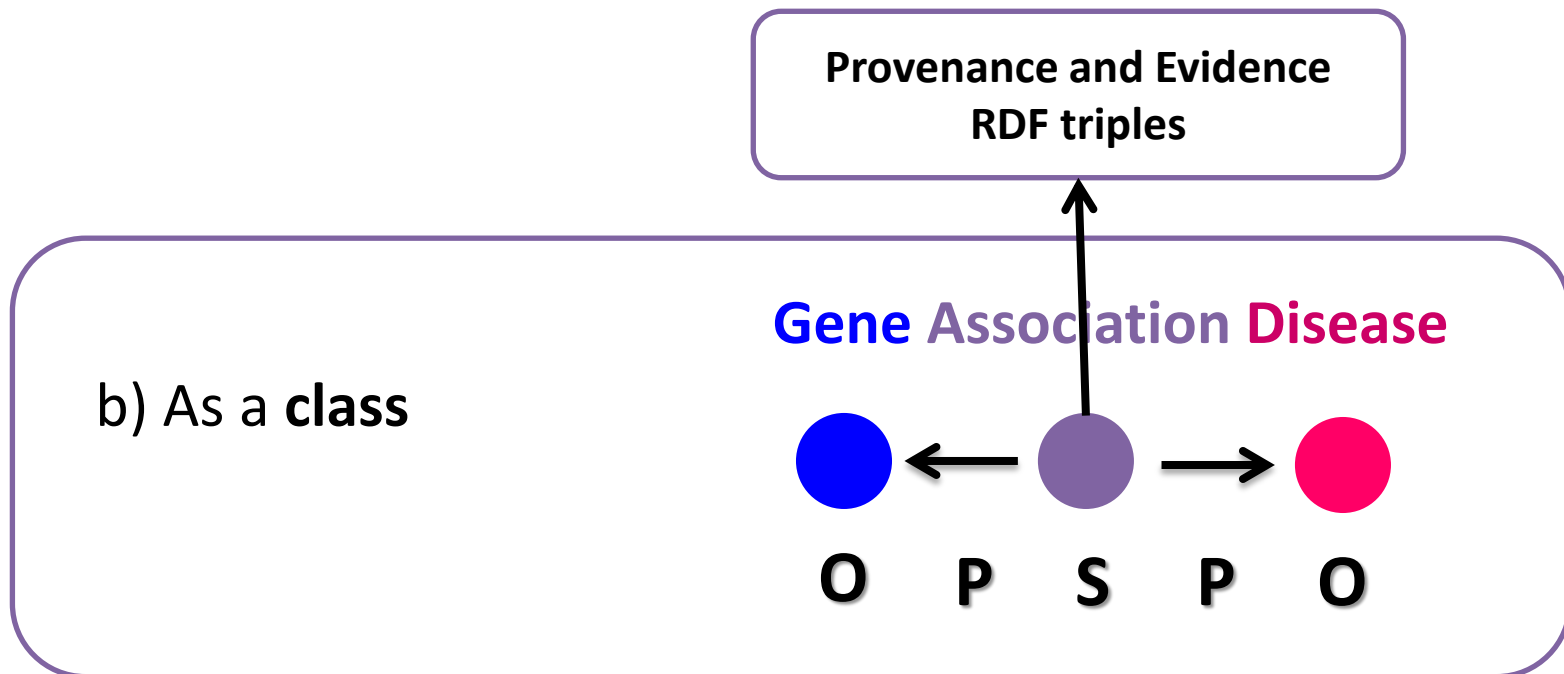
Gene Association Disease





Data Model

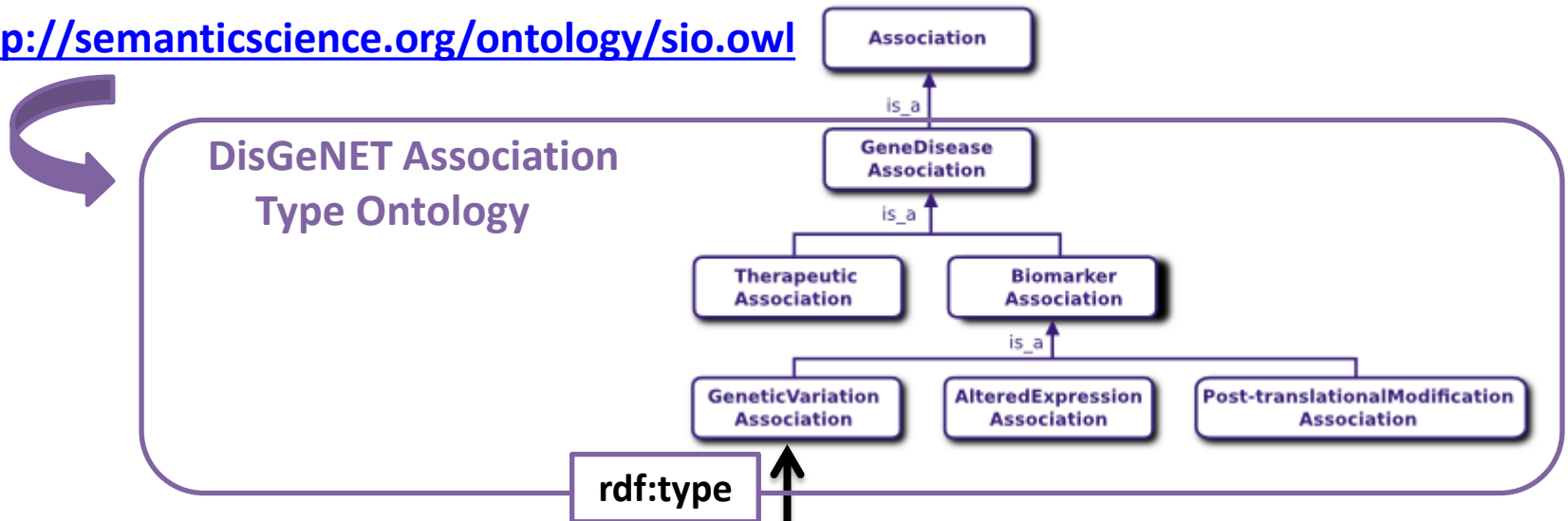
- How to describe an **association**?



Data Model

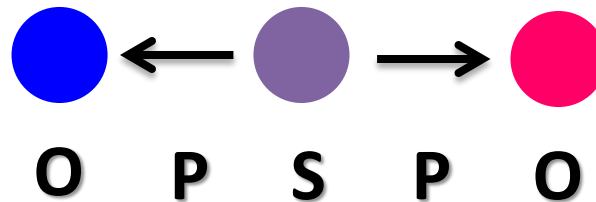
- Ontology-based integration

<http://semanticscience.org/ontology/sio.owl>



Gene Association Disease

- DisGeNET Standards
 - Shared IDs
 - Standard ontologies



Data Model

- Semantic Annotation: **Standard ontologies**

Prefix	Namespace	Vocabularies
ncit	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#	NCI Thesaurus
sio	http://semanticscience.org/resource/	SIO
up	http://purl.uniprot.org/core/	UniProt
void	http://rdfs.org/ns/void#	VOID
foaf	http://xmlns.com/foaf/0.1/	FOAF Vocabulary
dcterms	http://purl.org/dc/terms/	DCMI Terms
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	RDF
rdfs	http://www.w3.org/2000/01/rdf-schema#	RDF Schema
xsd	http://www.w3.org/2001/XMLSchema#	XML Schema
owl	http://www.w3.org/2002/07/owl#	OWL
skos	http://www.w3.org/2004/02/skos/core#	SKOS

Data Model

- Semantic Annotation: **Standard ontologies**

Prefix	Namespace	Vocabularies
ncit	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#	NCI Thesaurus
sio	http://semanticscience.org/resource/	SIO
up	http://purl.uniprot.org/core/	UniProt
void	http://rdfs.org/ns/void#	VOID
foaf	http://xmlns.com/foaf/0.1/	FOAF Vocabulary
dcterms	http://purl.org/dc/terms/	DCMI Terms
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	RDF
rdfs	http://www.w3.org/2000/01/rdf-schema#	RDF Schema
xsd	http://www.w3.org/2001/XMLSchema#	XML Schema
owl	http://www.w3.org/2002/07/owl#	OWL
skos	http://www.w3.org/2004/02/skos/core#	SKOS

RDF Structure

Data Model

- Semantic Annotation: **Standard ontologies**

Prefix	Namespace	Vocabulary	
ncit	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#	NCI Thesaurus	Biomedical entities
sio	http://semanticscience.org/resource/	SIO	
up	http://purl.uniprot.org/core/	UniProt	Relationships
void	http://rdfs.org/ns/void#	VOID	
foaf	http://xmlns.com/foaf/0.1/	FOAF Vocabulary	
dcterms	http://purl.org/dc/terms/	DCMI Terms	
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	RDF	
rdfs	http://www.w3.org/2000/01/rdf-schema#	RDF Schema	
xsd	http://www.w3.org/2001/XMLSchema#	XML Schema	
owl	http://www.w3.org/2002/07/owl#	OWL	
skos	http://www.w3.org/2004/02/skos/core#	SKOS	RDF Structure

Data Model

- Semantic Annotation: **Standard ontologies**

Prefix	Namespace	Vocabulary	
ncit	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#	NCI Thesaurus	Biomedical entities
sio	http://semanticscience.org/resource/	SIO	
up	http://purl.uniprot.org/core/	UniProt	Relationships
void	http://rdfs.org/ns/void#	VOID	
foaf	http://xmlns.com/foaf/0.1/	FOAF Vocabulary	
dcterms	http://purl.org/dc/terms/	DCMI Terms	Metadata
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	RDF	
rdfs	http://www.w3.org/2000/01/rdf-schema#	RDF Schema	
xsd	http://www.w3.org/2001/XMLSchema#	XML Schema	
owl	http://www.w3.org/2002/07/owl#	OWL	
skos	http://www.w3.org/2004/02/skos/core#	SKOS	RDF Structure

Data Model

- **URIs** in DisGeNET: **shared, cool & dereferenceable**

- ID Normalization

- DisGeNET URIs:

<http://rdf.disgenet.org/resource/entity/ID>

Unique
association
attributes

- Establish URIs from **primary data providers**

- *Identifiers.org*

<http://identifiers.org/data-collection-namespace/ID>

Data Model

- URIs in DisGeNET: **shared, cool & dereferenceable**
 - ID Normalization
 - **Gene-Disease Association::DisGeNET ID**

Entity	URI	Semantics
Gene-Disease Association	http://rdf.disgenet.org/resource/gda/DGNf5cb3969d75871f05a5d5f984f8dfc34	sio:SIO_001122
PubMed article	http://identifiers.org/pubmed/9837812	ncit:C47902
Source	http://rdf.disgenet.org/v3.0.0/void/uniprot-20150221	dctypes:Dataset, dcat:Distribution
Score	http://rdf.disgenet.org/resource/gda/ncbigene:4728_umls:C0023264_association_DisGeNET_Score	ncit:C25338
SNP	http://identifiers.org/dbsnp/rs28939679	ncit:C18279

Data Model

- URIs in DisGeNET: **shared, cool & dereferenceable**
 - ID Normalization
 - **Gene::NCBI Gene ID**

Entity	URI	Semantics
Gene	http://identifiers.org/ncbigene/4728	ncit:C16612
HGNC Gene Symbol	http://identifiers.org/hgnc.symbol/NDUFS8	ncit:C43568
Protein	http://identifiers.org/uniprot/O00217	ncit:C17021
Panther Class	http://rdf.disgenet.org/resource/panther.classification/PC00211	rdfs:Class
Pathway	http://identifiers.org/reactome/REACT_111217	ncit:C20633

Data Model

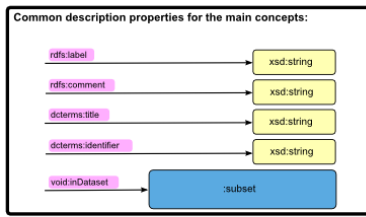
- URIs in DisGeNET: **shared, cool & dereferenceable**
 - ID Normalization
 - **Disease::UMLS Concept Unique Identifier (CUI)**

Entity	URI	Semantics
Disease	http://linkedlifedata.com/resource/umls/id/C0023264	ncit:C7057
MeSH Class	http://rdf.imim.es/rh-mesh.owl#C18	rdfs:Class
UMLS Semantic Type	http://biotop.googlecode.com/svn/trunk/umlssn.owl#T047	rdfs:Class
Phenotype	http://purl.obolibrary.org/obo/HP_0004633	sio:SIO_010056
Cross References	http://identifiers.org/vocab-namespace/ID	Human Disease Ontology, MeSH, OMIM, Orphanet, Decipher, NCIt, ICD9, Human Phenotype Ontology

