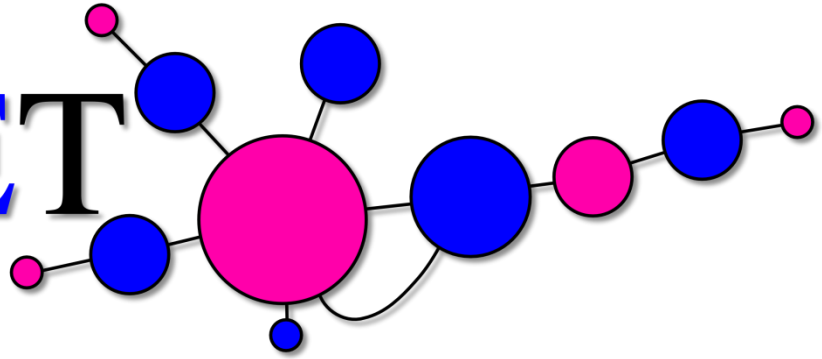
DisGeNET



A discovery platform for translational research

Usage Tutorial

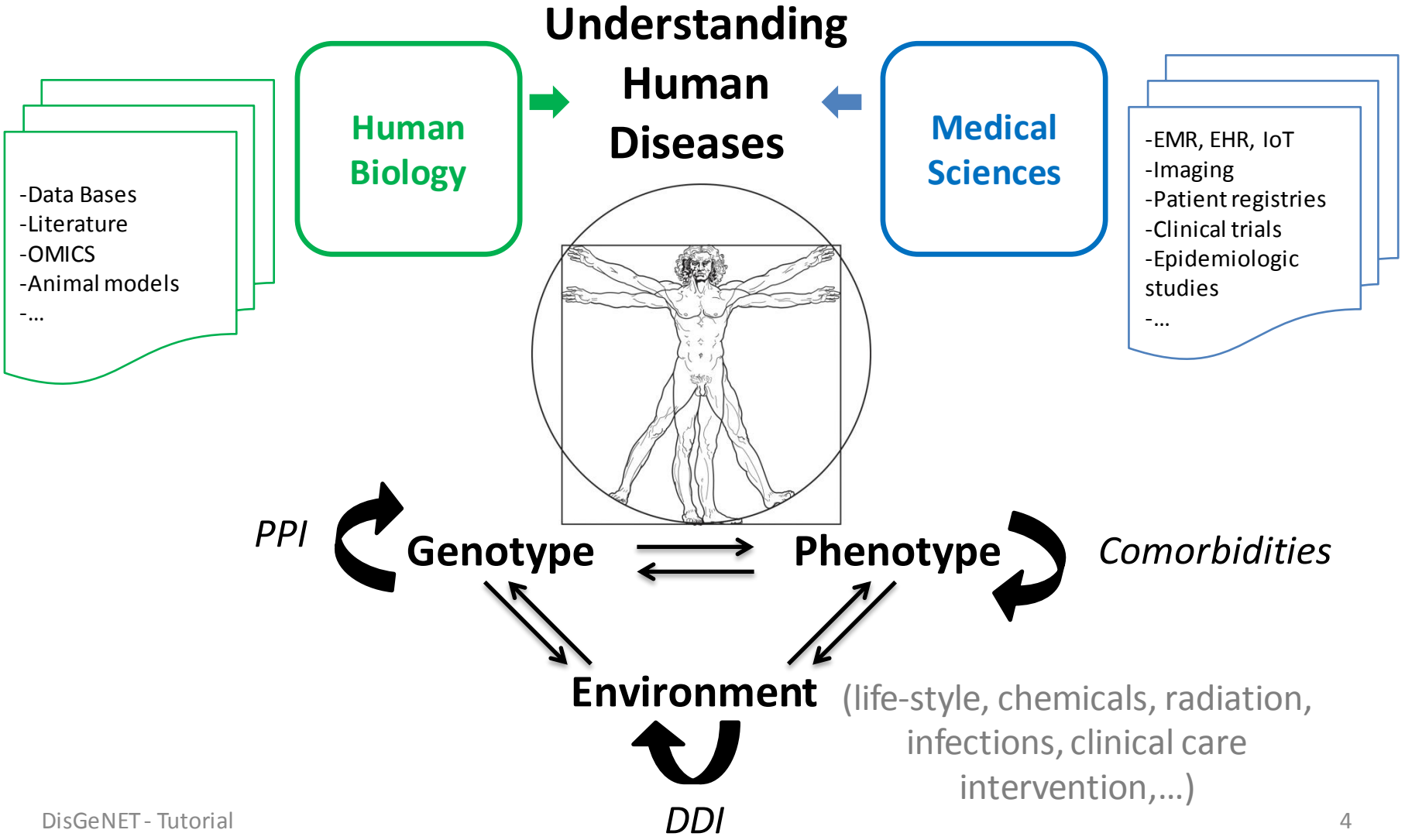
Núria Queralt Rosinach
Integrative Biomedical Informatics Group (IBI)
Research Programme on Biomedical Informatics (GRIB)
Hospital del Mar Research Institute (IMIM)
Pompeu Fabra University (UPF)
Barcelona

Outline

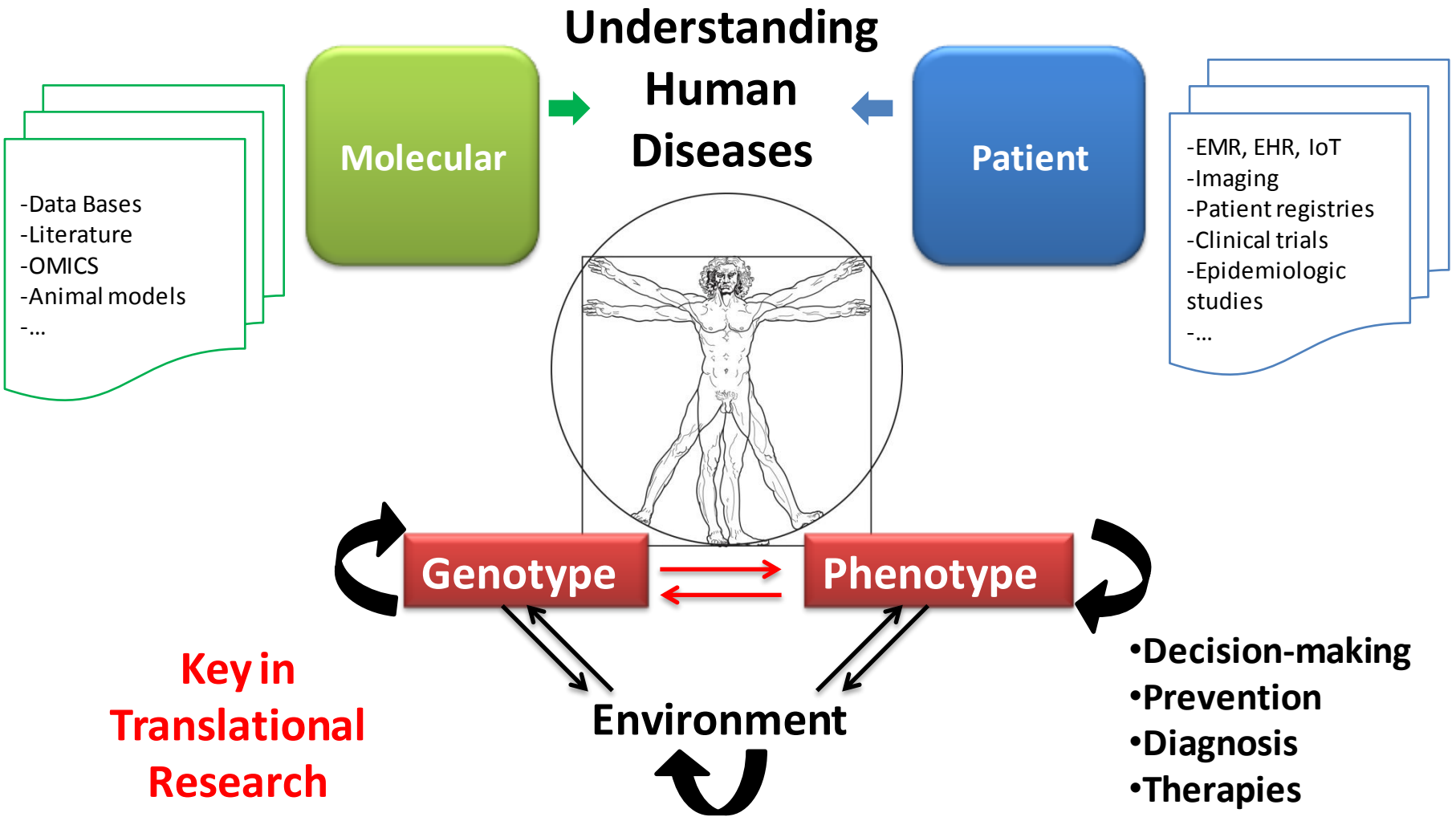
- How can DisGeNET help your research?
- DisGeNET Discovery Platform Overview
- DisGeNET Linked Open Data
  - Introduction
    - RDF-LD Description: Data Model, VoID, Interlinking
    - Implementation
    - Accessibility
    - Documentation
    - Use Cases
  - Querying the DisGeNET-RDF
    - Hands-on

How can DisGeNET help your research?

Big Questions 4 Big Data



Translational Research



Access to Gene-Disease Associations

Mental retardation - ? - SOX3

SOX3

OMIM Online Mendelian Inheritance in Man — OMIM:300123; OMIM:312000

orphanet — ORPHA393; ORPHA90695; ORPHA3157; ORPHA79495; ORPHA67045

UniProt — Mental Retardation; Panhypopituitarism; 46,XX sex reversal 3

ClinVar — No Data

ctd — MESH:C538613; MESH:C538613

Access to Gene-Disease Associations

Mental retardation - ? - SOX3

SOX3

OMIM
Online Mendelian Inheritance in Man

OMIM:300123; OMIM:312000

orphanet

ORPHA393; HA67045

UniProt

Mental Retardation

TATATCT ACCTCAC ClinVar

No Data

ctd™

MESH:C538613; MESH:C538613

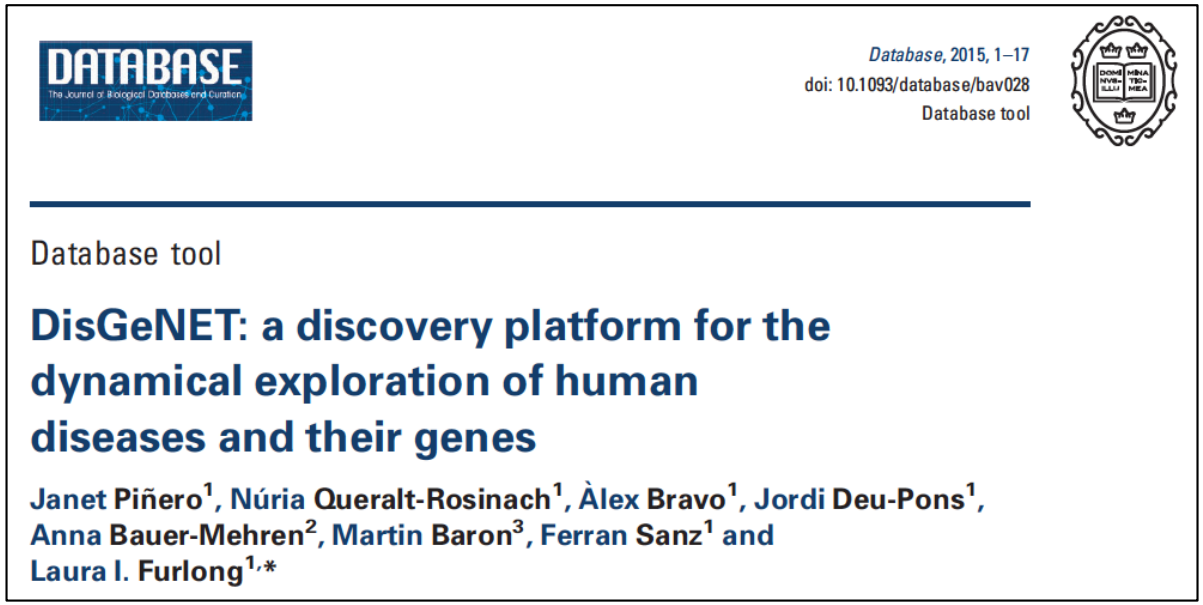
Lack of:

- **Normalization**
- **Semantic integration**
- **Data model harmonization**
- **Unified access**

DisGeNET

http://www.disgenet.org/

- **Knowledge platform** on human **gene-disease associations (GDAs)**
- Integrates information from expert-**curated** databases and from the **literature** (text mining)
- All disease areas
- Supporting **evidence**
- Analysis tools



DATABASE
The Journal of Biological Databases and Curation

Database, 2015, 1–17
doi: 10.1093/database/bav028
Database tool

Database tool

DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes

Janet Piñero[1], Núria Queralt-Rosinach[1], Àlex Bravo[1], Jordi Deu-Pons[1], Anna Bauer-Mehren[2], Martin Baron[3], Ferran Sanz[1] and Laura I. Furlong[1,*]

•Piñero *et al*. **DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes**. Database (2015) Vol. 2015: article ID bav028, (2015)

Research Questions

- Which genes are associated to **Marfan syndrome**?
- Which disease genes have **approved drugs** annotated?
- Which disease genes have **differential expression**?
- Which disease genes share a **pathway**?
- Is there **genetic variation** related to the MECP2 and Rett Syndrome association?
- What **evidence** supports the association between *APP* gene and *Alzheimer Disease*?
- Which genes and evidence support the **comorbidity** between *Chronic Kidney* disease and *Diabetes Mellitus, Type 2*?