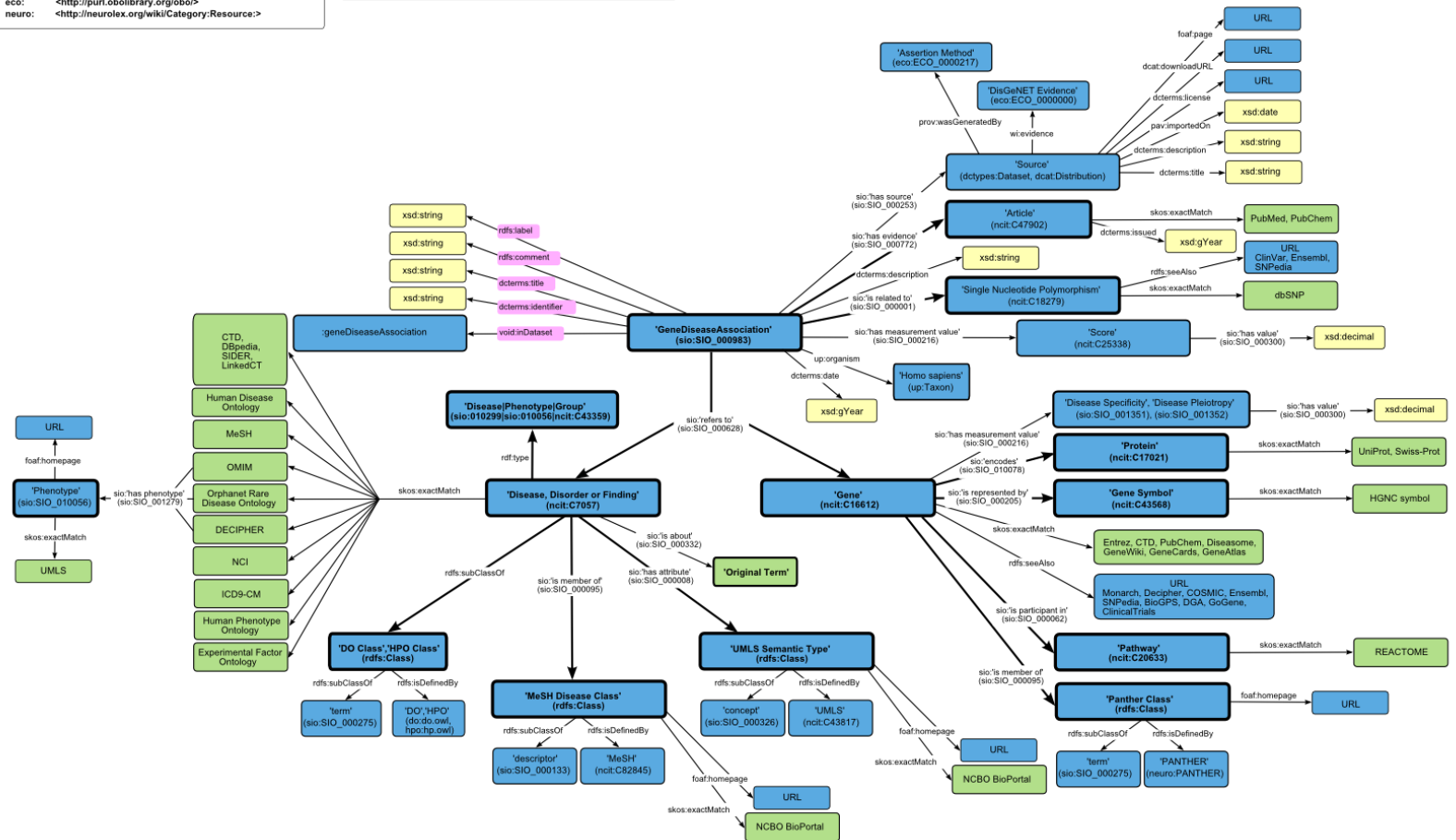
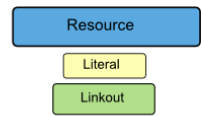
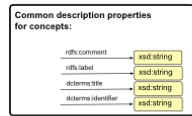
Data Model

Namespaces:

- rdflib: <http://rdflib.org/v2.1.0/void.ttl#>
- rdfs: <http://www.w3.org/2000/01/rdf-schema#>
- dcterms: <http://purl.org/dc/terms/>
- skos: <http://www.w3.org/2004/02/skos/core#>
- xsd: <http://www.w3.org/2001/XMLSchema#>
- ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
- sio: <http://semanticscience.org/resource/>
- up: <http://purl.uniprot.org/wp#>
- foaf: <http://xmlns.com/foaf/0.1/>
- void: <http://rdfs.org/ns/void#>
- dctypes: <http://purl.org/dc/dcmitype/>
- dc: <http://www.w3.org/ns/dc/atom#>
- pav: <http://purl.org/pav/2.0/>
- prov: <http://www.w3.org/ns/prov#>
- eco: <http://purl.obolibrary.org/obo/>
- neuro: <http://neurolex.org/wiki/Category:Resource/>



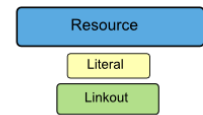
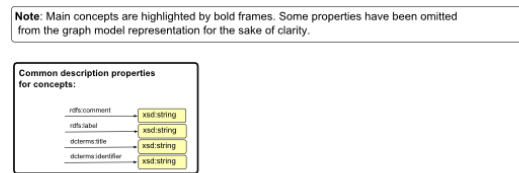
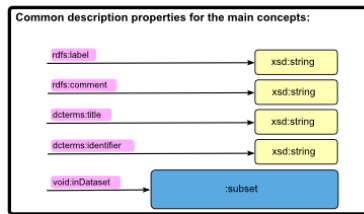
Note: Main concepts are highlighted by bold frames. Some properties have been omitted from the graph model representation for the sake of clarity.



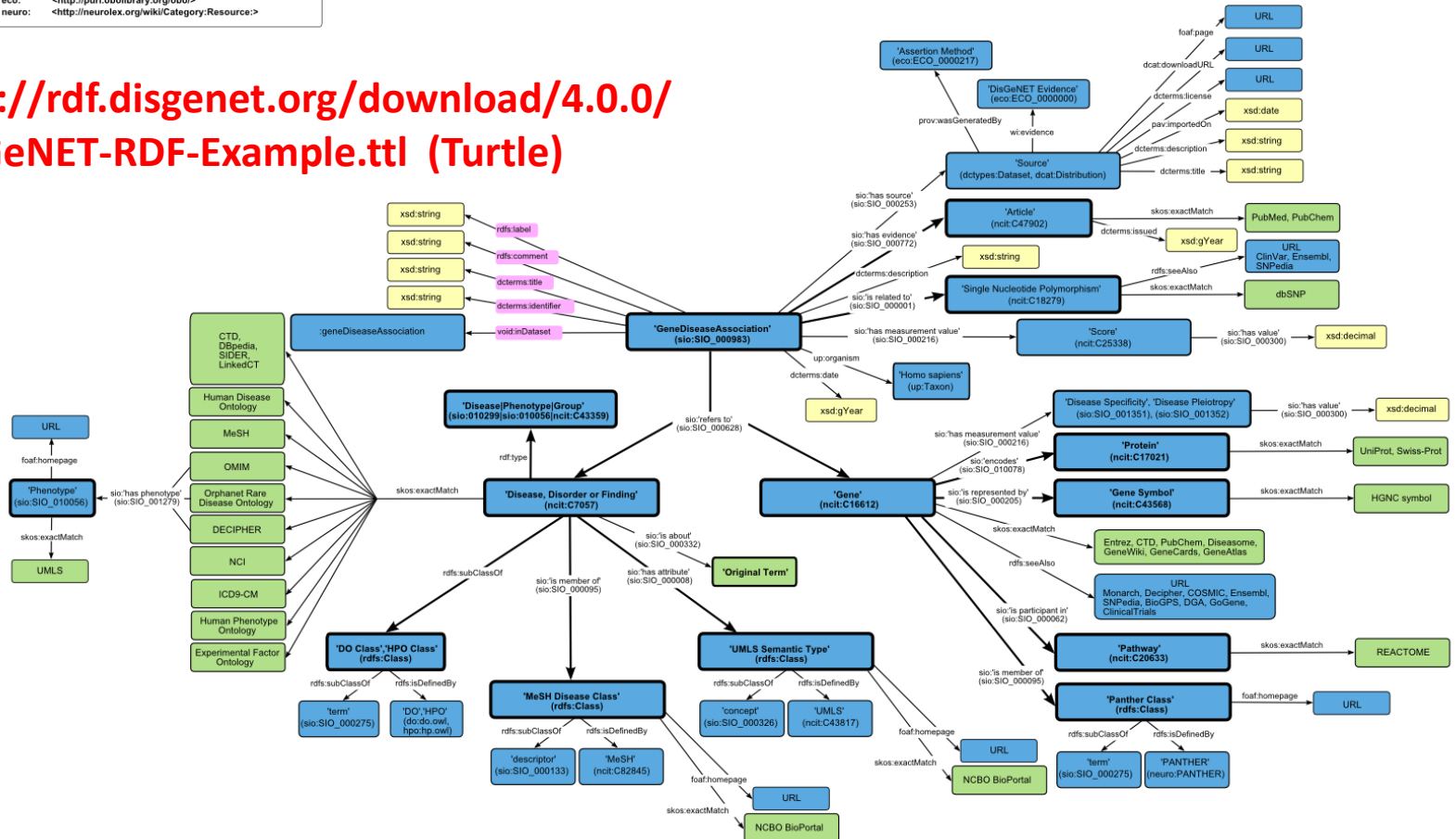
Data Model

Namespaces:

- rdflib: <http://rdflib.org/v2.1.0/void.ttl#>
- rdfs: <http://www.w3.org/2000/01/rdf-schema#>
- dcterms: <http://purl.org/dc/terms/>
- skos: <http://www.w3.org/2004/02/skos/core#>
- xsd: <http://www.w3.org/2001/XMLSchema#>
- ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
- sio: <http://semanticscience.org/resource/>
- up: <http://purl.uniprot.org/wp#>
- foaf: <http://xmlns.com/foaf/0.1/>
- void: <http://rdfs.org/void#>
- dctypes: <http://purl.org/dc/dcmitype/>
- dcats: <http://www.w3.org/ns/dcat#>
- pav: <http://purl.org/pav/2.0/>
- prov: <http://www.w3.org/ns/prov#>
- eco: <http://purl.obolibrary.org/obo/>
- neuro: <http://neurolex.org/wiki/Category:Resource/>



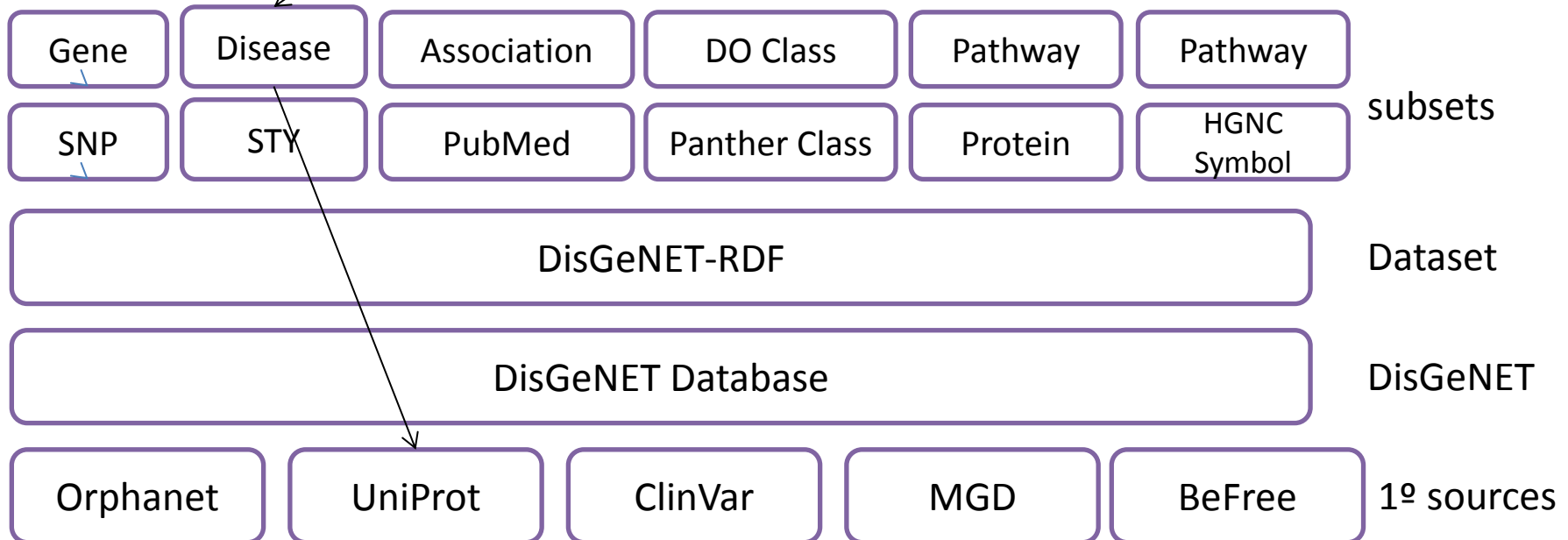
<http://rdf.disgenet.org/download/4.0.0/DisGeNET-RDF-Example.ttl> (Turtle)



Metada Dataset Description

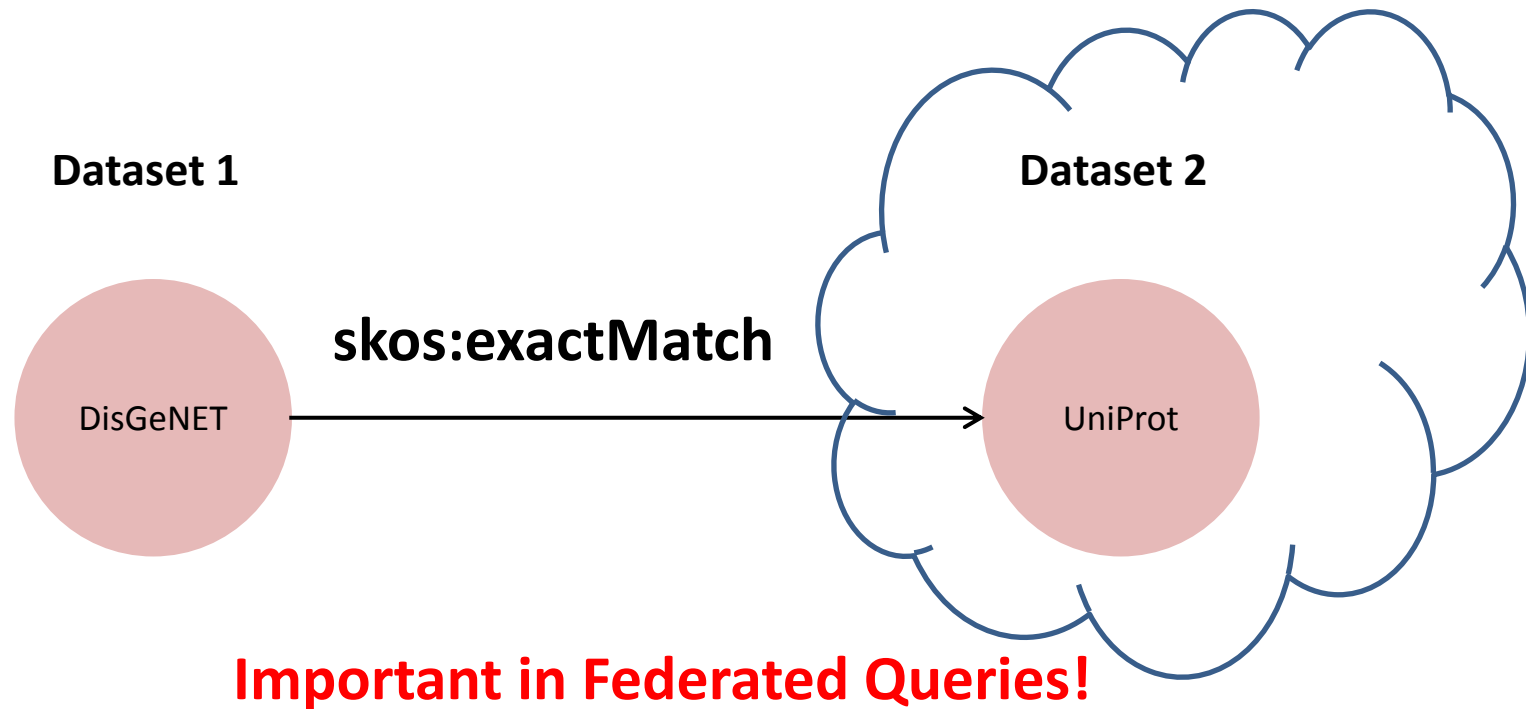
DisGeNET-RDF VoID file (Vocabulary of Interlinked Datasets)

```
<http://linkedlifedata.com/resource/umls/id/C0023264>  
  a          ncid:C7057 ;  
  rdfs:comment "Disease [umls:C0023264] associated with at least one gene in DisGeNET. Diseases are identified by the UMLS CUI." ;  
  rdfs:label "Leigh Disease [umls:C0023264]" ;  
  dct:identifier "umls:C0023264"^^xsd:string ;  
  dct:title "Leigh Disease" ;  
  void:inDataset <http://rdf.disgenet.org/v3.0.0/void/disease> ;  
  sio:SIO_000000 <http://biotop.googlecode.com/svn/trunk/umlsn.owl#T047> ;  
  sio:SIO_000095 <http://rdf.imim.es/rh-mesh.owl#C18> , <http://rdf.imim.es/rh-mesh.owl#C16> , <http://rdf.imim.es/rh-mesh.owl#C10>
```