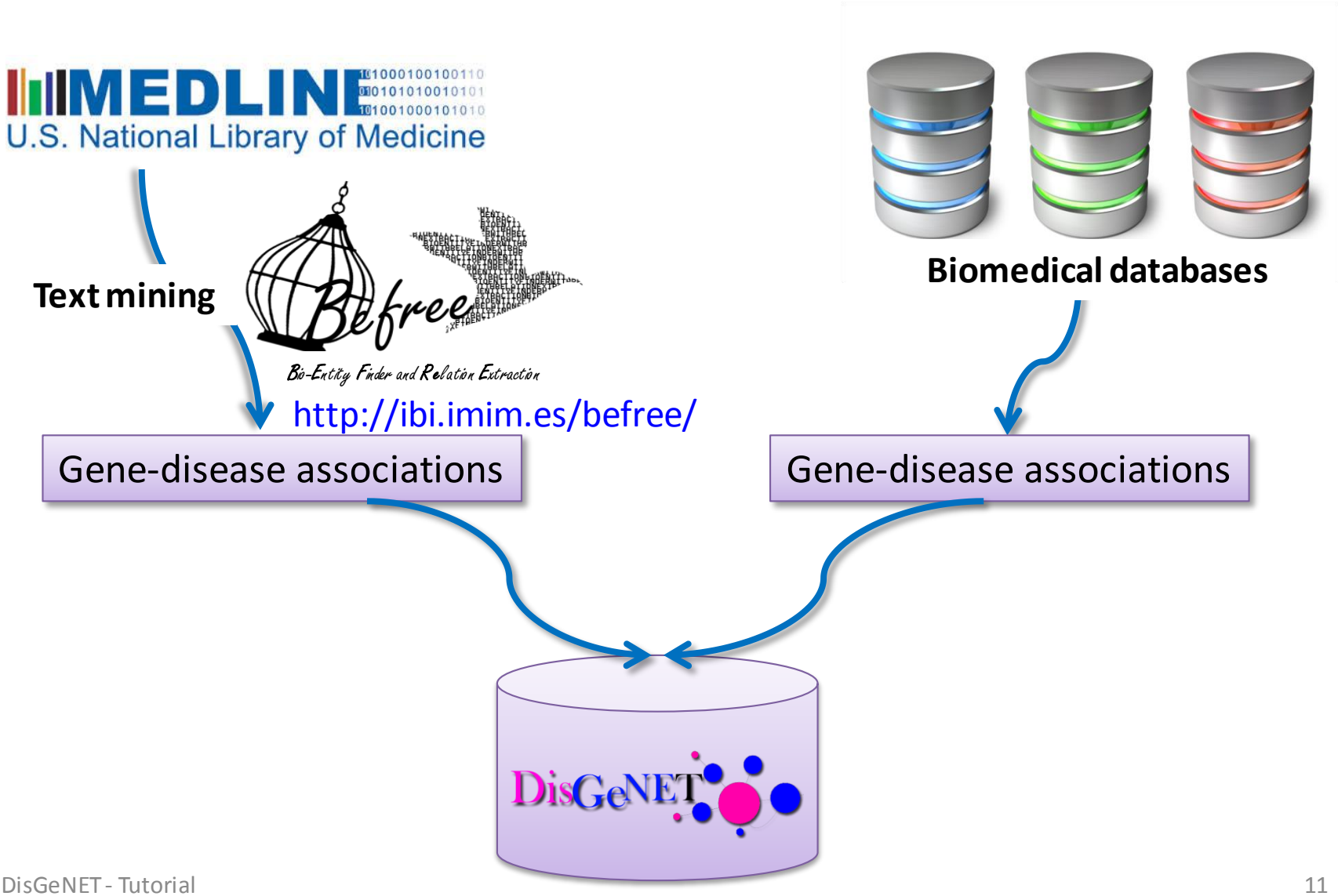
Answer biomedical questions requires search, collection and integration of data

KNOWLEDGE DISCOVERY

ACTIONABLE INFORMATION

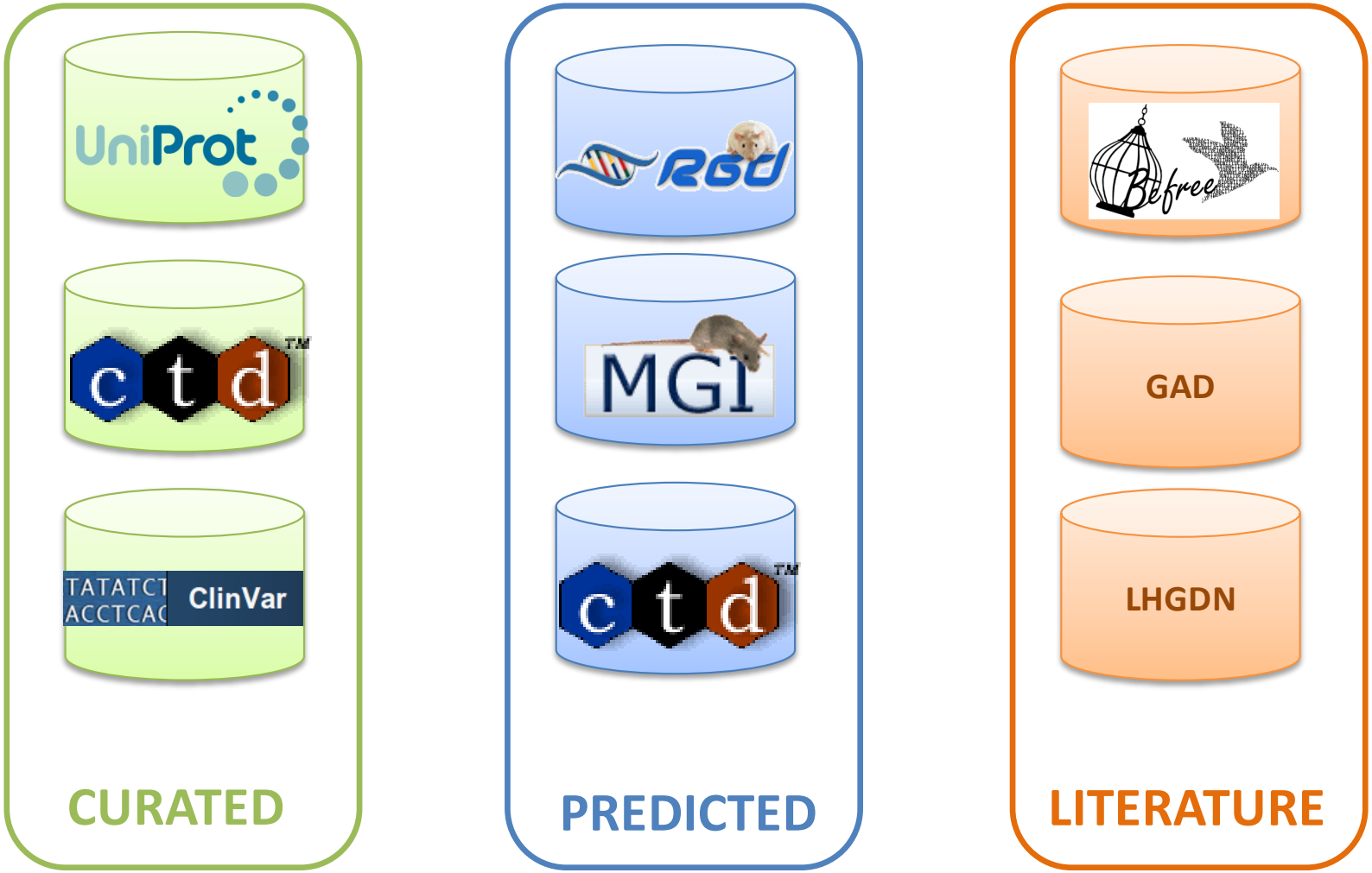
DisGeNET Discovery Platform Overview

DisGeNET Implementation

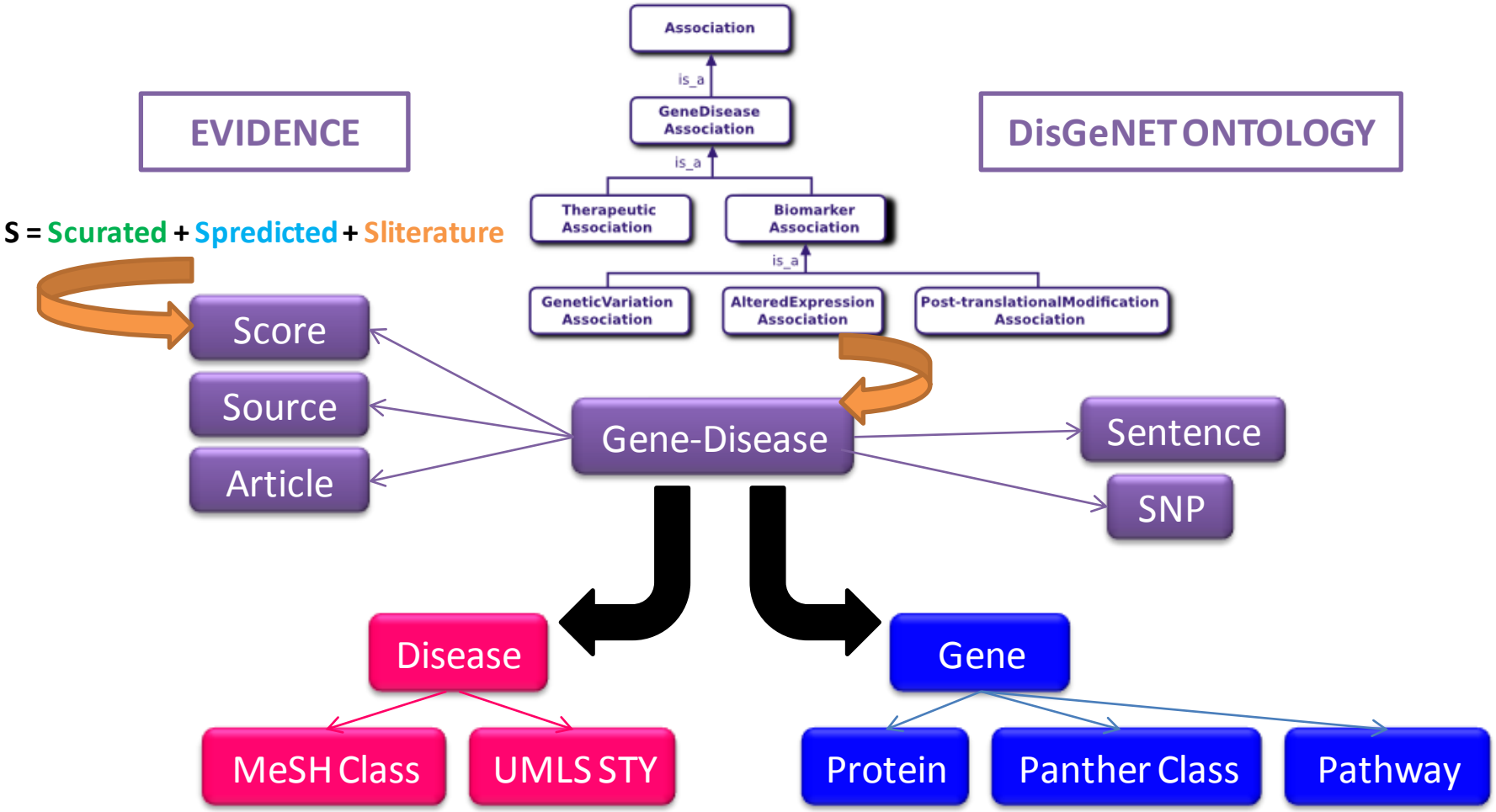


DisGeNET Sources

DisGeNET v3.0

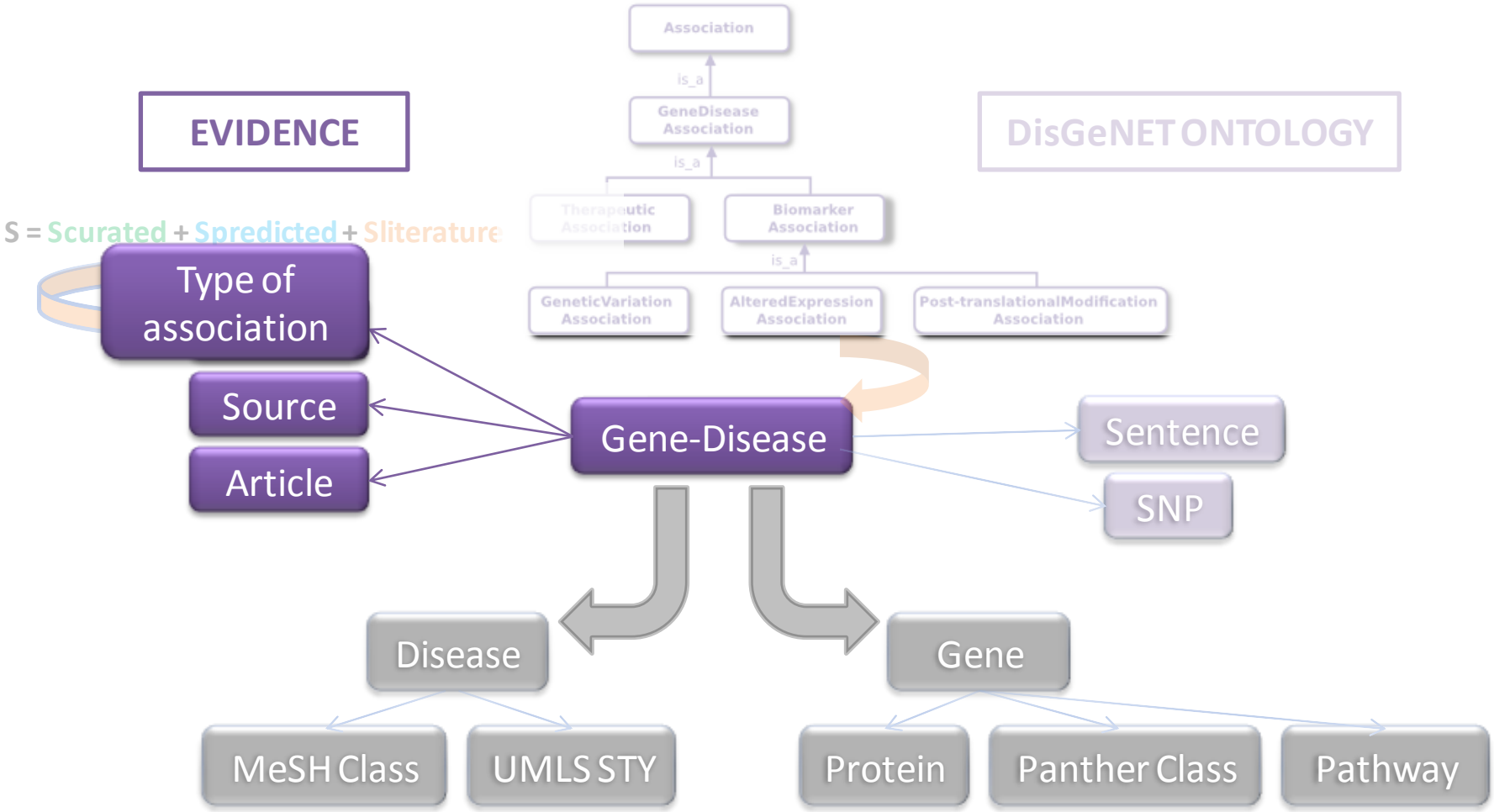


Data Integration



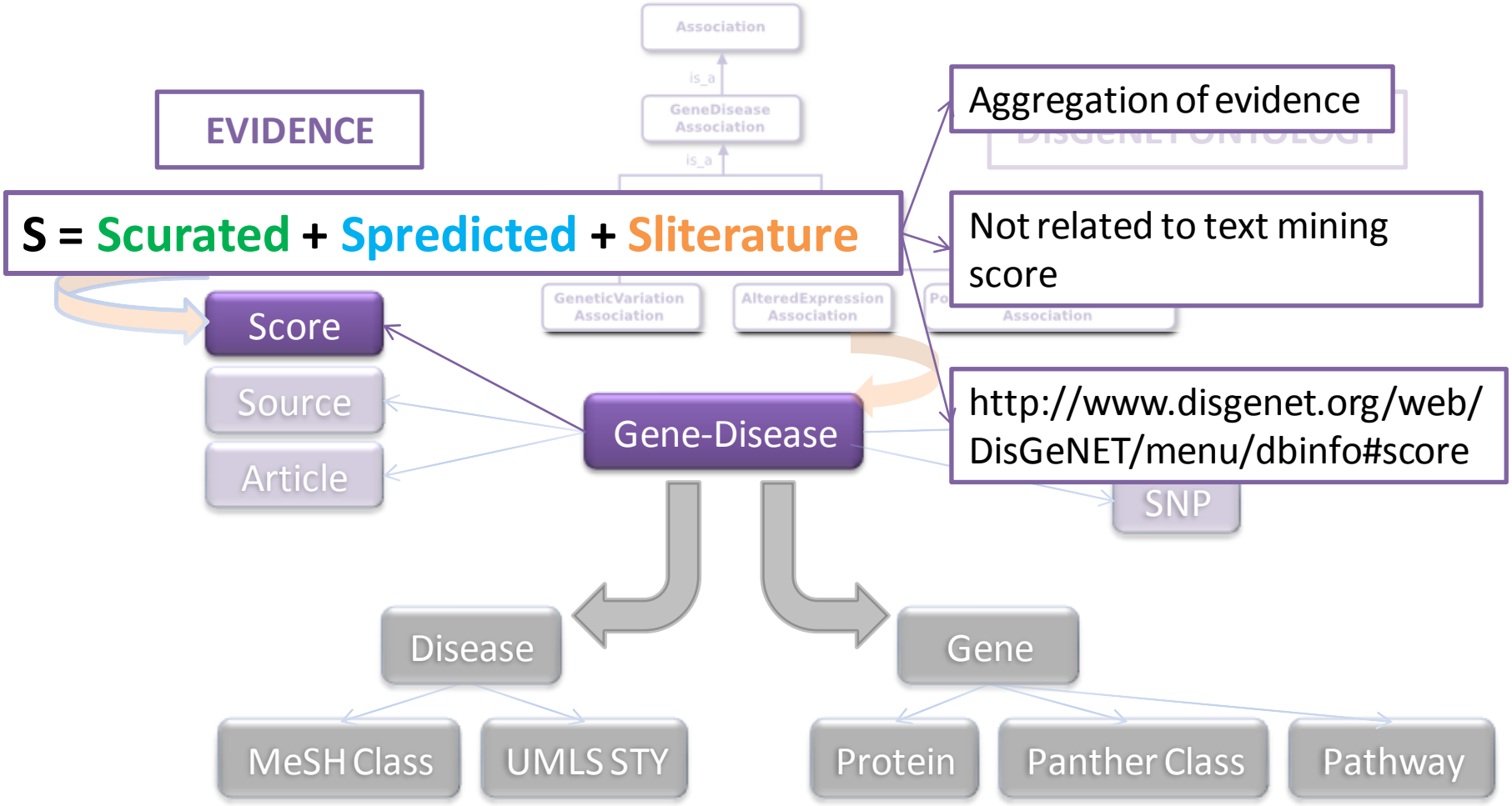
- Use of Standards and controlled vocabularies

Data Integration



- Use of Standards and controlled vocabularies

Data Integration



- Use of Standards and controlled vocabularies

DisGeNET Statistics (May 15th, 2015)

Annual Release

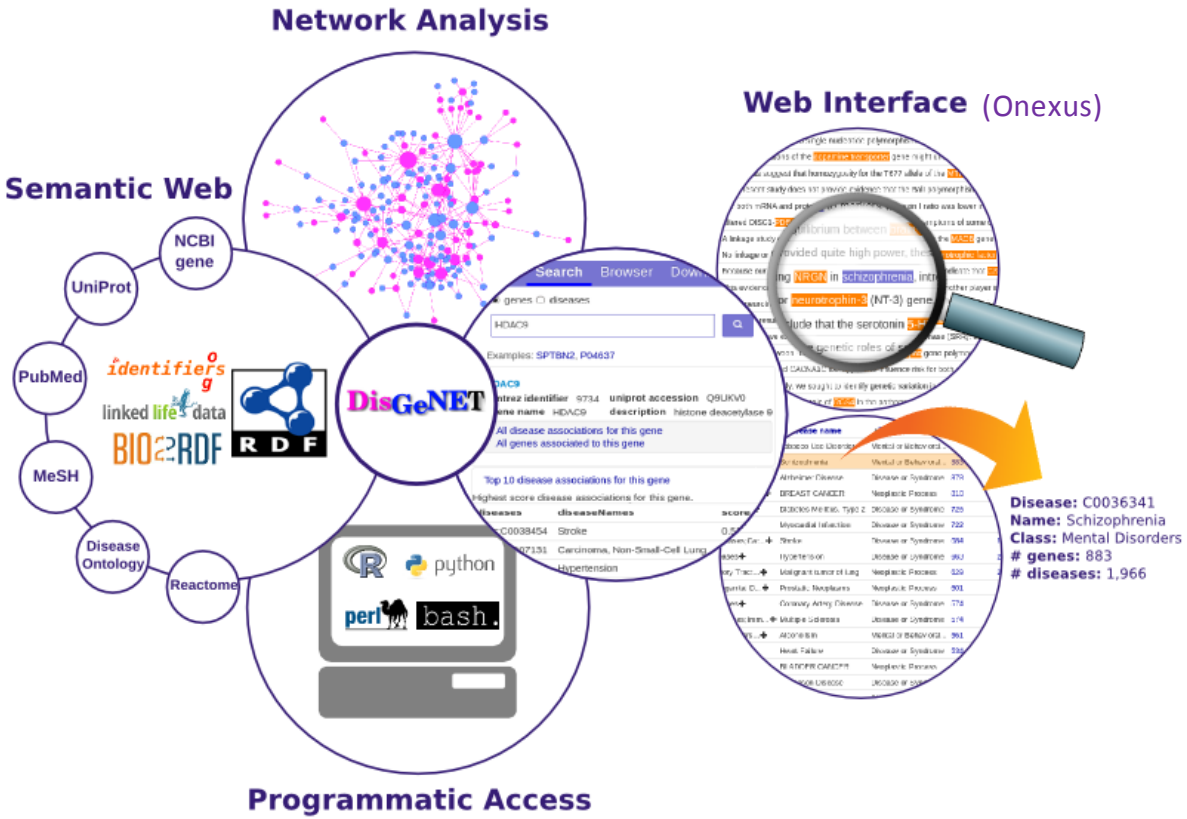
DisGeNET v3.0

Source	Genes	Diseases	Associations
Curated	7,878	6,761	26,522
Predicted	2,557	2,003	9,536
Literature	16,298	11,374	408,175
All	**17,181**	**14,619**	**429,111**

82 %

Large volume of information unlocked by text mining the literature

Tools for exploration



Usage stats (Ago2014-Ago2015):

- 12,040 **users**, 22,696 sessions (4:33 min/session)
- 14,494 **downloads** (database, Cytoscape plugin, RDF/Nanopubs)
- DisGeNET **used in +20** publications, **cited in +60** articles

Web Interface

SOX3
NCBI:6658

Panhypopituitarism
UMLS:C0342376

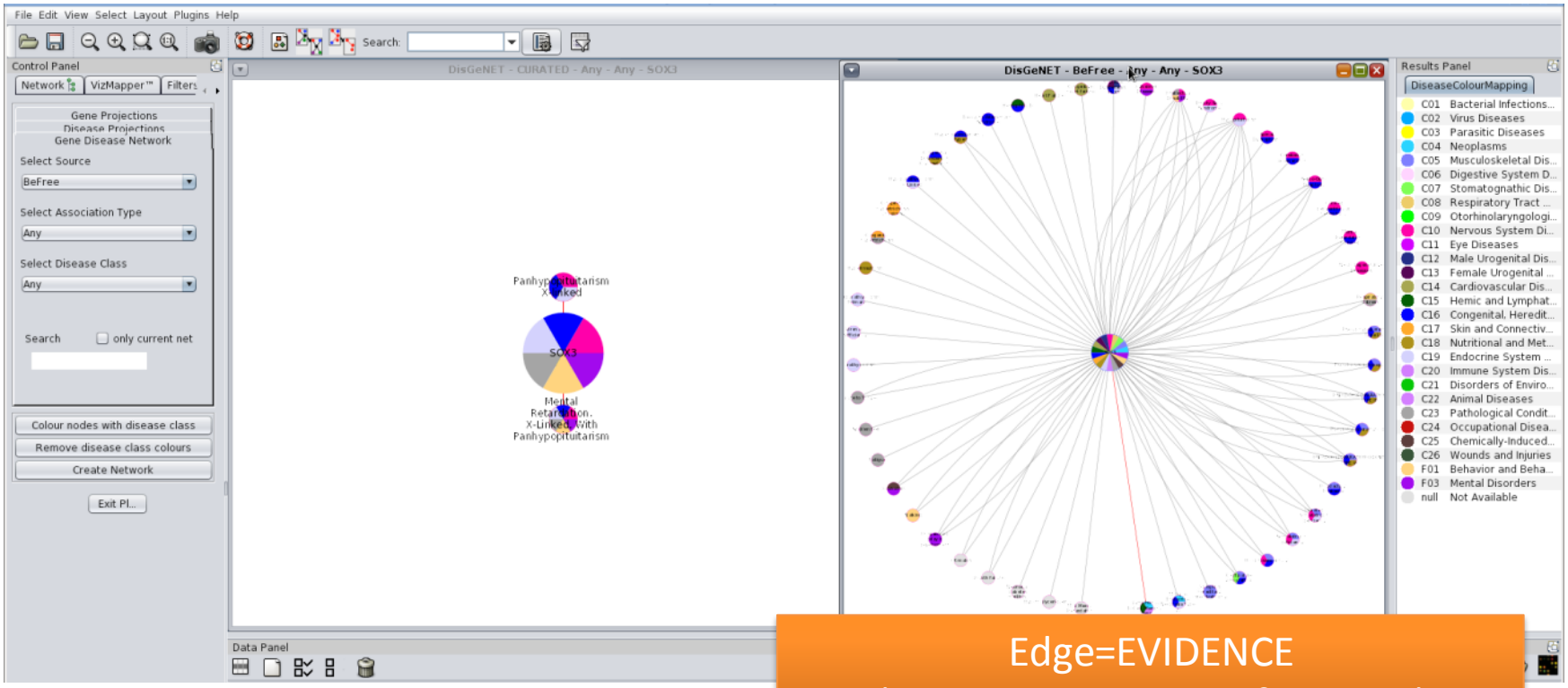
Symbol	Uniprot	Gene Name	Pathway	Panther Protein Class	# Of Diseases
SOX3	P41225	SRY (sex determining region Y)-box 3		nucleic acid binding; transcr...	47

Disease	Score	Disease Name	Origina db	Association Type	PMID
umls:C0342376	0.41	Panhypopituitarism X-linked	CLINVAR	GeneticVariation	
umls:C0342376	0.41	Panhypopituitarism X-linked	CTD_human	Biomarker	
umls:C2678223	0.31	Mental Retardation, X-Linked, With Panhypopituitarism	CTD_human	Biomarker	
umls:C2678223	0.31	Mental Retardation, X-Linked, With Panhypopituitarism	MGD	Biomarker	2350537
umls:C1848068	0.2	Mental Retardation, X-Linked, with Isolated Growth Hormone Deficiency	CLINVAR	GeneticVariation	
umls:C0265216	0.1	Hydrocephalus, X-linked	MGD	Biomarker	
umls:C0020635	0.01	Hypopituitarism	BeFree	Biomarker	1533827
umls:C0020635	0.01	Hypopituitarism	BeFree	Biomarker	1534269

Cytoscape Plugin

SOX3
NCBI:6658

Panhypopituitarism
UMLS:C0342376



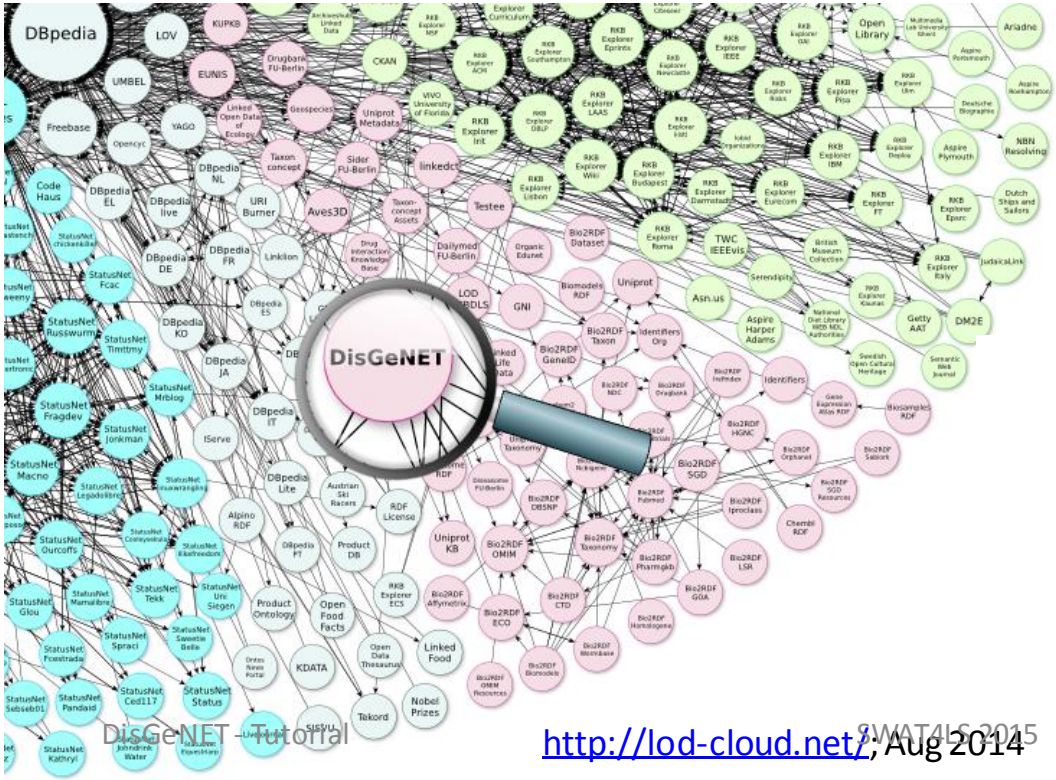
Edge=EVIDENCE
(Source, PMID, type of relation)

DisGeNET Linked Open Data

DisGeNET as Linked Open Data

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

BioPortal



http://lod-cloud.net/, Aug 2014

- **Metadata** description (W3C HCLS)
- **Interlinking**
  - BIO2RDF
  - linked life data
- Access
  - **Download** Data Dump
  - **SPARQL** Endpoint
  - **Faceted** Browser
  - Open PHACTS Discovery Platform
  - Nanopublication Network
- **Open** license
- Datahub
- Software

D2RQ

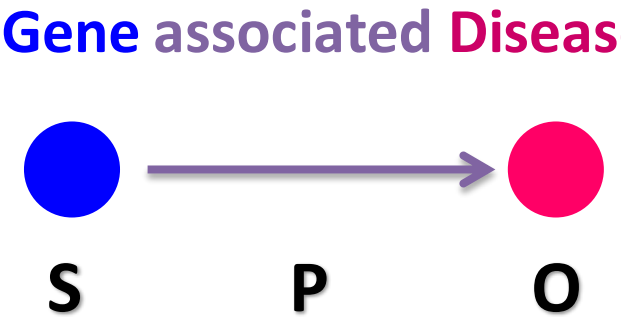
Virtuoso Universal Server — OPENLINK SOFTWARE Making Technology Work For You®

DisGeNET-RDF

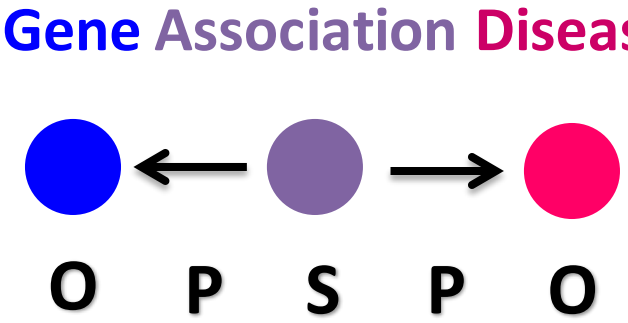
Data Model

- How to describe an **association**?

a) As a **property**



b) As a **class**

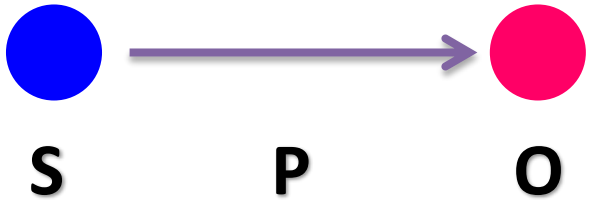


Data Model

- How to describe an **association**?