Interlinking

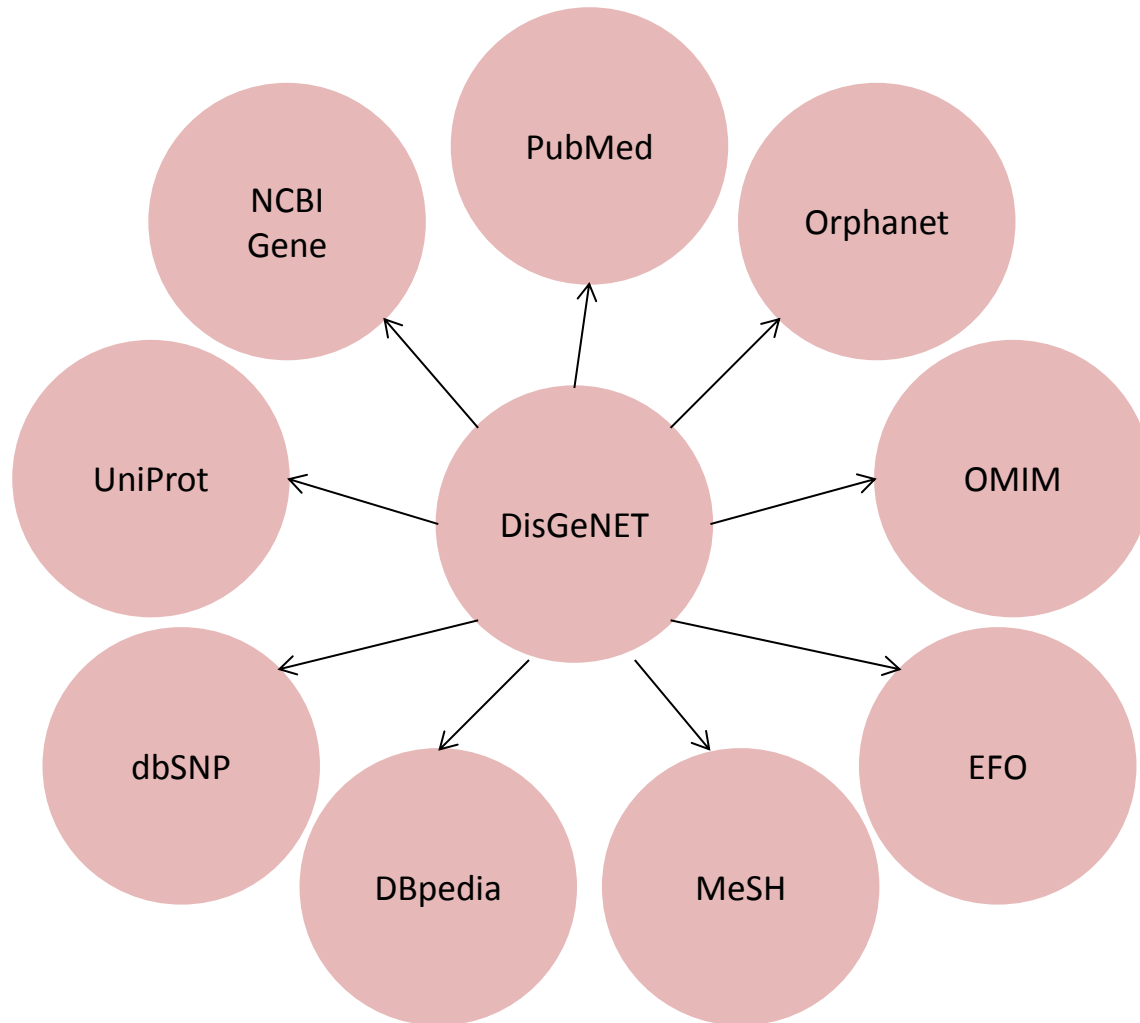
DisGeNET -- RDF link -> LOD cloud



Interlinking

?s skos:exactMatch ?o

**Biomedical
Databases
and
Disease
Terminologies**

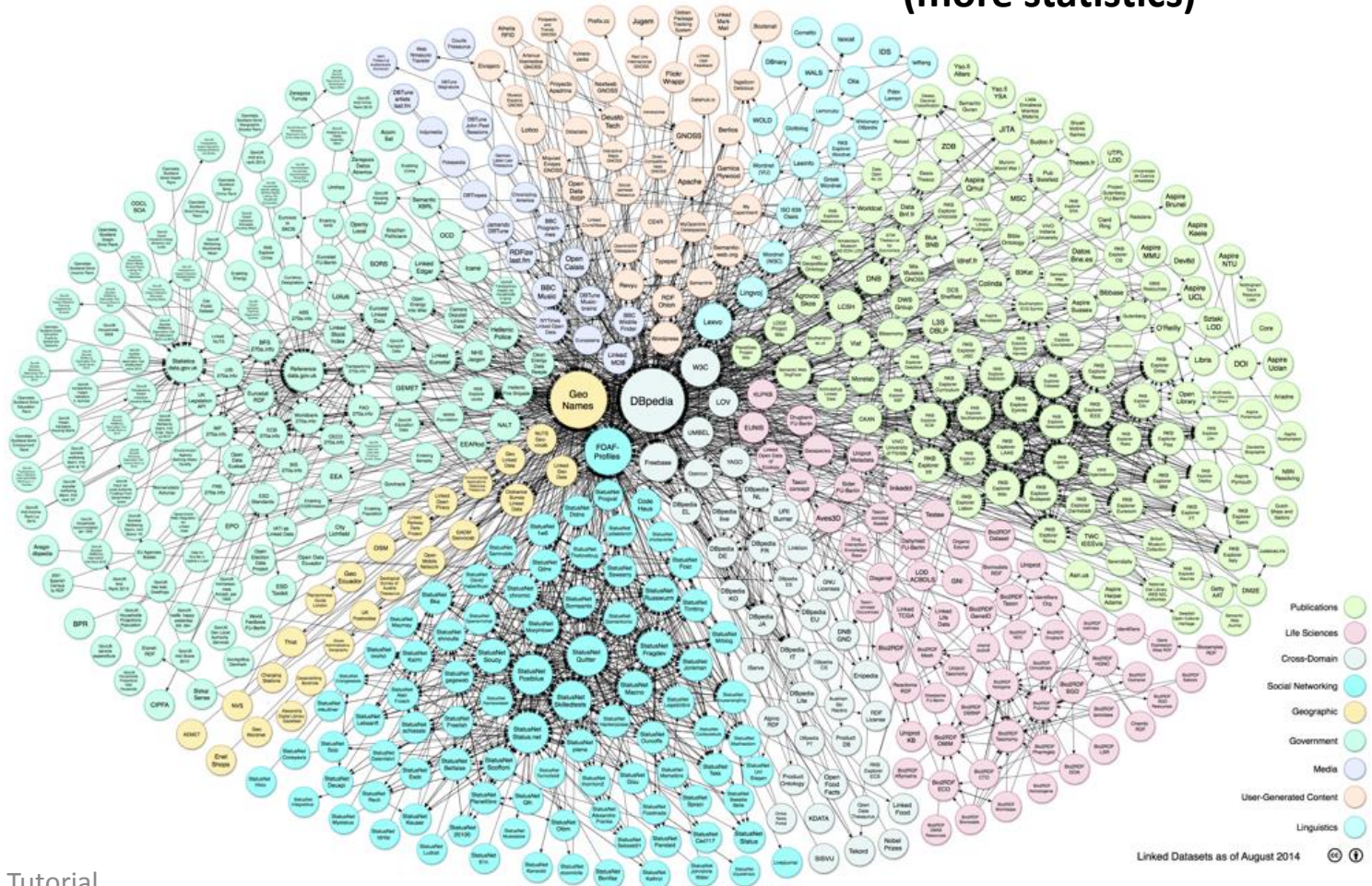


DisGeNET as Linked Open Data

- Interlinking: 4,962,315 RDF links to RDF datasets in the LOD

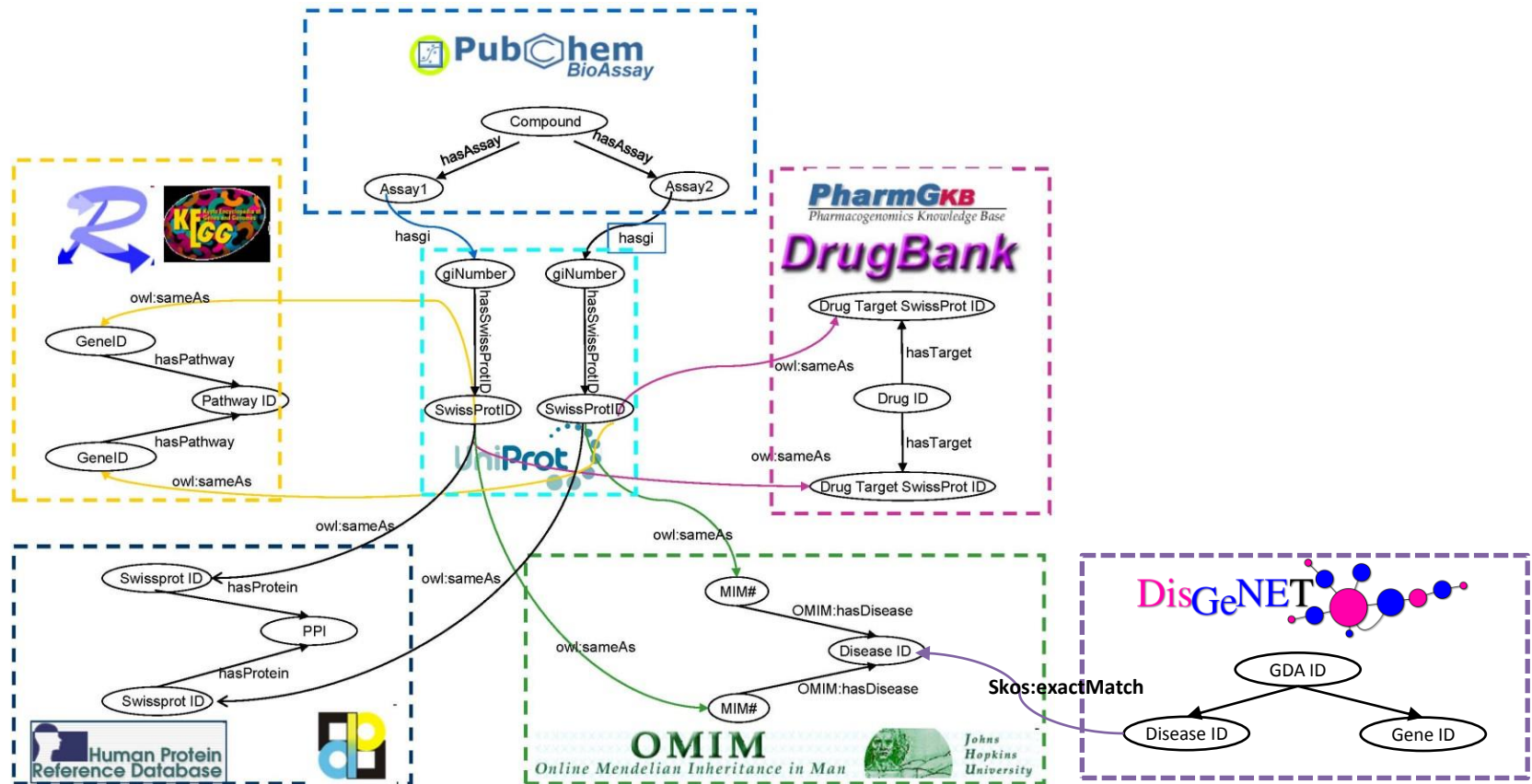


<https://datahub.io/dataset/disgenet>
(more statistics)



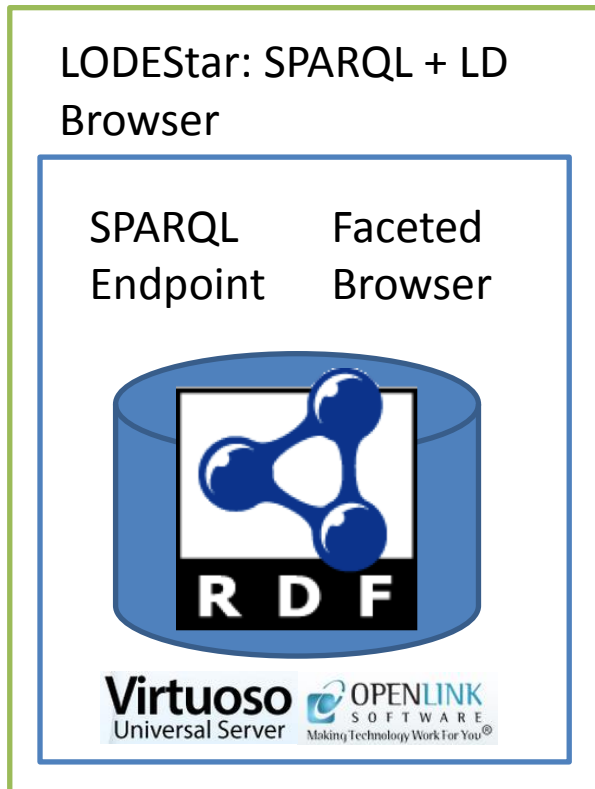
Federated Query Support

- *SPARQL 1.1*: SERVICE <sparql endpoint> {}



Implementation

- DisGeNET RDF data, VoID dataset description, and seven OWL ontologies loaded into the RDF Store
- Total number of triples: 30,506,021 (13G)



Hardware: 7.1.0 Usage Restrictions

- SPARQL:
 - only **SELECT, DESCRIBE, ASK, CONSTRUCT**
 - performance opt:
 - Max # of rows per result
 - Max query cost estimation time
 - Max query execution time

Security: basic setup

Accessibility

- Download: RDF dump + linksets
 - <http://rdf.disgenet.org/download/>
- Faceted Browser
 - <http://rdf.disgenet.org/fct/>
- SPARQL endpoint
 - <http://rdf.disgenet.org/sparql/>
- EBI::LODEStar SPARQL + Linked Data Browser
 - <http://rdf.disgenet.org/lodestar/sparql>
- Open PHACTS APIs
 - <https://dev.openphacts.org/docs/2.1>

Documentation

- Descriptions
- RDF Schema
- Points of access
- SPARQL query examples @:



<http://rdf.disgenet.org/>

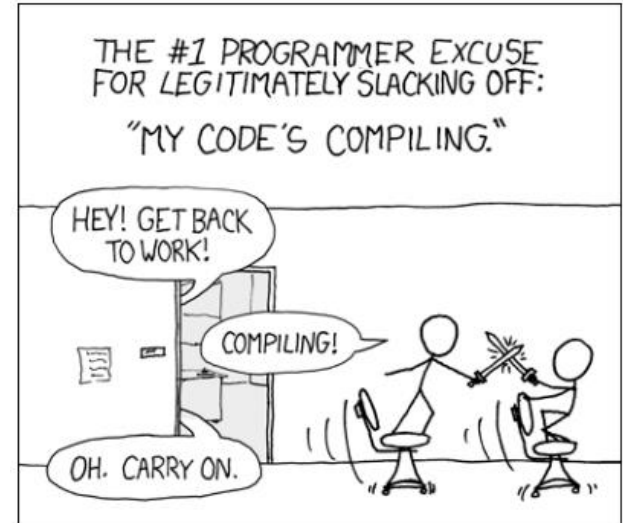
- Support @:

support@disgenet.org

Querying the DisGeNET-RDF

SPARQL QUERIES

- Not easy
- RDF Schema-aware
- Performance issues
 - Optimal queries: there is a trade off between the amount of time you spend analyzing and transforming the query and the performance gains of those transformations
 - Technology-dependant
 - crossing a lot of information decrease speed (making the system fails): better local



Querying DisGeNET

- SPARQL Queries over DisGeNET data

<http://rdf.disgenet.org/sparql/>

<http://rdf.disgenet.org/lodestar/sparql>

- Contains all DisGeNET data
- Free access
- SPARQL 1.1 Standard

Enter SPARQL Query

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX void: <http://rdfs.org/ns/void#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX nci: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
PREFIX up: <http://purl.uniprot.org/core/>

SELECT DISTINCT ?gda ?type ?label FROM <http://rdf.disgenet.org> WHERE {
  ?gda rdf:type ?type .
  ?type rdfs:subClassOf+ sio:SIO_000983 .
  ?type rdfs:label ?label
}
LIMIT 50
```

Results per page:

Output:

Example Queries

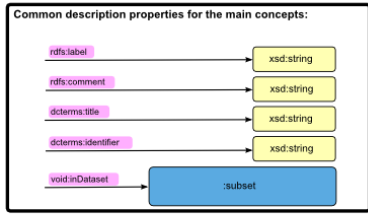
- Query 1
Get all gene-disease associations integrated in DisGeNET
- Query 2
Get all gene-disease associations integrated in DisGeNET searching by the 'Ovarian cancer' class in DO (DOID:2394)
- Query 3
Get all diseases in DisGeNET searching by 'Familial prostate cancer' class in ORDO (Orphanet_1331).

Generated by the [LODEStar](#) linked data browser from the Functional Genomics Production Team (FGPT)

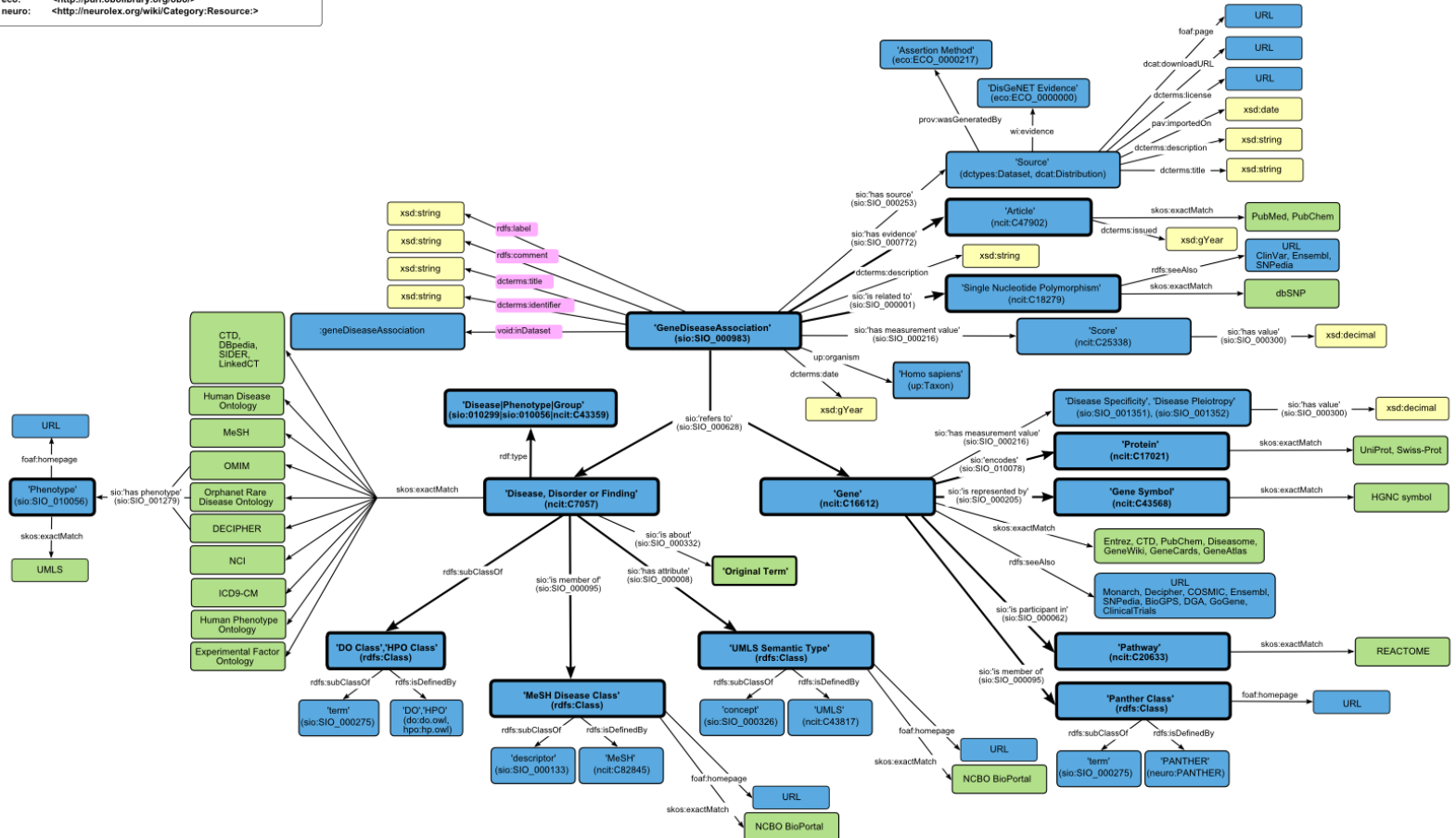
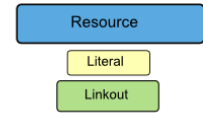
Data Model

Namespaces:

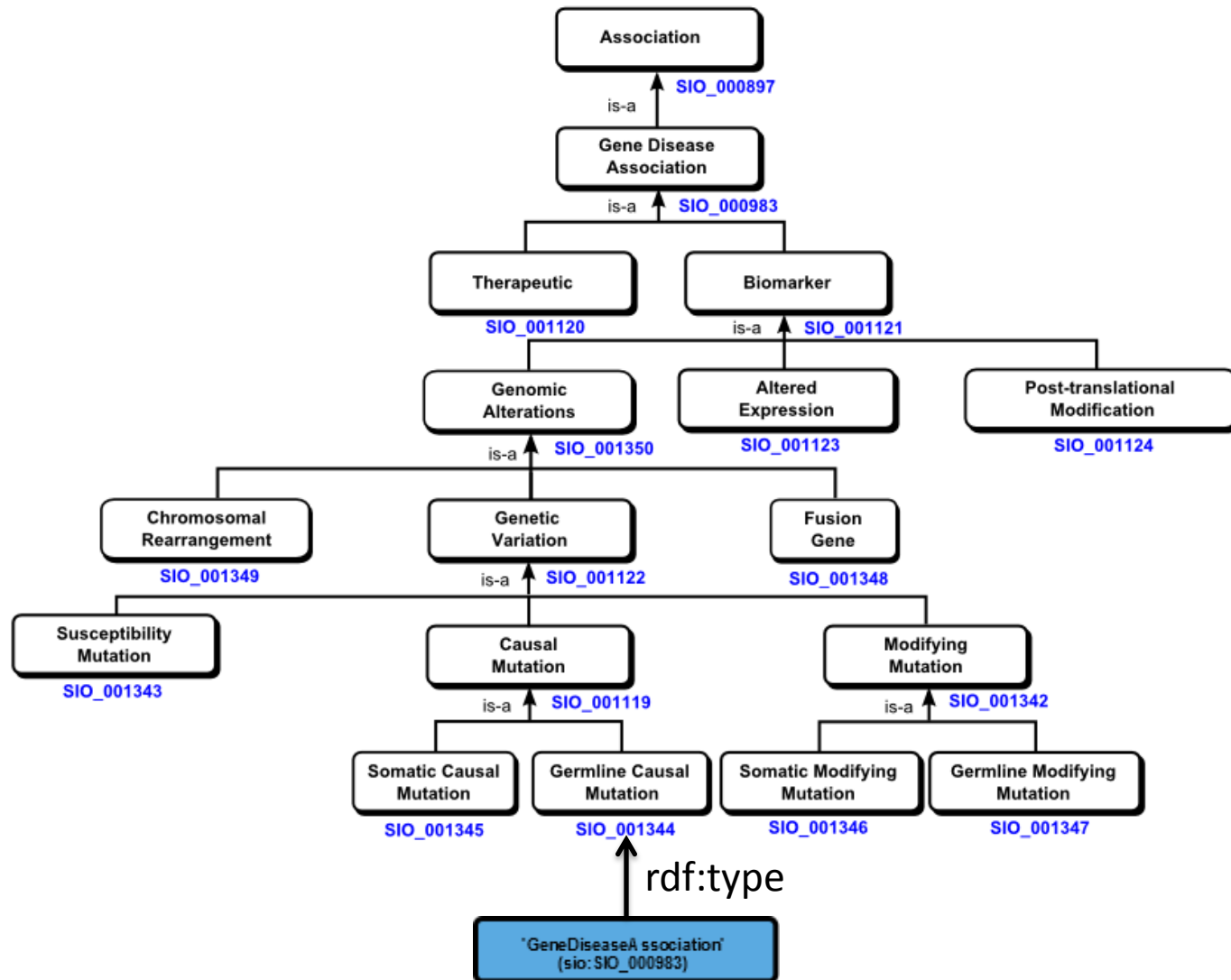
- rdflib: <http://rdflib.org/v2.1.0/void.ttl#>
- rdfs: <http://www.w3.org/2000/01/rdf-schema#>
- dcterms: <http://purl.org/dc/terms/>
- skos: <http://www.w3.org/2004/02/skos/core#>
- xsd: <http://www.w3.org/2001/XMLSchema#>
- ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
- sio: <http://semanticscience.org/resource/>
- up: <http://purl.uniprot.org/up/#>
- foaf: <http://xmlns.com/foaf/0.1/>
- void: <http://rdfs.org/ns/void#>
- dctypes: <http://purl.org/dc/dcmitype/>
- dc: <http://www.w3.org/ns/dc/atom#>
- pav: <http://purl.org/pav/2.0/>
- prov: <http://www.w3.org/ns/prov#>
- eco: <http://purl.obolibrary.org/obo/>
- neuro: <http://neurolex.org/wiki/Category:Resource/>



Note: Main concepts are highlighted by bold frames. Some properties have been omitted from the graph model representation for the sake of clarity.



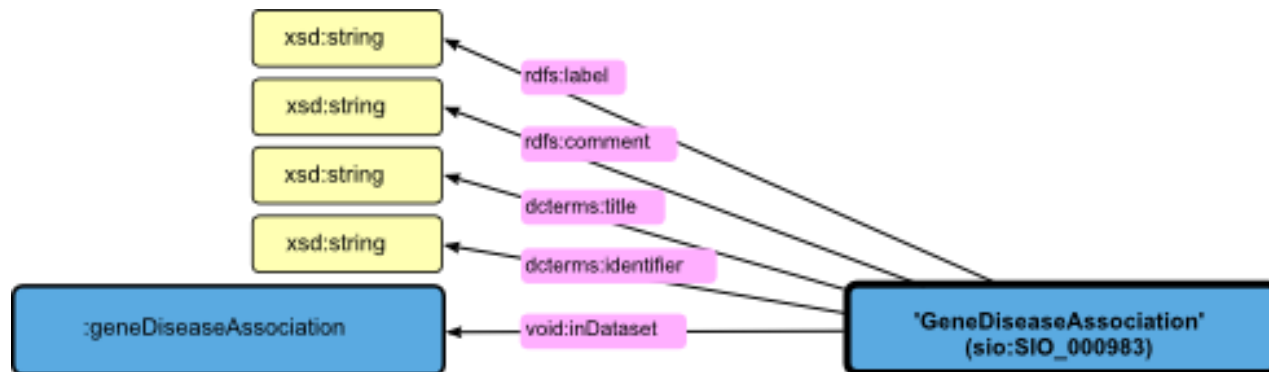
Data Model



Querying DisGeNET

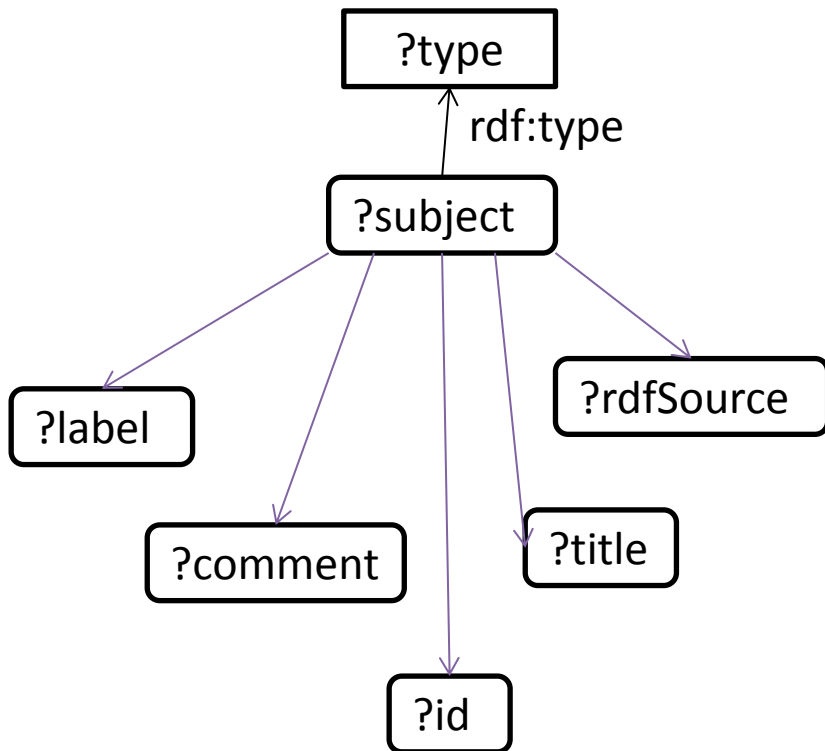
- SPARQL Queries over DisGeNET data
- *Minimal Resource Description Graph*

- `rdfs:label`: name + identifier
- `rdfs:comment`: human-readable description
- `dcterms:title`: resource name
- `dcterms:identifier`: namespace:identifier
- `void:inDataset`: RDF subset provenance



Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Minimal Resource Description Graph*



```
SELECT DISTINCT *
FROM <http://rdf.disgenet.org>
WHERE{
    ?subject rdf:type ?type ;
    rdfs:label ?label;
    rdfs:comment ?comment ;
    dcterms:identifier ?id ;
    dcterms:title ?title ;
    void:inDataset ?rdfSource .
}
LIMIT 100
```

Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Minimal Resource Description Graph*

Enter SPARQL Query

```

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX void: <http://rdfs.org/ns/void#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX nci: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
PREFIX up: <http://purl.uniprot.org/core/>

SELECT DISTINCT *
FROM <http://rdf.disgenet.org>
WHERE{
  ?subject rdf:type ?type ;
    rdfs:label ?label ;
    rdfs:comment ?comment ;
    dcterms:identifier ?id ;
    dcterms:title ?title ;
    void:inDataset ?rdfSource .
}
    