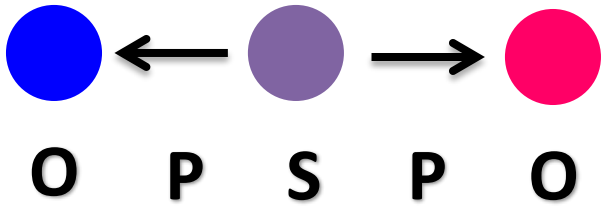
a) As a **property**

Gene associated Disease



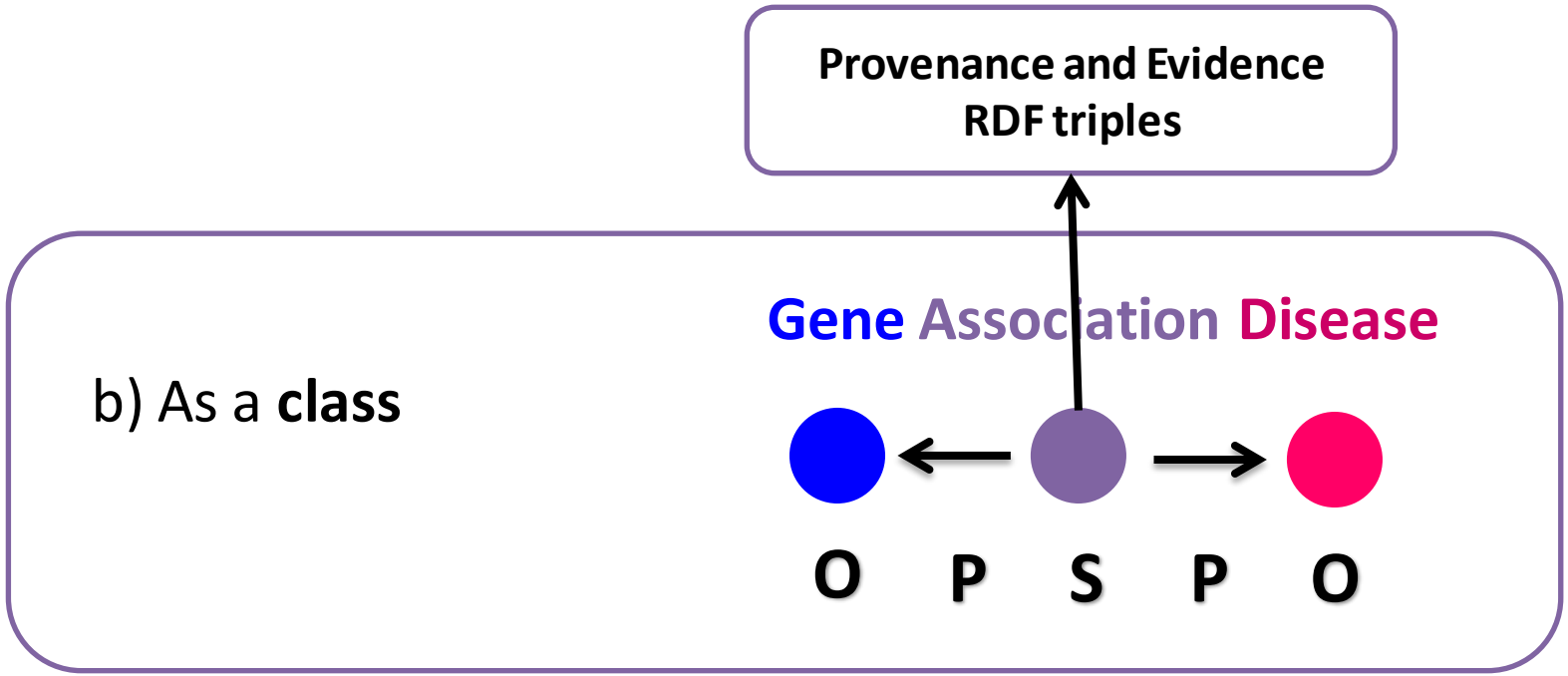
b) As a **class**

Gene Association Disease



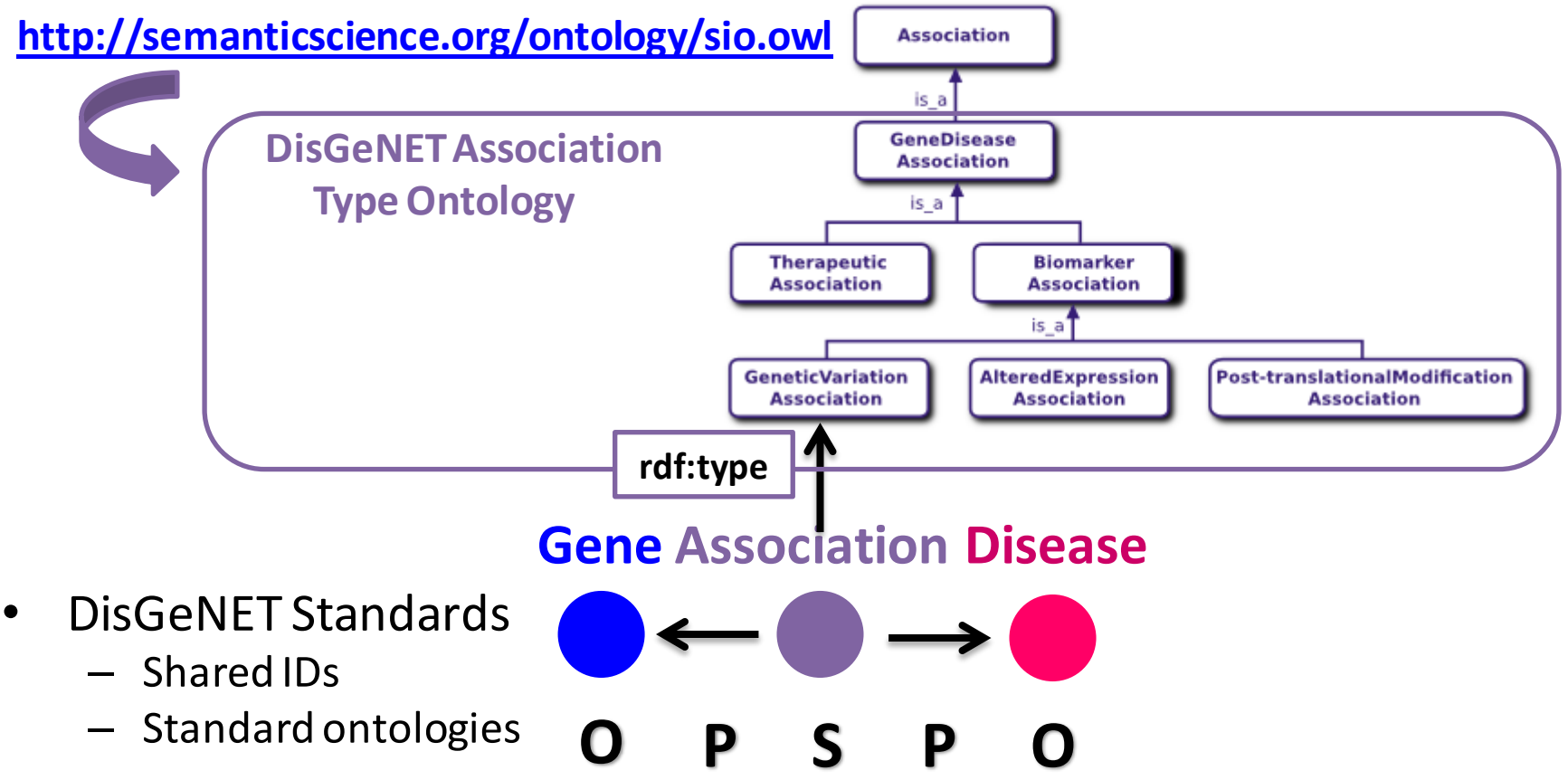
Data Model

- How to describe an **association**?



Data Model

- Ontology-based **integration**



- 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	Vocabula
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

Biomedical entities

Relationships

RDF Structure

Data Model

- Semantic Annotation: **Standard ontologies**

Prefix	Namespace	Vocabula
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

Biomedical entities

Relationships

Metadata

RDF Structure

Data Model

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

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

    Unique association attributes

  - Estable 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**

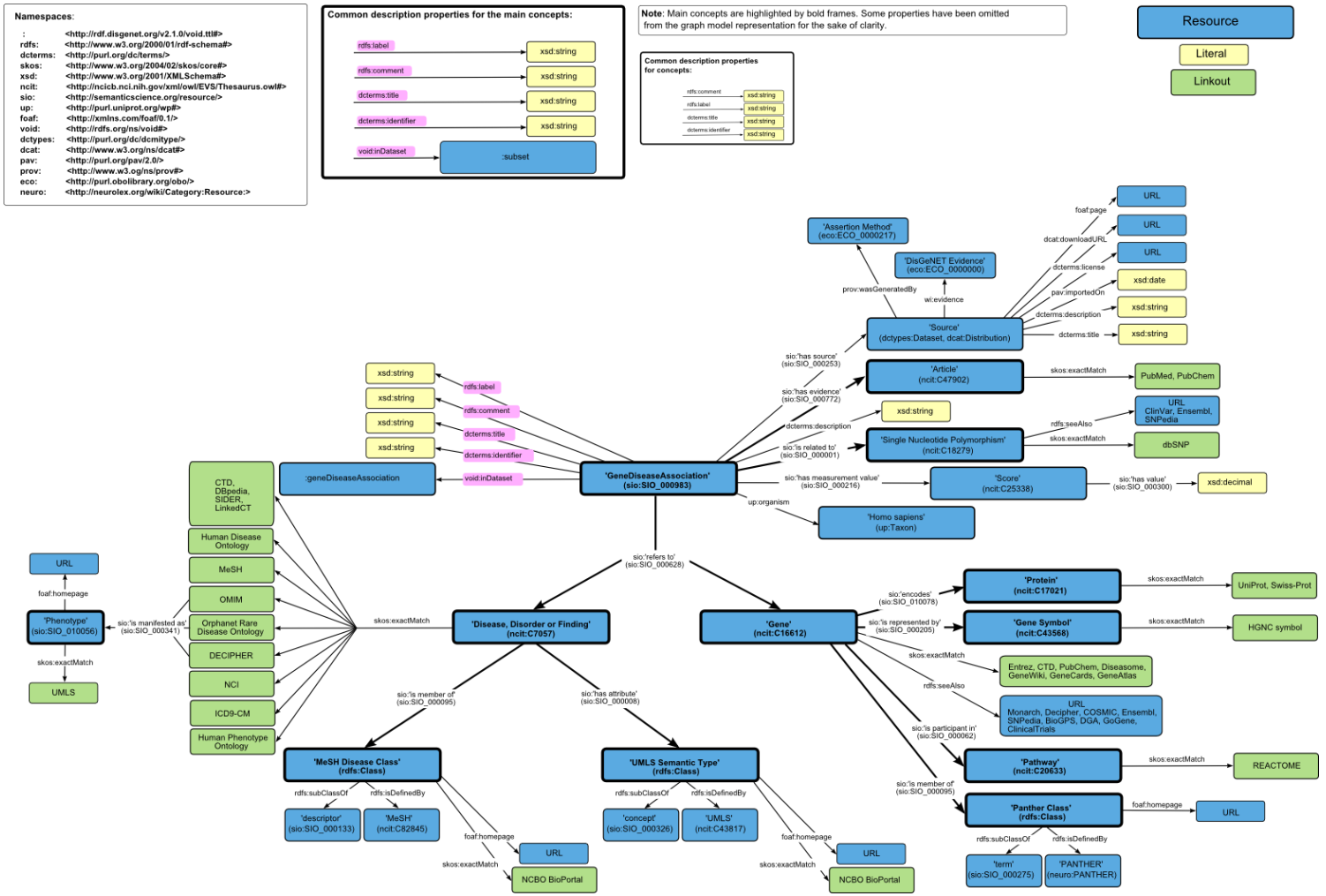
Entity	URI	Semantics
Gene	http://identifiers.org/ncbigene/**4728**	ncit:C16612
HGNC Gene Symbol	http://identifiers.org/hgnc.symbol/**NDUFS8**	ncit:C43568
Protein	http://identifiers.org/uniprot/**O00217**	ncit:C17021
Panther Class	http://rdf.disgenet.org/resource/panther.classification/**PC00211**	rdfs:Class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

Data Model

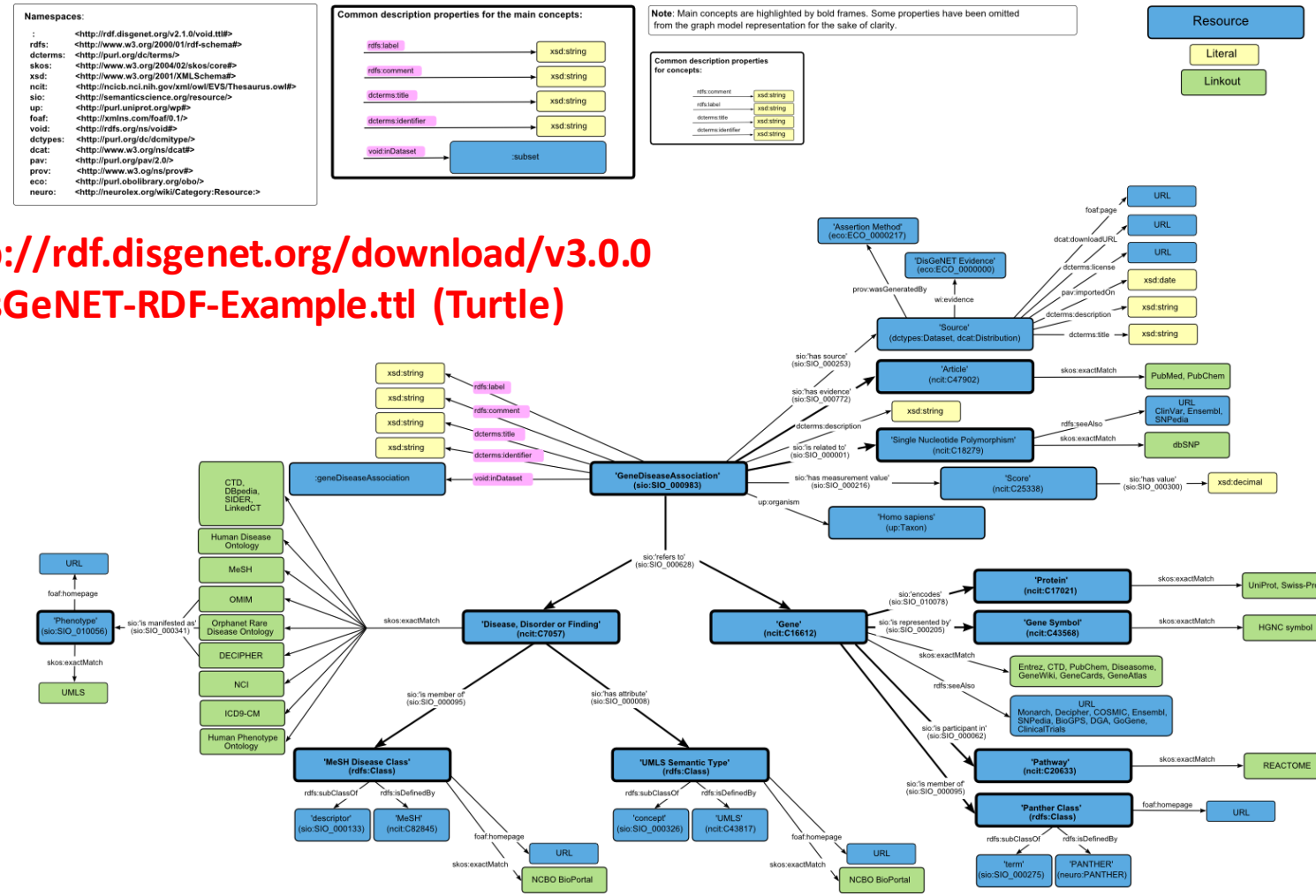
- URIs in DisGeNET: **shared, cool & dereferenceable**
  - **ID Normalization**
  - **Disease::UMLS Concept Unique Identifier (CUI)**

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:
:
rdfs:
dcterms:
skos:
xsd:
ncit:
sio:
up:
foaf:
void:
dctypes:
dcat:
pav:
prov:
eco:
neuro:
Common description properties for the main concepts:
rdfs:label
rdfs:comment
dcterms:title
dcterms:identifier
void:inDataset
xsd:string
:subset
Note: Main concepts are highlighted by bold frames. Some properties have been omitted from the graph model representation for the sake of clarity.
Common description properties for concepts:
Resource
Literal
Linkout
'GeneDiseaseAssociation' (sio:SIO_000983)
:geneDiseaseAssociation
'Source' (dctypes:Dataset, dcat:Distribution)
'Assertion Method' (eco:ECO_0000217)
'DisGeNET Evidence' (eco:ECO_0000000)
'Article' (ncit:C47902)
'Single Nucleotide Polymorphism' (ncit:C18279)
'Score' (ncit:C25338)
'Homo sapiens' (up:Taxon)
CTD, DBpedia, SIDER, LinkedCT
Human Disease Ontology
MeSH
OMIM
Orphanet Rare Disease Ontology
DECIPHER
NCI
ICD9-CM
Human Phenotype Ontology
'Phenotype' (sio:SIO_010056)
'Disease, Disorder or Finding' (ncit:C7057)
'Gene' (ncit:C16612)
'Protein' (ncit:C17021)
'Gene Symbol' (ncit:C43568)
'Pathway' (ncit:C20633)
'Panther Class' (rdfs:Class)
'MeSH Disease Class' (rdfs:Class)
'UMLS Semantic Type' (rdfs:Class)
NCBO BioPortal
URL
UMLS
PubMed, PubChem
dbSNP
UniProt, Swiss-Prot
HGNC symbol
REACTOME
Entrez, CTD, PubChem, Diseasome, GeneWiki, GeneCards, GeneAtlas
URL Monarch, Decipher, COSMIC, Ensembl, SNPedia, BioGPS, DGA, GoGene, ClinicalTrials
URL ClinVar, Ensembl, SNPedia
xsd:date
xsd:decimal

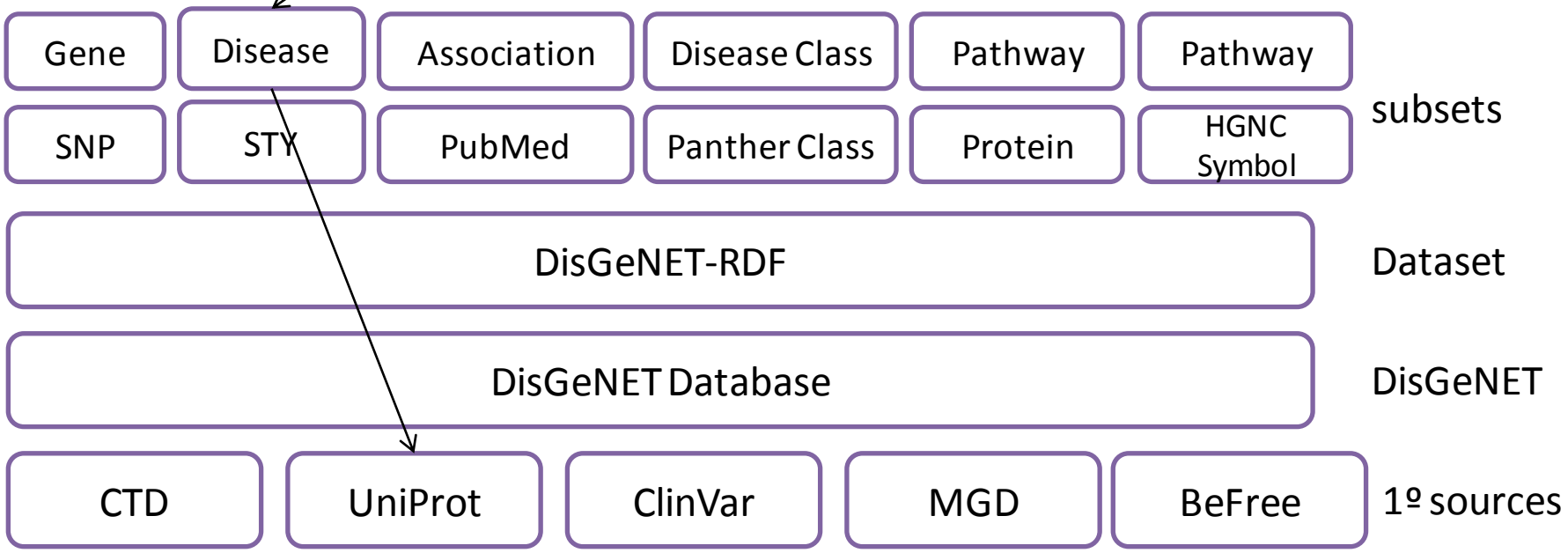
Data Model


Namespaces:
:
rdfs:
dcterms:
skos:
xsd:
ncit:
sio:
up:
foaf:
void:
dctypes:
dcat:
pav:
prov:
eco:
neuro:
Common description properties for the main concepts:
rdfs:label
rdfs:comment
dcterms:title
dcterms:identifier
void:inDataset
xsd:string
:subset
Note: Main concepts are highlighted by bold frames. Some properties have been omitted from the graph model representation for the sake of clarity.
Common description properties for concepts:
Resource
Literal
Linkout
http://rdf.disgenet.org/download/v3.0.0/DisGeNET-RDF-Example.ttl (Turtle)
'Assertion Method' (eco:ECO_0000217)
'DisGeNET Evidence' (eco:ECO_0000000)
'Source' (dctypes:Dataset, dcat:Distribution)
'Article' (ncit:C47902)
'Single Nucleotide Polymorphism' (ncit:C18279)
'Score' (ncit:C25338)
'GeneDiseaseAssociation' (sio:SIO_000983)
:geneDiseaseAssociation
CTD, DBpedia, SIDER, LinkedCT
Human Disease Ontology
MeSH
OMIM
Orphanet Rare Disease Ontology
DECIPHER
NCI
ICD9-CM
Human Phenotype Ontology
'Phenotype' (sio:SIO_010056)
'Disease, Disorder or Finding' (ncit:C7057)
'Gene' (ncit:C16612)
'Protein' (ncit:C17021)
'Gene Symbol' (ncit:C43568)
'Pathway' (ncit:C20633)
'Panther Class' (rdfs:Class)
'MeSH Disease Class' (rdfs:Class)
'UMLS Semantic Type' (rdfs:Class)
'Homo sapiens' (up:Taxon)
URL
UMLS
NCBO BioPortal
PubMed, PubChem
dbSNP
UniProt, Swiss-Prot
HGNC symbol
REACTOME
Entrez, CTD, PubChem, Diseasome, GeneWiki, GeneCards, GeneAtlas
URL Monarch, Decipher, COSMIC, Ensembl, SNPedia, BioGPS, DGA, GoGene, ClinicalTrials
URL ClinVar, Ensembl, SNPedia
xsd:date
xsd:decimal