```

Results per page:

Submit Query

Output:

Example Queries

- Query 1
Get all gene-disease associations integrated in DisGeNET
- Query 2
Get all gene-disease associations integrated in DisGeNET searching by the 'Ovarian cancer' class in DO (DOID:2394)
- Query 3
Get all diseases in DisGeNET searching by 'Familial prostate cancer' class in ORDO (Orphanet_1331).

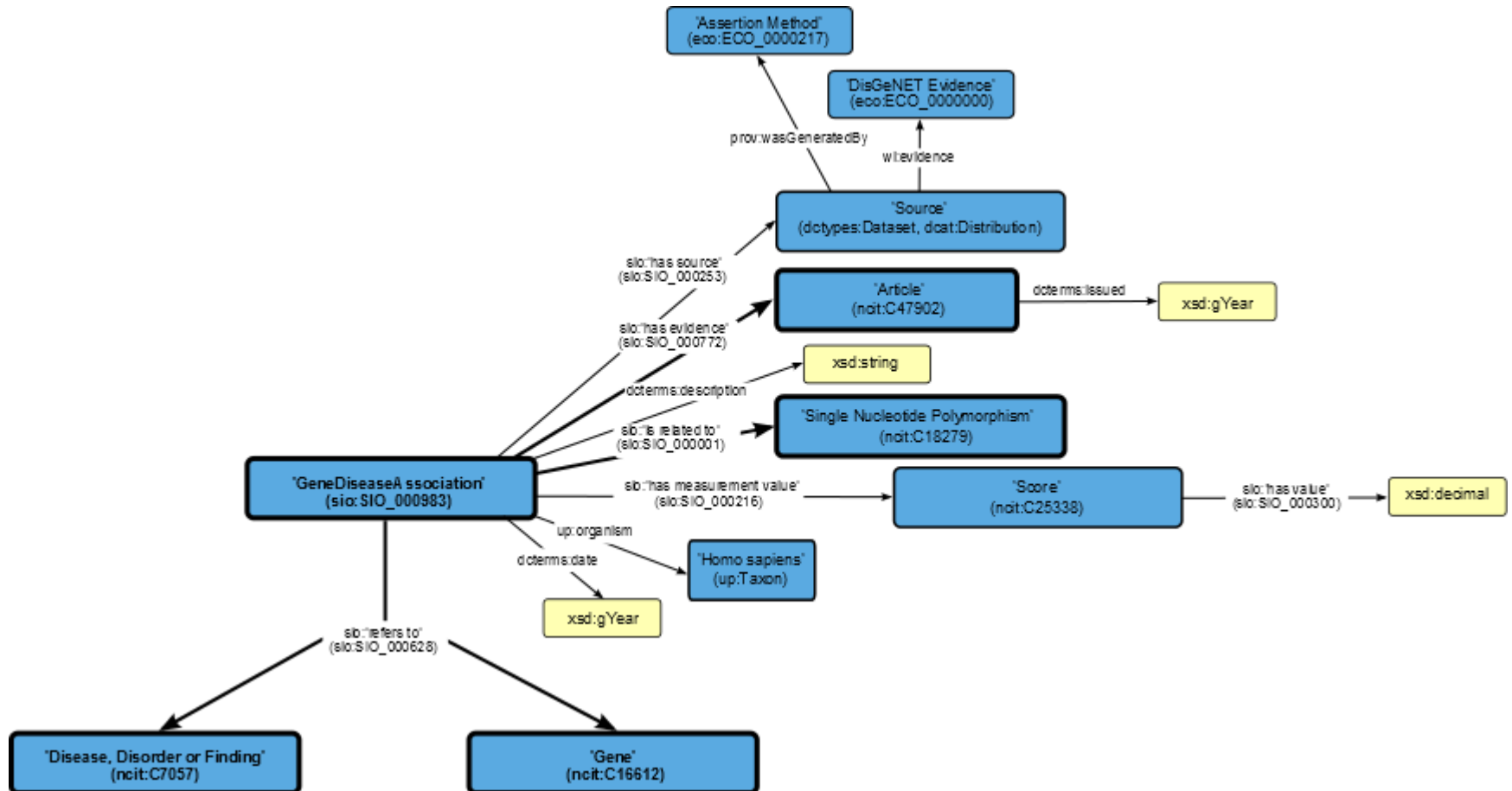
Previous

25 results per page (offset 0)

subject	type	label	comment	id	title	rdfSource
http://identifiers.org/ncbigene/100192455	ncit:C16612	fertility associated sperm antigen [ncbigene:100192455]	Gene [ncbigene:100192455] associated with at least one disease in DisGeNET. Genes are identified by the NCBI Entrez Gene ID from the NCBI Gene Database, a database of the U.S. National Library of Medicine.	ncbigene:100192455	fertility associated sperm antigen	http://rdf.disgenet.org/v3.0.0/void/gene
http://identifiers.org/ncbigene/10290	ncit:C16612	SPEG complex locus [ncbigene:10290]	Gene [ncbigene:10290] associated with at least one disease in DisGeNET. Genes are identified by the NCBI Entrez Gene ID from the NCBI Gene Database, a database of the U.S. National Library of Medicine.	ncbigene:10290	SPEG complex locus	http://rdf.disgenet.org/v3.0.0/void/gene
http://identifiers.org/ncbigene/1043	ncit:C16612	CD52 molecule [ncbigene:1043]	Gene [ncbigene:1043] associated with at least one disease in DisGeNET. Genes are identified by the NCBI Entrez Gene ID from the NCBI Gene Database, a database of the U.S. National Library of Medicine.	ncbigene:1043	CD52 molecule	http://rdf.disgenet.org/v3.0.0/void/gene
http://identifiers.org/ncbigene/10795	ncit:C16612	zinc finger protein 268 [ncbigene:10795]	Gene [ncbigene:10795] associated with at least one disease in DisGeNET. Genes are identified by the NCBI Entrez Gene ID from the NCBI Gene Database, a database of the U.S. National Library of Medicine.	ncbigene:10795	zinc finger protein 268	http://rdf.disgenet.org/v3.0.0/void/gene

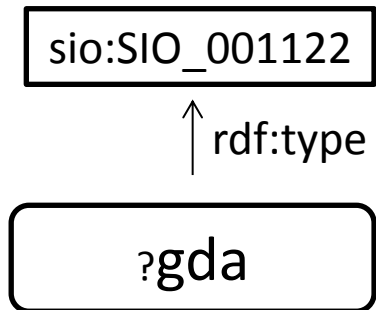
Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Gene-Disease Association Graph*



Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Gene-Disease Association Graph*



```
SELECT DISTINCT ?gda
FROM <http://rdf.disgenet.org>
WHERE{
    ?gda rdf:type sio:SIO_001122 .
}
LIMIT 100
```

Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Gene-Disease Association Graph*

Enter SPARQL Query

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX void: <http://rdfs.org/ns/void#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX nci: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
PREFIX up: <http://purl.uniprot.org/core/>

SELECT DISTINCT ?gda
FROM <http://rdf.disgenet.org>
WHERE{
  ?gda rdf:type sio:SIO_001122 .
}
LIMIT 100
```

Results per page: 25 ▼

Submit Query Reset

```
SELECT DISTINCT ?gda
FROM <http://rdf.disgenet.org>
WHERE{
  ?gda rdf:type sio:SIO_001122 .
}
LIMIT 100
```

[Previous](#)

gda

- <http://rdf.disgenet.org/resource/gda/DGN005116da025752508c8a8a711352233c>
- <http://rdf.disgenet.org/resource/gda/DGN0128c33d2d7d3611593dd3e364e8a5f8>
- <http://rdf.disgenet.org/resource/gda/DGN01a0e0bdf932c18c9604603bfe9ed7b>
- <http://rdf.disgenet.org/resource/gda/DGN01deea6a83200cf0c7da8c96cd95990f>
- <http://rdf.disgenet.org/resource/gda/DGN0201abd09acb60a9d9c6a5e9773dc99c>
- <http://rdf.disgenet.org/resource/gda/DGN04e13acc904b1092e07ee6119ac48579>
- <http://rdf.disgenet.org/resource/gda/DGN05310900eab4a0389626cc804963e69a>
- <http://rdf.disgenet.org/resource/gda/DGN079147e11ef2be0656cbb1ddfa1e928b>
- <http://rdf.disgenet.org/resource/gda/DGN0801812c587e464042a1885d53ac8a73>
- <http://rdf.disgenet.org/resource/gda/DGN08d09605f6eb3f6f7e715d73fe2587c3>

Querying DisGeNET

- **SPARQL Queries over DisGeNET data**
- *Gene-Disease Association Graph*

- Which is the `sio:SIO_001122` class?

Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Gene-Disease Association Graph*
- Which is the `sio:SIO_001122` class?

```
SELECT DISTINCT ?gda ?type ?label
FROM <http://rdf.disgenet.org>
WHERE {
  ?gda rdf:type ?type .
  FILTER(?type = sio:SIO_001122)
  ?type rdfs:label ?label
}
LIMIT 100
```

Querying DisGeNET

- **SPARQL Queries over DisGeNET data**
- *Gene-Disease Association Graph*
- For each ?gda, show me the ?gene and the ?disease associated, and the ?typeOfAssociation

Querying DisGeNET

- SPARQL Queries over DisGeNET data
 - *Gene-Disease Association Graph*
- For each ?gda, show me the ?gene and the ?disease associated, and the ?typeOfAssociation

```
SELECT DISTINCT ?gda ?gene ?disease ?type ?label
FROM <http://rdf.disgenet.org>
WHERE {
  ?gda rdf:type ?type ;
       sio:SIO_000628 ?gene, ?disease .
  ?type rdfs:label ?label .
  ?gene a ncit:C16612 .
  ?disease a ncit:C7057
}
LIMIT 50
```

Querying DisGeNET

- **SPARQL Queries over DisGeNET data**
- *Gene-Disease Association Graph*
- For each ?gda, show me the ?gene and the ?disease associated, the ?paper, and the ?sentence description of the relationship in the paper

Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Gene-Disease Association Graph*

• For each ?gda, show me the ?gene and the ?disease associated, the ?paper, and the ?sentence description of the relationship in the paper

```
SELECT DISTINCT ?gda ?gene ?disease ?paper ?sentence
FROM <http://rdf.disgenet.org>
WHERE {
  ?gda sio:SIO_000628 ?gene, ?disease ;
      sio:SIO_000772 ?paper ;
      dcterms:description ?sentence .
  ?gene a ncit:C16612 .
  ?disease a ncit:C7057
}
LIMIT 50
```

Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Gene-Disease Association Graph*

• For each ?gda, show me the ?gene and the ?disease associated, the ?paper, and the ?sentence description of the relationship in the paper

```
SELECT DISTINCT ?gda ?gene ?disease ?paper ?sentence
FROM <http://rdf.disgenet.org>
WHERE {
  ?gda sio:SIO_000628 ?gene, ?disease ;
      sio:SIO_000772 ?paper ;
      dcterms:description ?sentence .
  FILTER(regex(str(?sentence), "syndrome", "i"))
  ?gene a ncit:C16612 .
  ?disease a ncit:C7057
}
LIMIT 50
```

Querying DisGeNET

- **SPARQL Queries over DisGeNET data**
 - *Gene-Disease Association Graph*
- For each ?gda show me the ?gene, ?disease, ?source, and the level of ?evidence of the association

Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Gene-Disease Association Graph*

• For each ?gda show me the ?gene, ?disease, ?source, and the level of ?evidence of the association

```
PREFIX wi: <http://purl.org/ontology/wi/core#>

SELECT DISTINCT ?gda ?gene ?disease ?source ?evidence
FROM <http://rdf.disgenet.org>
WHERE {
  ?gda sio:SIO_000628 ?gene, ?disease ;
      sio:SIO_000253 ?source .
  ?gene a ncit:C16612 .
  ?disease a ncit:C7057 .
  ?source wi:evidence ?evidence
}
LIMIT 50
```

Querying DisGeNET

- SPARQL Queries over DisGeNET data
 - *Gene-Disease Association Graph*
- For each **gene-disease pair** show me the ?number of evidences and the score ?value

Querying DisGeNET

- **SPARQL Queries over DisGeNET data**
 - *Gene-Disease Association Graph*
- For each **gene-disease pair** show me the ?number of evidences and the score ?value

```
SELECT DISTINCT ?gene ?disease count(DISTINCT ?gda) AS ?numberOfEvidences
?scoreValue
FROM <http://rdf.disgenet.org>
WHERE {
?gda sio:SIO_000628 ?gene, ?disease ;
    sio:SIO_000216 ?score .
?gene a ncit:C16612 .
?disease a ncit:C7057 .
?score sio:SIO_000300 ?scoreValue
}
ORDER BY DESC(?numberOfEvidences) DESC(?scoreValue)
LIMIT 50
```

Querying DisGeNET

- **SPARQL Queries over DisGeNET data**
- *Gene-Disease Association Graph*
- For each ?gda show me the ?snp

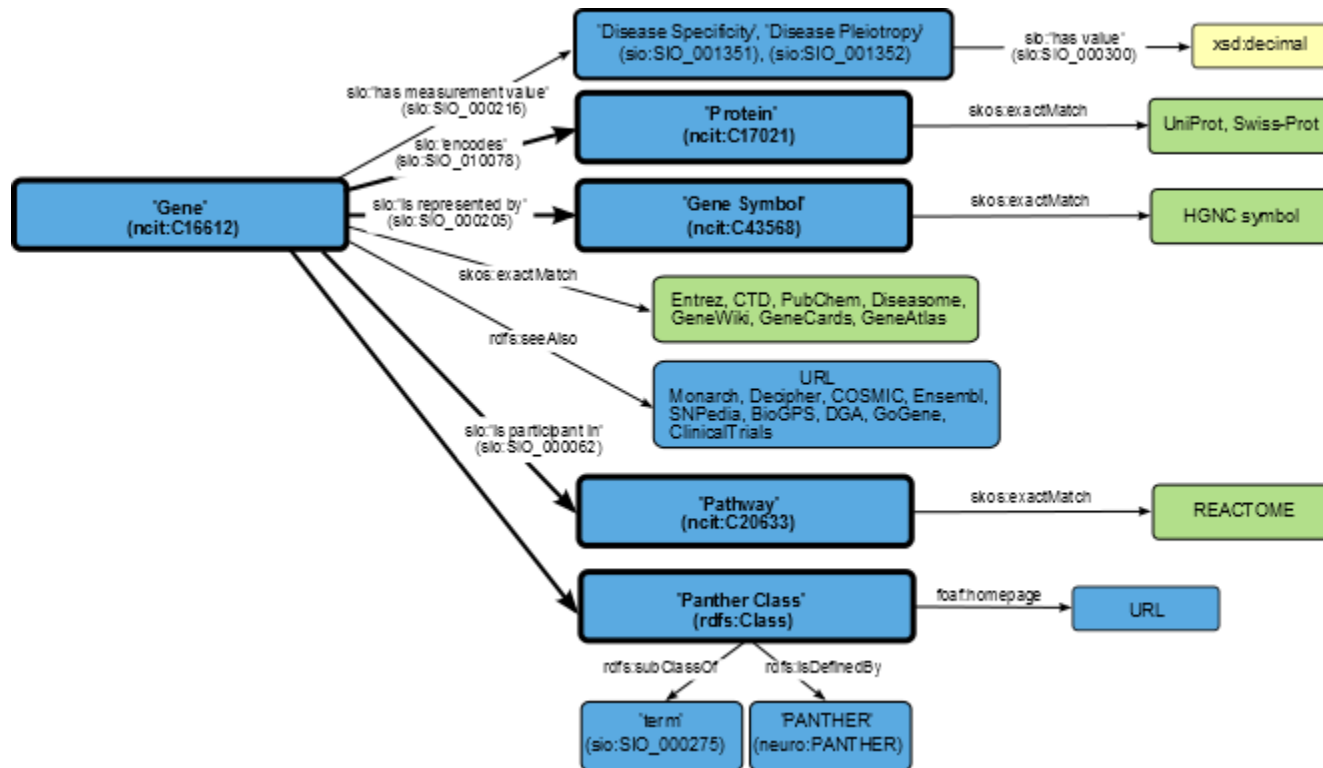
Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Gene-Disease Association Graph*
- For each ?gda show me the ?snp
- Go to the Web and understand and execute Q1.1-Q1.4

```
SELECT DISTINCT ?gda ?gene ?disease ?snp FROM
<http://rdf.disgenet.org>
WHERE {
?gda sio:SIO_000628 ?gene, ?disease ;
    sio:SIO_000001 ?snp .
?gene a ncit:C16612 .
?disease a ncit:C7057 .
}
LIMIT 50
```