Metada Dataset Description

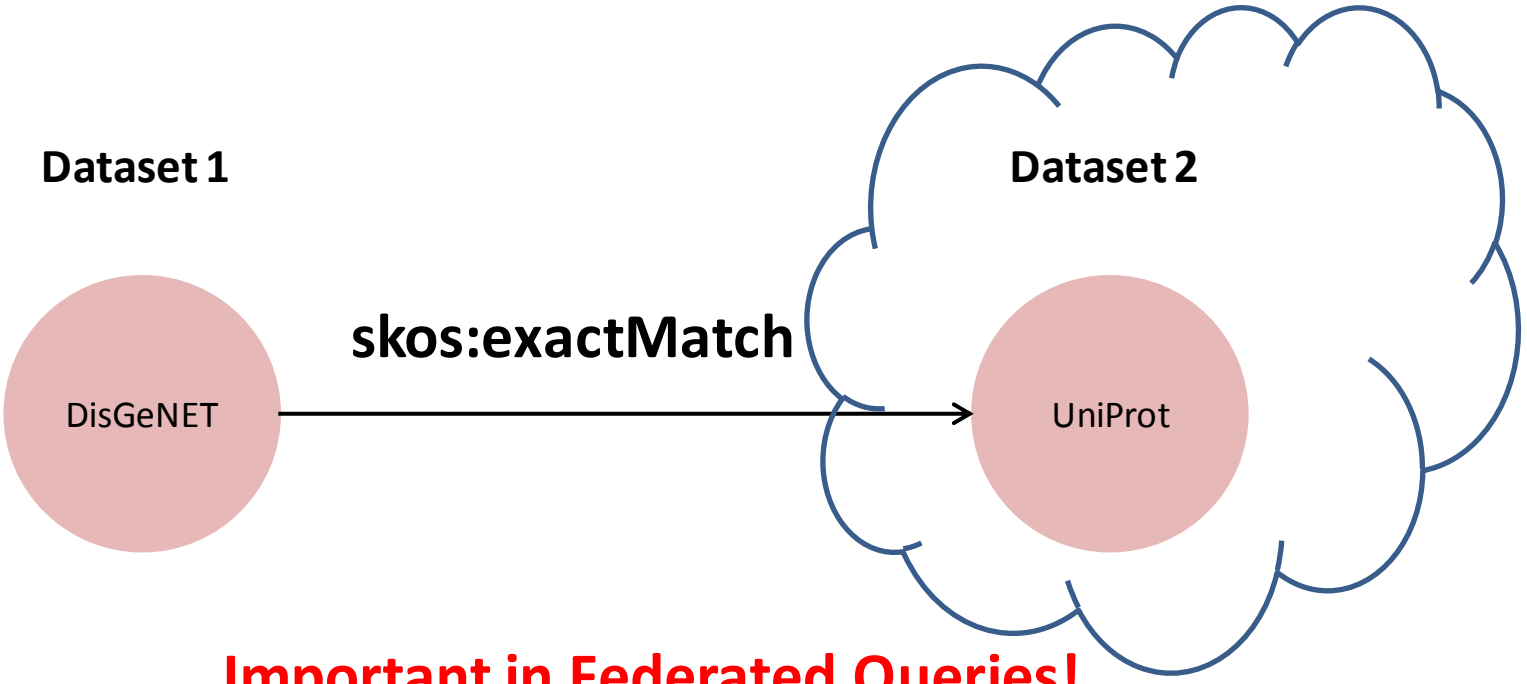
DisGeNET-RDF **VoID file (Vocabulary of Interlinked Datasets)**

```
<http://linkedlifedata.com/resource/umls/id/C0023264>
      a       ncit: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]" ;
      dcterms:identifier "umls:C0023264"^^xsd:string ;
      dcterms:title "Leigh Disease" ;
      void:inDataset <http://rdf.disgenet.org/v3.0.0/void/disease> ;
      sio:SIO_000008 <http://bioportal.bioontology.org/ontologies/umlssn.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

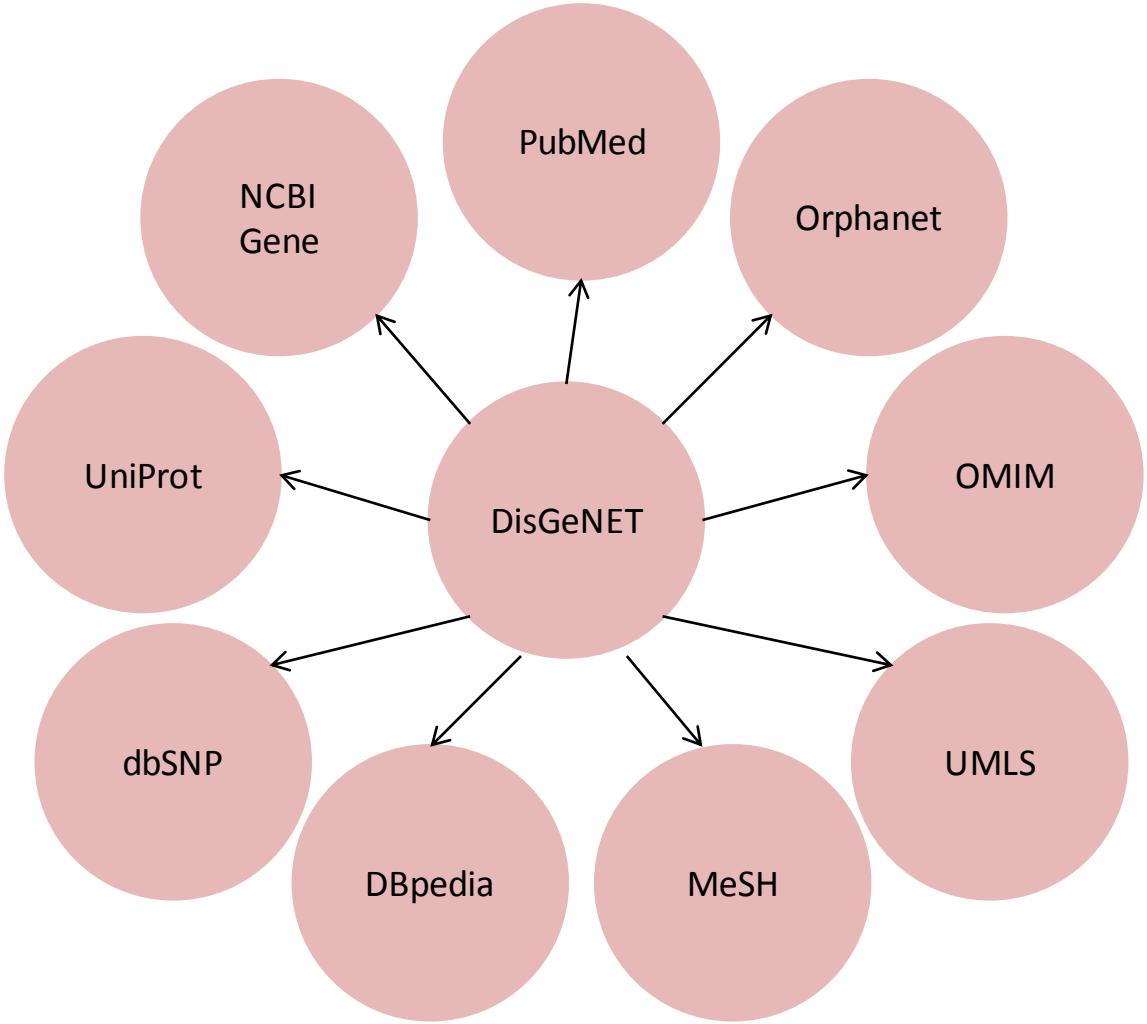


Important in Federated Queries!

Interlinking

?s skos:exactMatch ?o

Biomedical Databases and Disease Terminologies

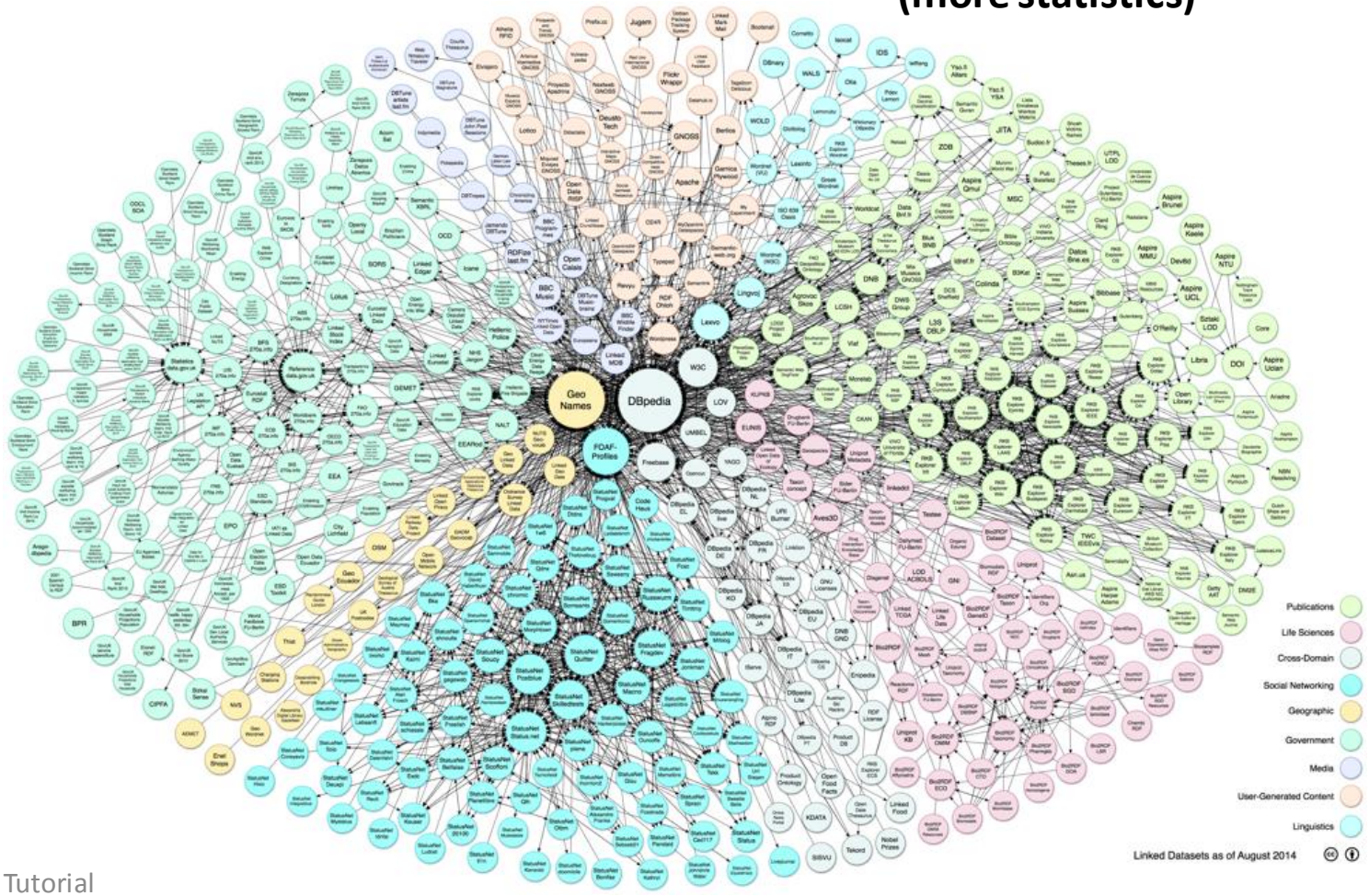


DisGeNET as Linked Open Data

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

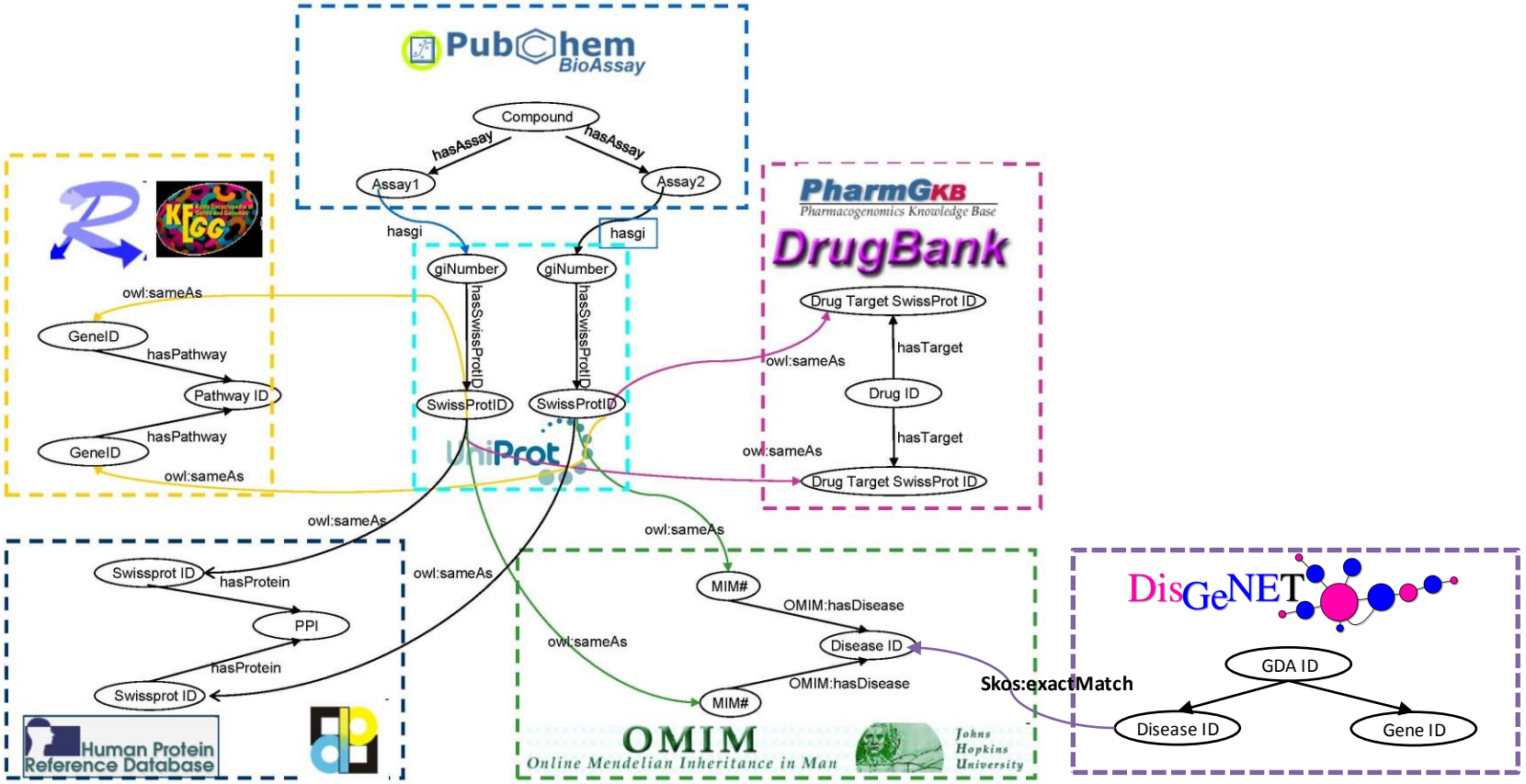
datahub
The easy way to get, use and share data

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



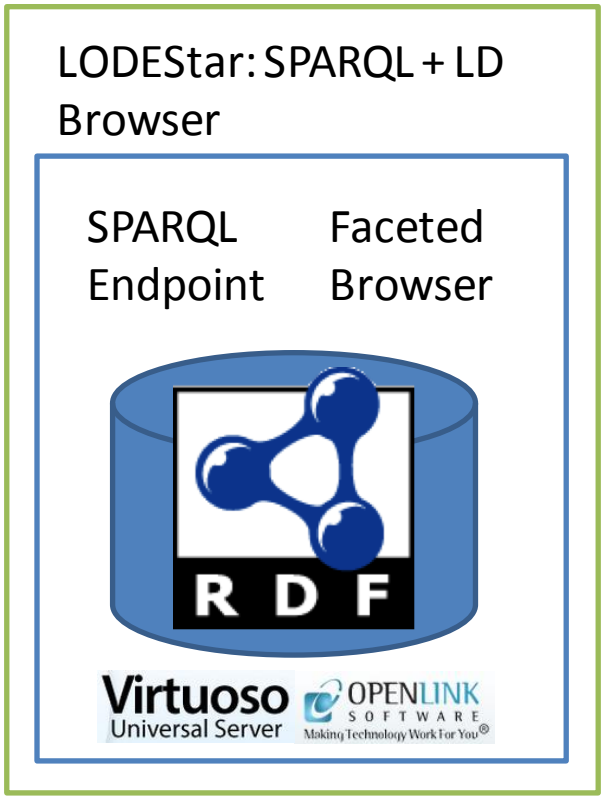
Federated Query Support

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



Implementation

- DisGeNET RDF data, VoID dataset description, and six OWL ontologies loaded into the RDF Store
- Total number of triples: 24,882,432 (8,5G)



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/1.5

Documentation

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

http://rdf.disgenet.org/

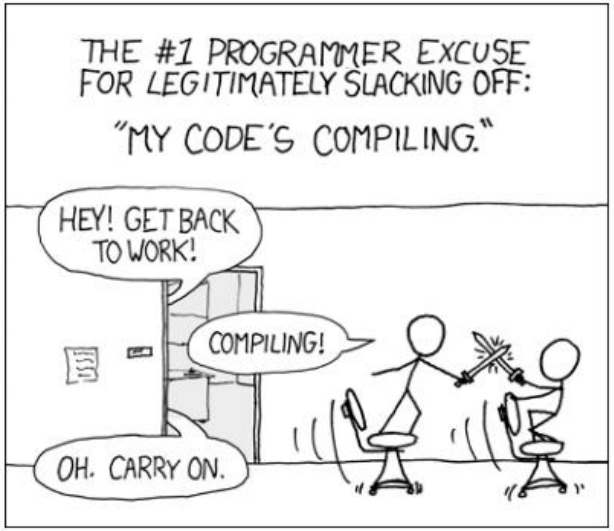
- Support @:

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
- Other approaches on development
  - Q/A based on natural language
  - Linked Data Fragments
  - ElasticSearch

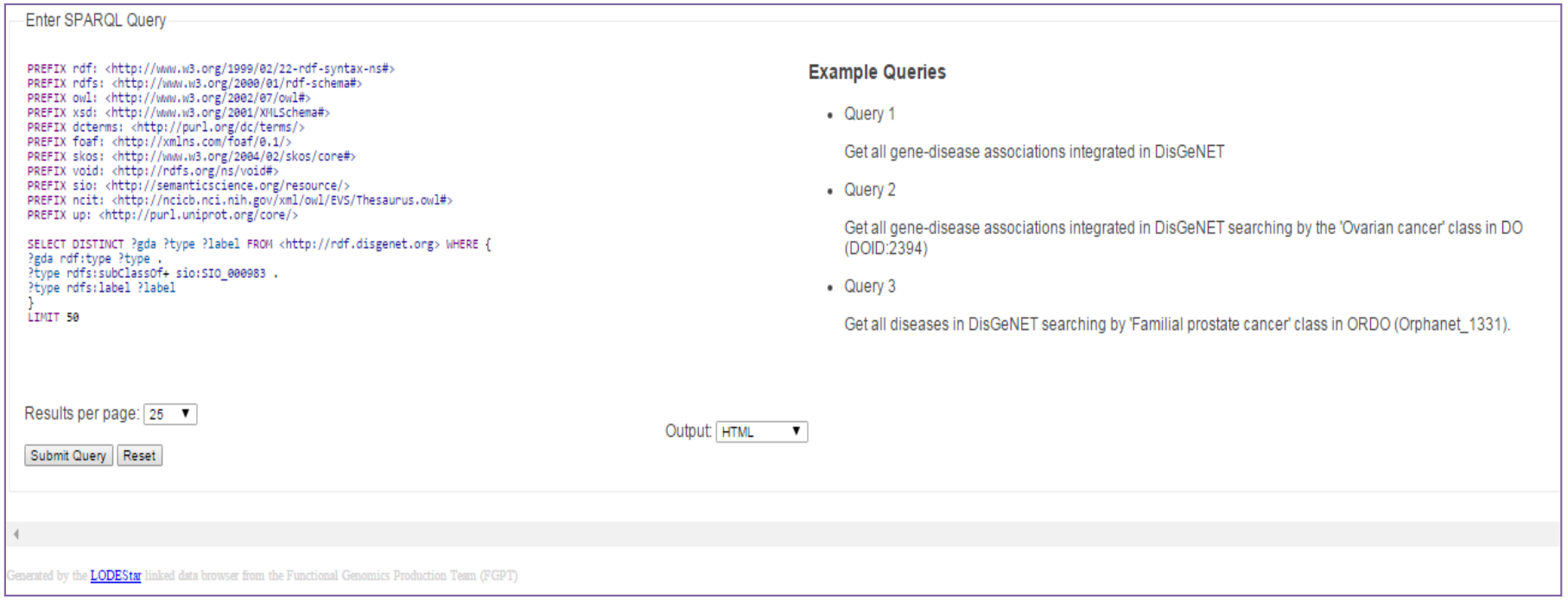


Querying DisGeNET

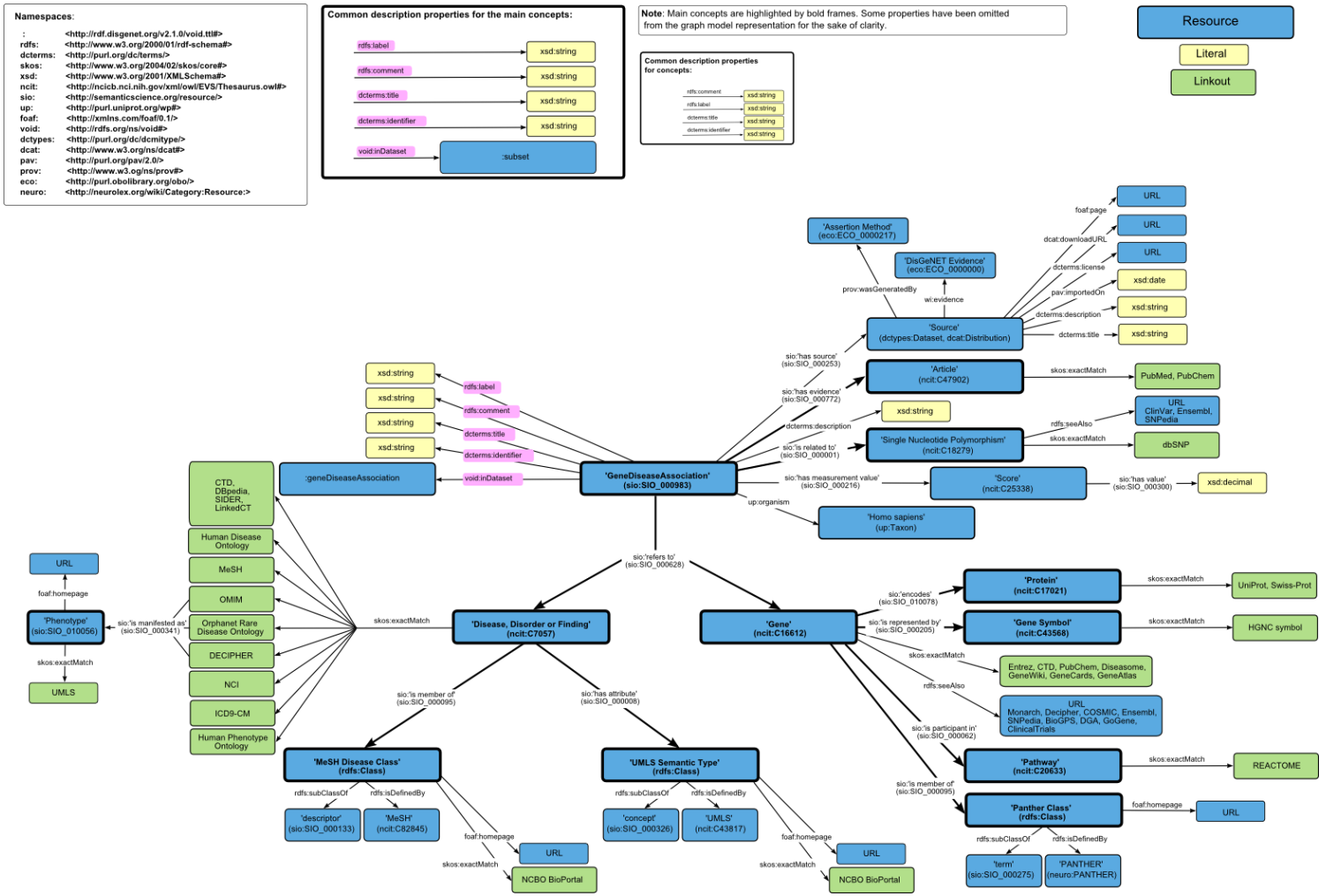
- **SPARQL Queries over DisGeNET data**

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

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



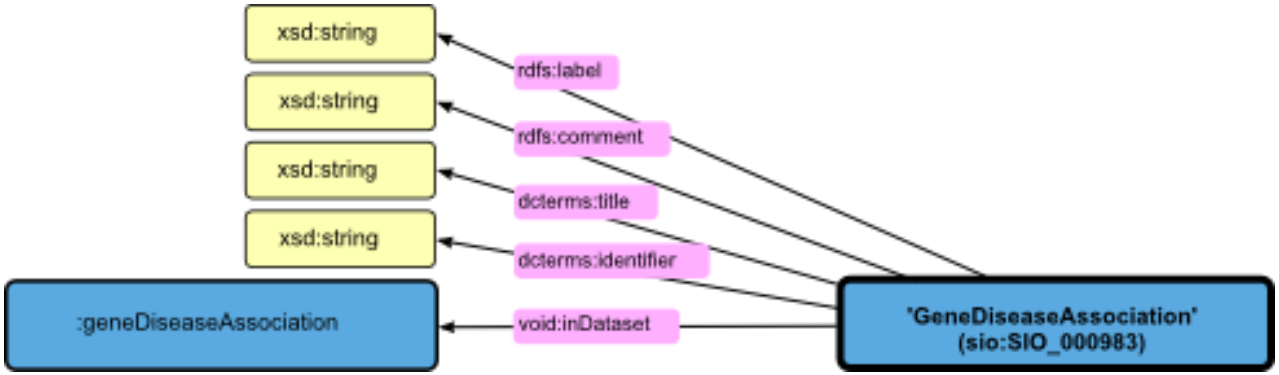
Data Model


Namespaces:
:
rdfs:
dcterms:
skos:
xsd:
ncit:
sio:
up:
foaf:
void:
dctypes:
dcat:
pav:
prov:
eco:
neuro:
Common description properties for the main concepts:
rdfs:label
rdfs:comment
dcterms:title
dcterms:identifier
void:inDataset
xsd:string
:subset
Note: Main concepts are highlighted by bold frames. Some properties have been omitted from the graph model representation for the sake of clarity.
Common description properties for concepts:
Resource
Literal
Linkout
'GeneDiseaseAssociation' (sio:SIO_000983)
:geneDiseaseAssociation
'Source' (dctypes:Dataset, dcat:Distribution)
'Assertion Method' (eco:ECO_0000217)
'DisGeNET Evidence' (eco:ECO_0000000)
'Article' (ncit:C47902)
'Single Nucleotide Polymorphism' (ncit:C18279)
'Score' (ncit:C25338)
'Homo sapiens' (up:Taxon)
CTD, DBpedia, SIDER, LinkedCT
Human Disease Ontology
MeSH
OMIM
Orphanet Rare Disease Ontology
DECIPHER
NCI
ICD9-CM
Human Phenotype Ontology
'Phenotype' (sio:SIO_010056)
URL
UMLS
'Disease, Disorder or Finding' (ncit:C7057)
'Gene' (ncit:C16612)
'Protein' (ncit:C17021)
'Gene Symbol' (ncit:C43568)
'Pathway' (ncit:C20633)
'Panther Class' (rdfs:Class)
'MeSH Disease Class' (rdfs:Class)
'UMLS Semantic Type' (rdfs:Class)
'descriptor' (sio:SIO_000133)
'MeSH' (ncit:C82845)
'concept' (sio:SIO_000326)
'UMLS' (ncit:C43817)
'term' (sio:SIO_000275)
'PANTHER' (neuro:PANTHER)
NCBO BioPortal
PubMed, PubChem
URL ClinVar, Ensembl, SNPedia
dbSNP
UniProt, Swiss-Prot
HGNC symbol
Entrez, CTD, PubChem, Diseasome, GeneWiki, GeneCards, GeneAtlas
URL Monarch, Decipher, COSMIC, Ensembl, SNPedia, BioGPS, DGA, GoGene, ClinicalTrials
REACTOME

Querying DisGeNET

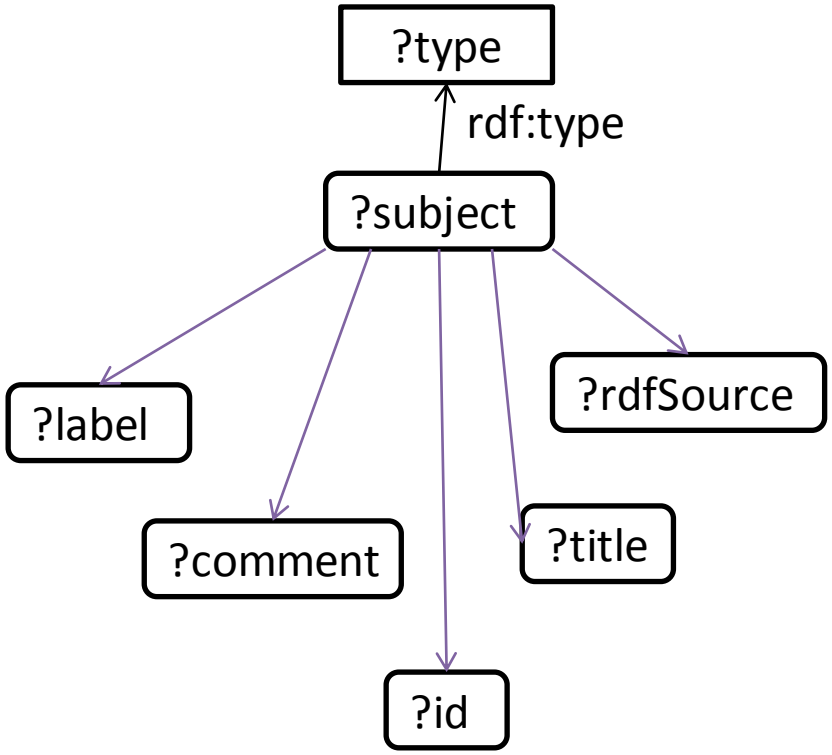
- **SPARQL Queries over DisGeNET data**
- *Minimal Resource Description* Graph

- 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 ncit: <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: 25

Submit Query | Reset

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).

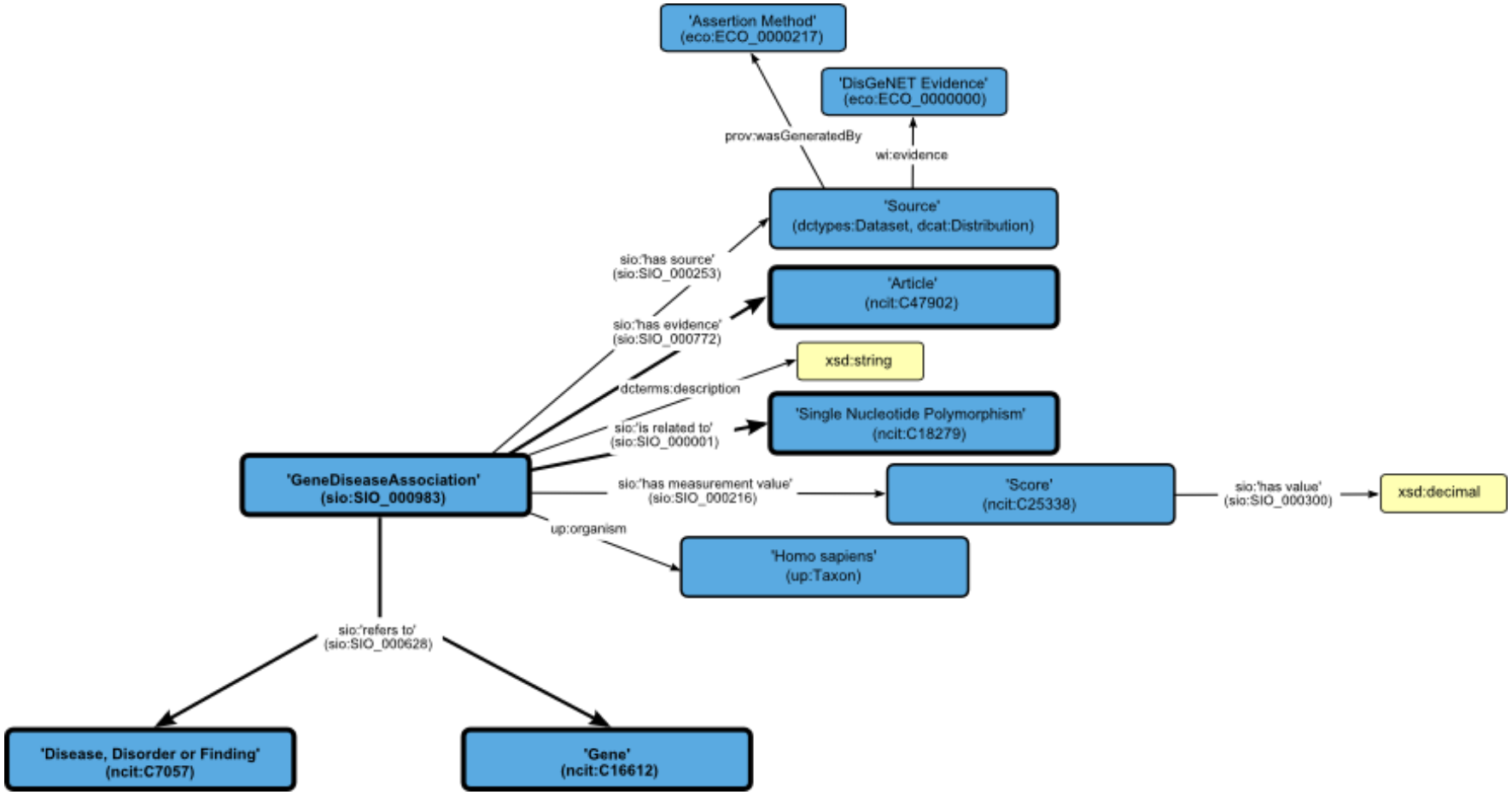
Output: HTML

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>

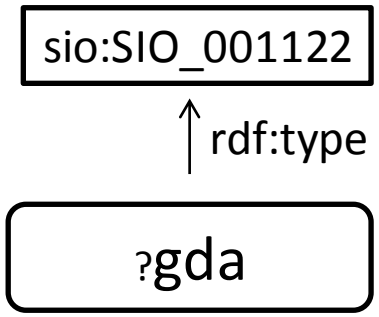
Querying DisGeNET

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



Querying DisGeNET

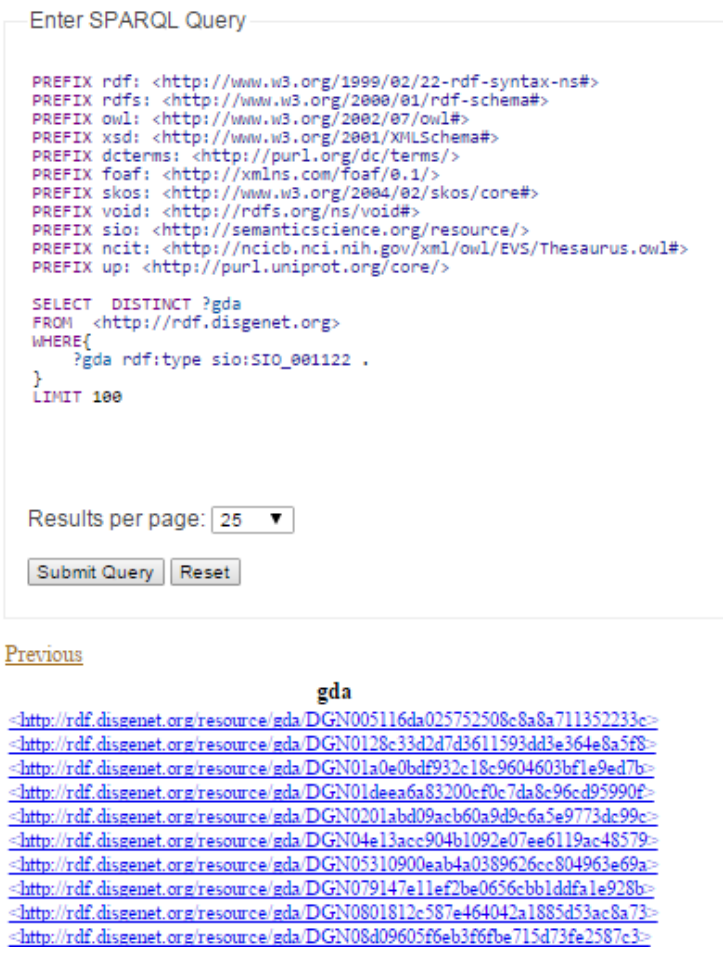
- **SPARQL Queries over DisGeNET data**
- *Gene-Disease Association* Graph




```
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




```
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

  - 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

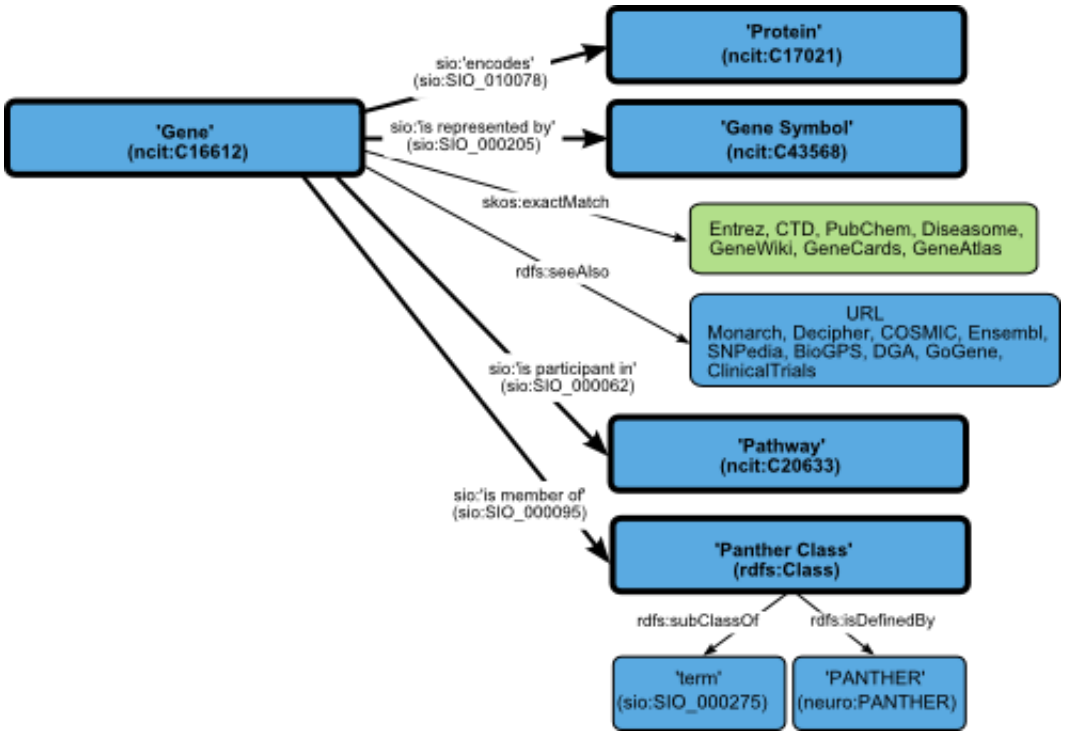
- **SPARQL Queries over DisGeNET data**
- *Gene-Disease Association* Graph

  - 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
```