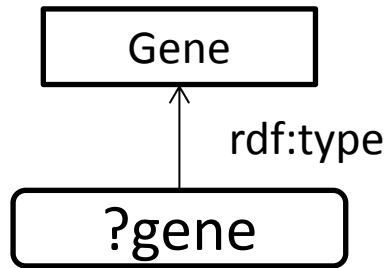
Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Gene Graph*



Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Gene Graph*



```
SELECT DISTINCT ?gene
FROM <http://rdf.disgenet.org>
WHERE{
    ?gene rdf:type ncit:C16612 .
}
LIMIT 100
```

Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Gene Graph*

Enter SPARQL Query

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX void: <http://rdfs.org/ns/void#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
PREFIX up: <http://purl.uniprot.org/core/>

SELECT DISTINCT ?gene
FROM <http://rdf.disgenet.org>
WHERE{
    ?gene rdf:type ncit:C16612 .
}
LIMIT 100
```

Results per page: 25 ▼

Submit Query Reset

Previous

gene

<http://identifiers.org/ncbigene/100192455>>
<http://identifiers.org/ncbigene/10290>>
<http://identifiers.org/ncbigene/1043>>
<http://identifiers.org/ncbigene/10795>>
<http://identifiers.org/ncbigene/11325>>
<http://identifiers.org/ncbigene/114769>>
<http://identifiers.org/ncbigene/117581>>
<http://identifiers.org/ncbigene/127343>>
<http://identifiers.org/ncbigene/130802>>
<http://identifiers.org/ncbigene/148713>>
<http://identifiers.org/ncbigene/165>>

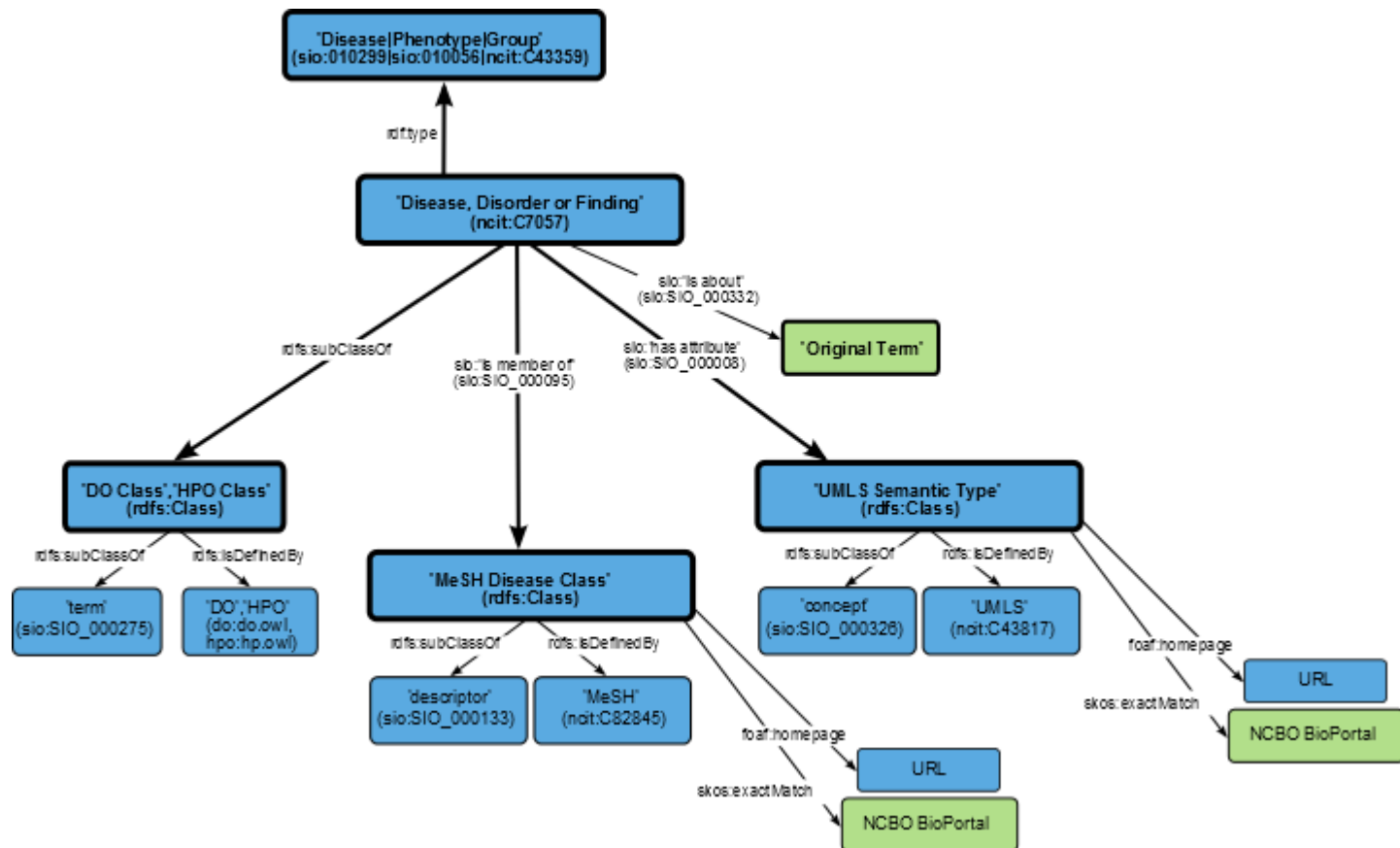
```
SELECT DISTINCT ?gene
FROM <http://rdf.disgenet.org>
WHERE{
    ?gene rdf:type ncit:C16612 .
}
LIMIT 100
```


Querying DisGeNET

- **SPARQL Queries over DisGeNET data**
- *Gene Graph*
 - For each ?gene show me:
 - ?identifier, ?name, ?geneSymbol
 - ?protein(s)
 - ?panther class(es) and ?pantherclassname
 - ?pathway(s) and ?pathwayname
 - Go to web and understand/execute Q1.5

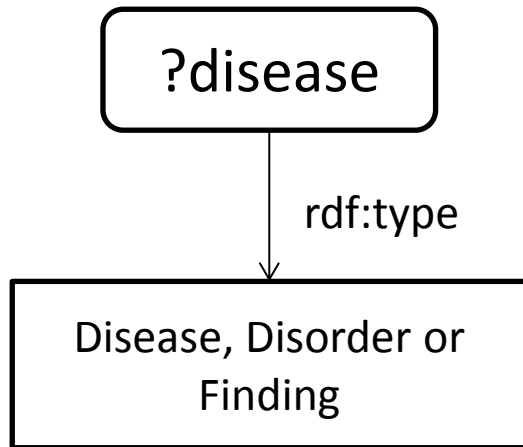
Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Disease Graph*



Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Disease Graph*



```
SELECT DISTINCT ?disease
FROM <http://rdf.disgenet.org>
WHERE{
    ?disease a ncit:C7057 .
}
LIMIT 100
```

Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Disease Graph*

Enter SPARQL Query

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX void: <http://rdfs.org/ns/void#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
PREFIX up: <http://purl.uniprot.org/core/>

SELECT DISTINCT ?disease
FROM <http://rdf.disgenet.org>
WHERE{
    ?disease a ncit:C7057 .
}
LIMIT 100
```

Results per page: 25 ▼

```
SELECT DISTINCT ?disease
FROM <http://rdf.disgenet.org>
WHERE{
    ?disease a ncit:C7057 .
}
LIMIT 100
```

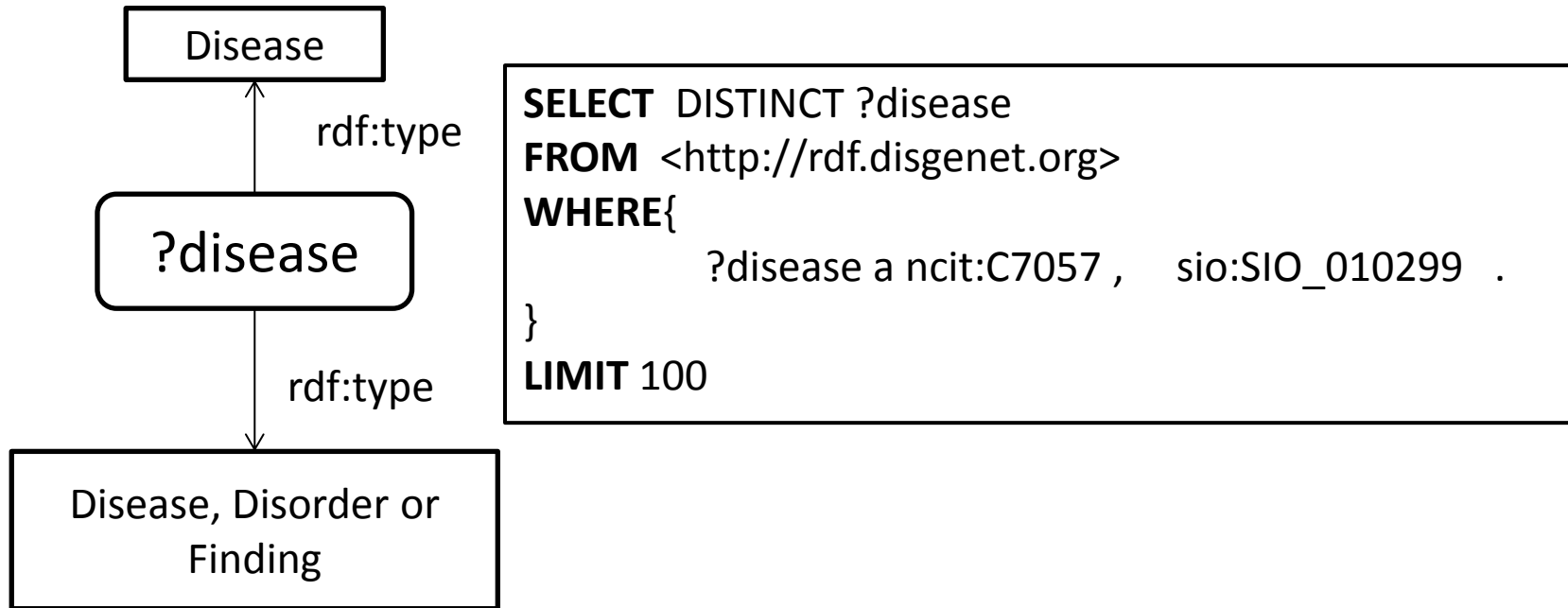
[Previous](#)

disease

[-<http://linkedlifedata.com/resource/umls/id/C0007133>](http://linkedlifedata.com/resource/umls/id/C0007133)
[-<http://linkedlifedata.com/resource/umls/id/C0011430>](http://linkedlifedata.com/resource/umls/id/C0011430)
[-<http://linkedlifedata.com/resource/umls/id/C0014078>](http://linkedlifedata.com/resource/umls/id/C0014078)
[-<http://linkedlifedata.com/resource/umls/id/C0018920>](http://linkedlifedata.com/resource/umls/id/C0018920)
[-<http://linkedlifedata.com/resource/umls/id/C0040100>](http://linkedlifedata.com/resource/umls/id/C0040100)
[-<http://linkedlifedata.com/resource/umls/id/C1300127>](http://linkedlifedata.com/resource/umls/id/C1300127)
[-<http://linkedlifedata.com/resource/umls/id/C2363142>](http://linkedlifedata.com/resource/umls/id/C2363142)
[-<http://linkedlifedata.com/resource/umls/id/C0009450>](http://linkedlifedata.com/resource/umls/id/C0009450)
[-<http://linkedlifedata.com/resource/umls/id/C0026948>](http://linkedlifedata.com/resource/umls/id/C0026948)
[-<http://linkedlifedata.com/resource/umls/id/C0080032>](http://linkedlifedata.com/resource/umls/id/C0080032)
[-<http://linkedlifedata.com/resource/umls/id/C0751774>](http://linkedlifedata.com/resource/umls/id/C0751774)

Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Disease Graph*



Querying DisGeNET

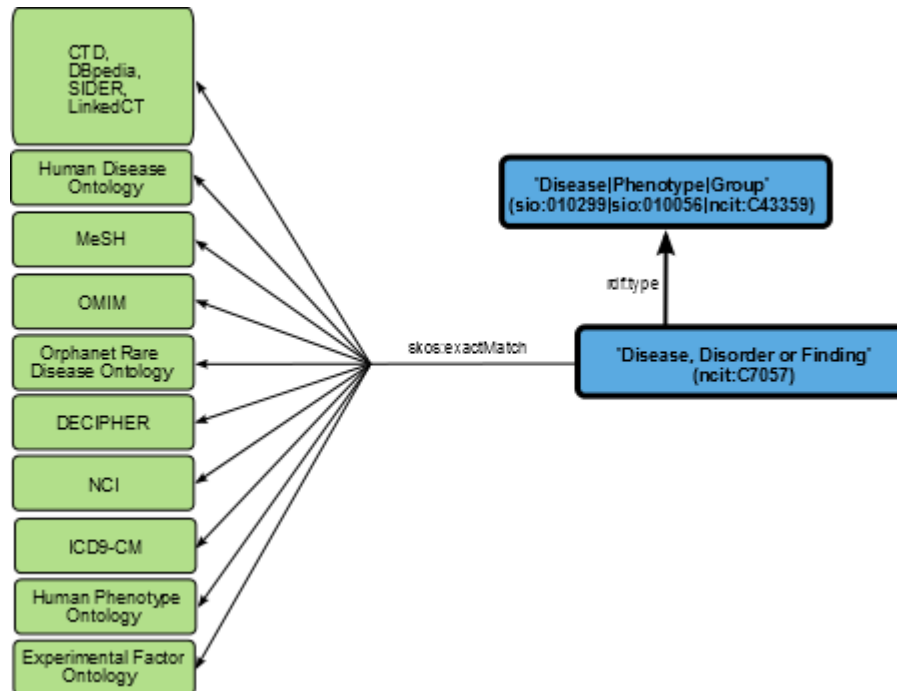
- **SPARQL Queries over DisGeNET data**
- *Disease Graph*

- For the disease `<http://linkedlifedata.com/resource/umls/id/C0596263>` show me:
 - the disease ?name, MeSH disease class ?label, and the umlsSTY ?title
 - show all cross-references to other disease terminologies

- Go to the Web and understand/execute Q1.6

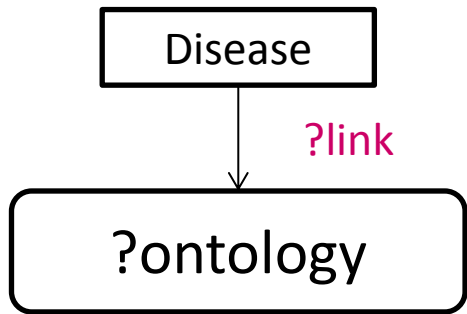
Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Disease* mapping to other ontologies



Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Disease* mapping to other ontologies



```
SELECT DISTINCT ?disease
FROM <http://rdf.disgenet.org>
WHERE{
    ?disease skos:exactMatch ?ontology .
}
```

COVERAGE

	UMLS	MeSH	OMIM	NCIt	DO	ORDO	ICD9CM	EFO	HPO	DECIPHER
% Diseases	100	57	40	34	20	14	11	11	8	0.4

Querying DisGeNET

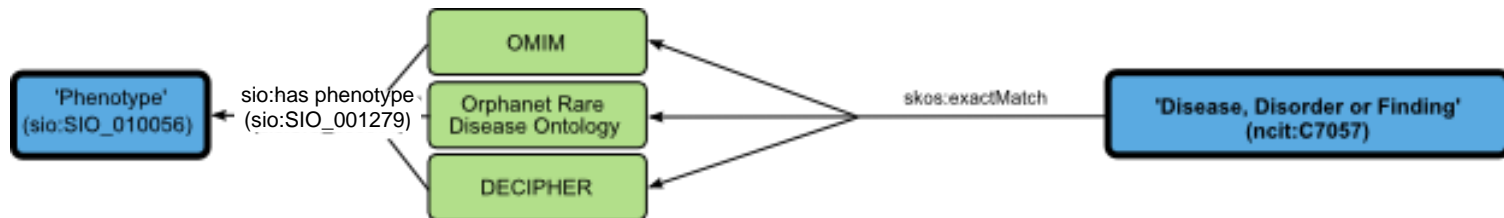
- **SPARQL Queries over DisGeNET data**
- *Ontology Walking queries*
 - Grouping of similar instances
 - Filtering data
 - Query data by classes

```
?child rdfs:subClassOf+ ?parent
```

- Ontologies loaded in our RDF triple store: SIO, DO, ORDO, NCIT, HPO, EFO, and ECO (OWL)
 - Go to the Web and understand/execute Q1.7 and Q1.11

Querying DisGeNET

- SPARQL Queries over DisGeNET data
- *Disease-Phenotype Association Graph* (curated from HPO)



Querying DisGeNET

- **SPARQL Queries over DisGeNET data**
- *Disease-Phenotype Association Graph* (curated from HPO)
 - **Why this model?**

```
SELECT DISTINCT ?disease count(distinct ?hpdisease) as ?hpdiseases count(distinct
?phenotype) as ?phenotypes WHERE {
    ?disease rdf:type ncit:C7057 .
    ?disease skos:exactMatch ?hpdisease .
    ?hpdisease sio:SIO_001279 ?phenotype .
}
ORDER BY DESC(?hpdiseases)
LIMIT 100
```

```
SELECT DISTINCT ?disease ?hpdisease count(distinct ?phenotype) as ?phenotypes
WHERE {
    ?disease rdf:type ncit:C7057 .
    ?disease skos:exactMatch ?hpdisease .
    ?hpdisease sio:SIO_001279 ?phenotype .
    FILTER (?disease = <http://linkedlifedata.com/resource/umls/id/C3280766>)
}
GROUP BY ?disease ?hpdisease
```

Querying DisGeNET

- **SPARQL Queries over DisGeNET data**
- *Disease-Phenotype Association Graph* (curated from HPO)
- How many phenotypes are associated with Orphanet:209

Querying DisGeNET

- **SPARQL Queries over DisGeNET data**
- *Disease-Phenotype Association Graph* (curated from HPO)
- How many phenotypes are associated with Orphanet:209

```
SELECT DISTINCT ?disease ?hpdisease count(distinct ?phenotype) as ?phenotypes
WHERE {
    ?disease rdf:type ncit:C7057 .
    ?disease skos:exactMatch ?hpdisease .
    ?hpdisease sio:SIO_001279 ?phenotype .
    FILTER (?hpdisease = <http://identifiers.org/orphanet/209>)
}
```

Querying DisGeNET

- **SPARQL Queries over DisGeNET data**
- *Disease-Phenotype Association Graph* (curated from HPO)
 - How many diseases are associated with a phenotype

Querying DisGeNET

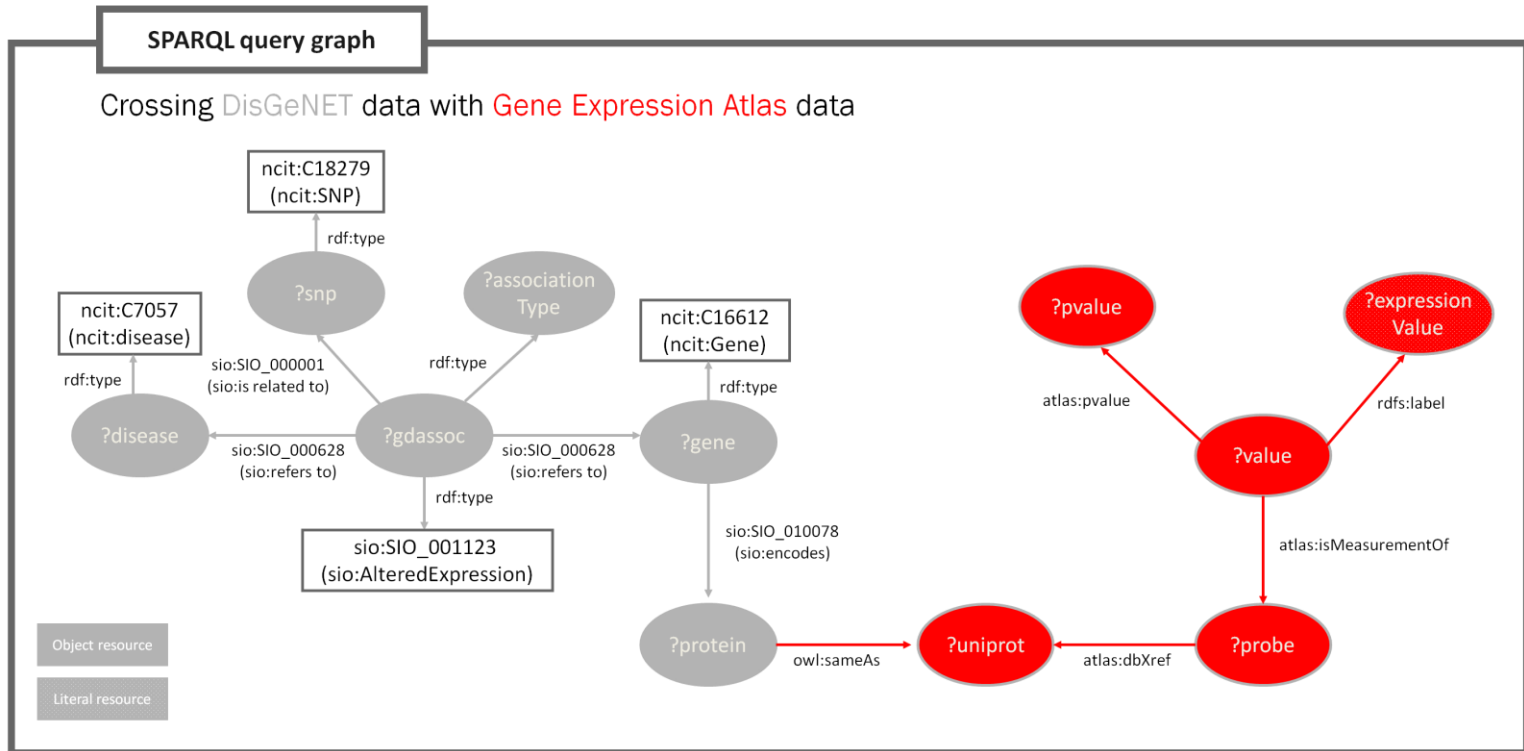
- **SPARQL Queries over DisGeNET data**
- *Disease-Phenotype Association Graph* (curated from HPO)
 - How many diseases are associated with a phenotype

```
SELECT DISTINCT ?phenotype ?phenotypeName count(distinct ?disease) as
?diseases
WHERE {
    ?hpdisease sio:SIO_001279 ?phenotype .
    ?phenotype dcterms:title ?phenotypeName .
    ?disease skos:exactMatch ?hpdisease .
    ?disease rdf:type ncit:C7057 ;
        dcterms:title ?diseaseName .
}
ORDER BY DESC(?diseases)
LIMIT 100
```

- Go to the Web and understand/execute Q1.10 and Q1.12

Querying DisGeNET + LOD cloud

- Federated Queries: DisGeNET + external datasets



- Go to the Web and understand/execute the Federated Queries

Use Cases

- What genes are associated to *Marfan syndrome*?
- What **evidence** supports the association between APP gene and Alzheimer Disease?
- What **disease classes** are associated with APP gene?
- Which genes and evidence support the **comorbidity** between Chronic Kidney disease and Diabetes Mellitus, Type 2?
- What **SNPs** are related to the MECP2 and Rett Syndrome association?
- Which diseases are associated to **post-translational modifications** type of association?
- What disease genes are hit by compounds in **ChEMBL**?
- What disease genes have differential expression in **Gene Expression Atlas**?
- What disease genes are in **WikiPathways**?
- Find **compounds** (from **ChEMBL**) that target **genes** (from **DisGeNET**) that participate in the same **pathway** (from **WikiPathways**)

Thanks for your attention!
Questions are welcome