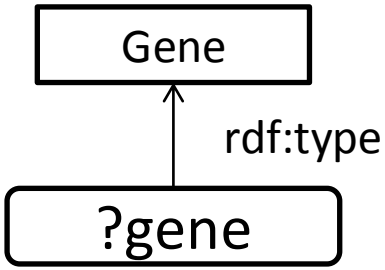
Querying DisGeNET

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



Querying DisGeNET

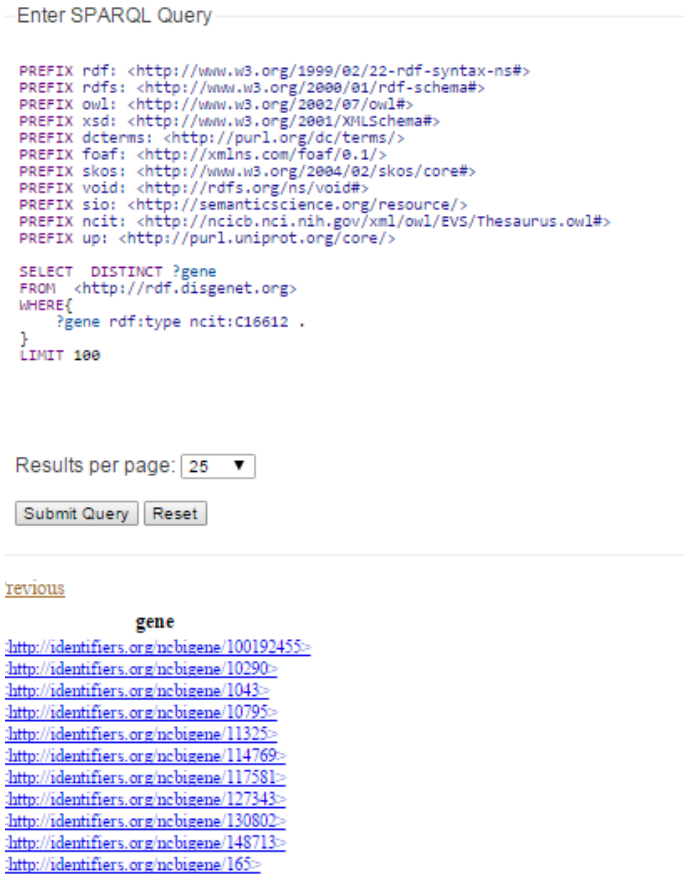
- **SPARQL Queries over DisGeNET data**
- *Gene* Graph




```
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




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

Querying DisGeNET

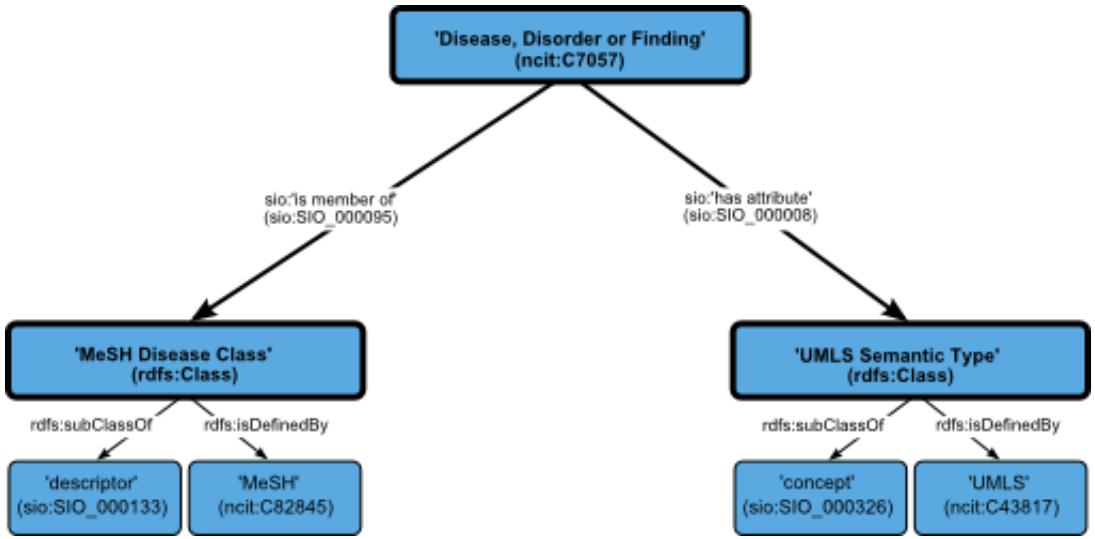
- **SPARQL Queries over DisGeNET data**
- *Gene* Graph

  - 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

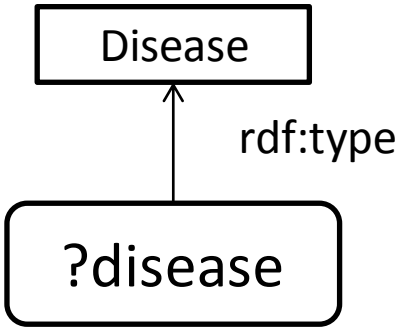
Querying DisGeNET

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



Querying DisGeNET

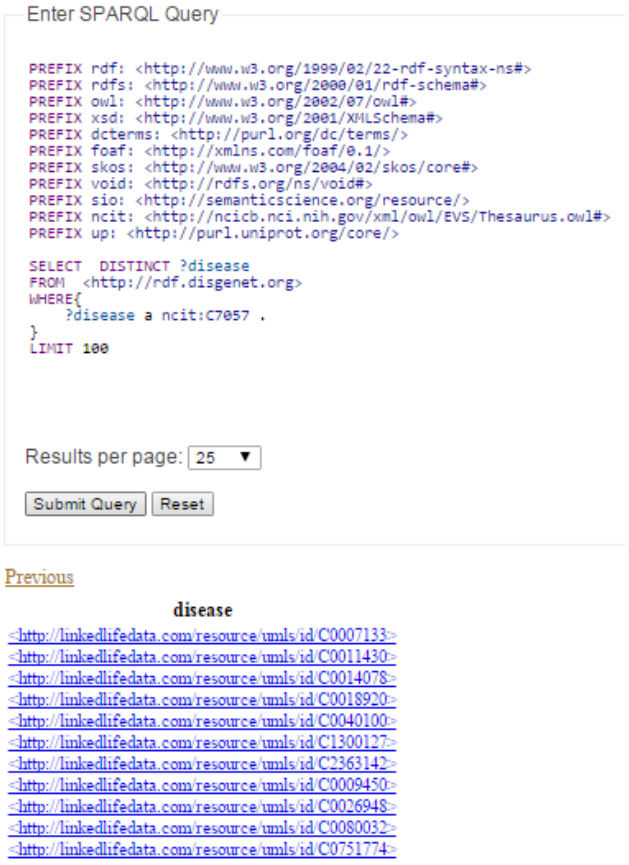
- **SPARQL Queries over DisGeNET data**
- *Disease* Graph




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

Querying DisGeNET

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




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

Querying DisGeNET

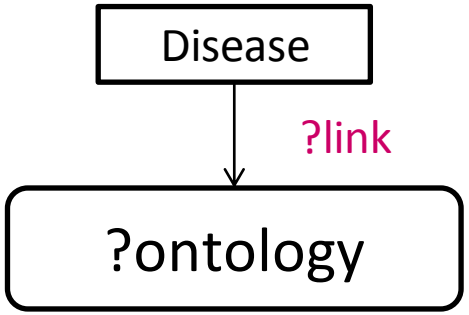
- **SPARQL Queries over DisGeNET data**
- *Disease* Graph

- 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




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

COVERAGE

Ontology	UMLS	MeSH	OMIM	NCI	DO	ORDO	ICD9CM	HPO	DECIPHER
% DisGeNET	100	58	38	33	19	13	12	9	0.4

Querying DisGeNET

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

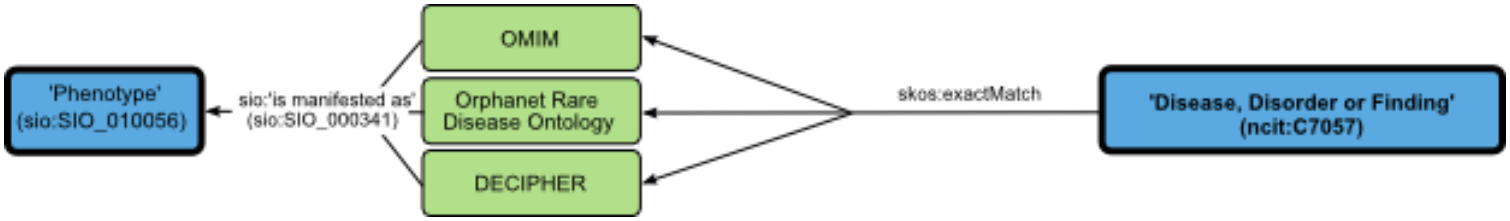
```
?child  rdfs:subClassOf+  ?parent
```

- Ontologies loaded in our RDF triple store: SIO, DO, ORDO, NCIT, HPO, 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_000341 ?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_000341 ?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_000341 ?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)*

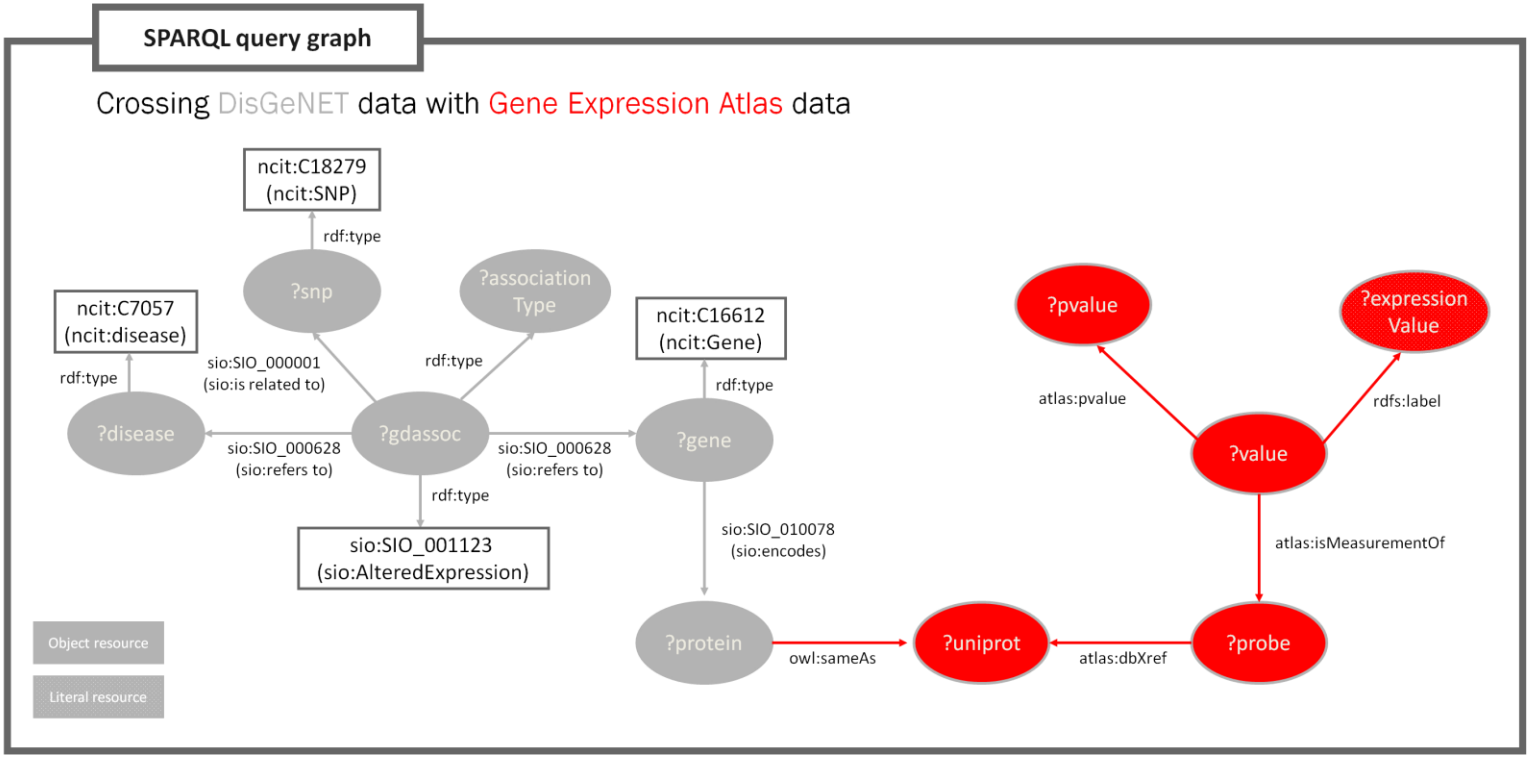
  - How many diseases are associated with a phenotype

```
SELECT DISTINCT ?phenotype ?phenotypeName count(distinct ?disease) as
?diseases
WHERE {
      ?hpdisease sio:SIO_000341 ?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 hitted 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**)

Acknowledgments

IBI Group

Alba Gutiérrez-Sacristán

Àlex Bravo

Janet Piñero

Núria Queralt Rosinach

Alexia Giannoula

Miguel A. Mayer

Laura I. Furlong

Ferran Sanz

Thanks for your attention!
Questions are welcome