PubChemRDF Tutorial

Presenter: Gang Fu, Evan Bolton
National Center for Biotechnology Information (NCBI)
National Library of Medicine (NLM)
National Institutes of Health (NIH)
2015 SWAT4LS

PubChem

OUTLINE

PubChem

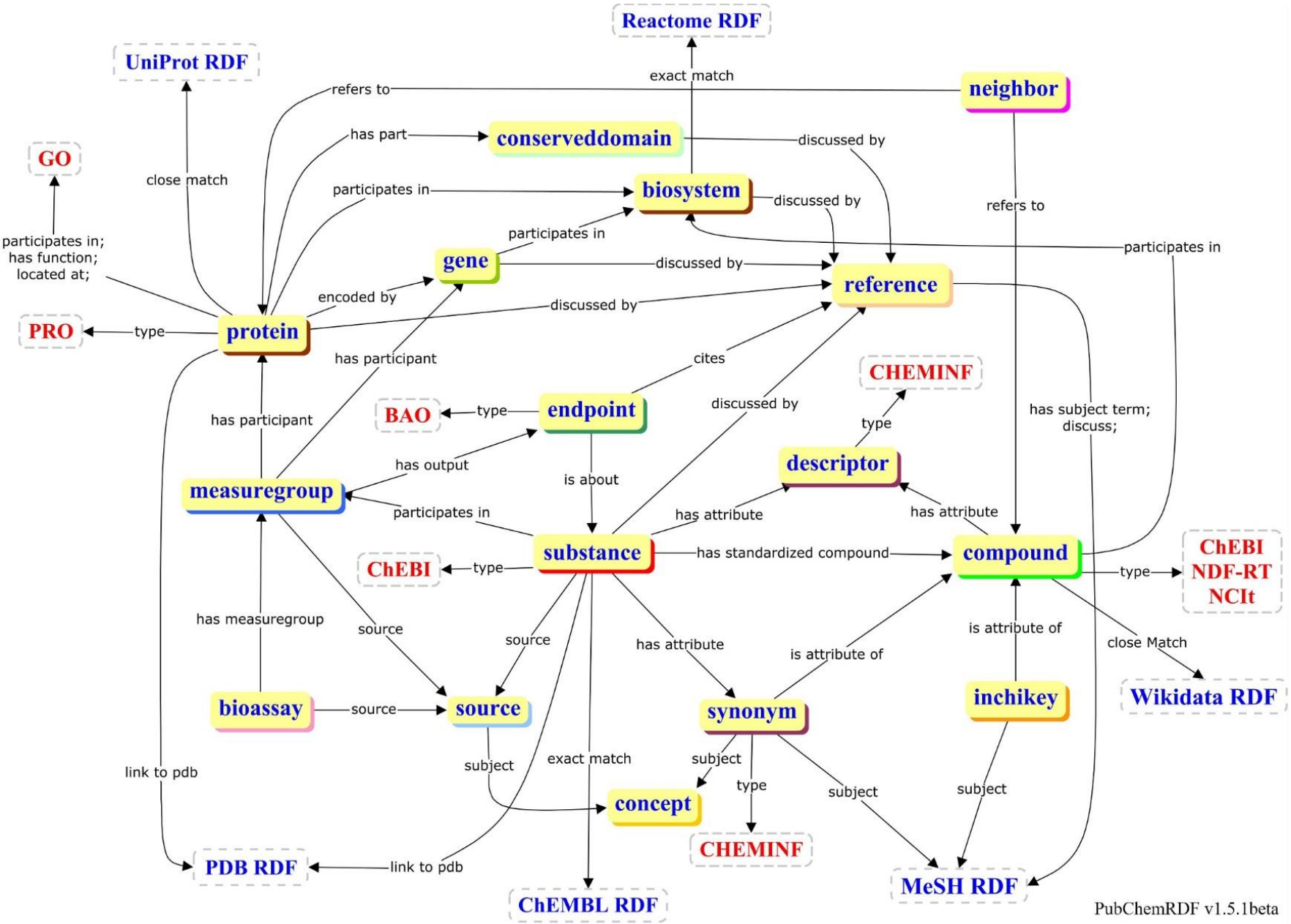
- **How the PubChemRDF is formulated?**
  - PubChemRDF URI schemes
  - PubChemRDF subdomains
  - Ontology-based data integration
- **How to Access the Data?**
- **How to answer scientific questions?**

PubChemRDF Subdomains

PubChem

Prefix	Namespace
compound	http://rdf.ncbi.nlm.nih.gov/pubchem/compound/
substance	http://rdf.ncbi.nlm.nih.gov/pubchem/substance/
descr	http://rdf.ncbi.nlm.nih.gov/pubchem/descriptor/
inchikey	http://rdf.ncbi.nlm.nih.gov/pubchem/inchikey/
syno	http://rdf.ncbi.nlm.nih.gov/pubchem/synonym/
bioassay	http://rdf.ncbi.nlm.nih.gov/pubchem/bioassay/
measuregroup	http://rdf.ncbi.nlm.nih.gov/pubchem/measuregroup/
endpoint	http://rdf.ncbi.nlm.nih.gov/pubchem/endpoint/
protein	http://rdf.ncbi.nlm.nih.gov/pubchem/protein/
conserveddomain	http://rdf.ncbi.nlm.nih.gov/pubchem/conserveddomain/
biosystem	http://rdf.ncbi.nlm.nih.gov/pubchem/biosystem/
gene	http://rdf.ncbi.nlm.nih.gov/pubchem/gene/
reference	http://rdf.ncbi.nlm.nih.gov/pubchem/reference/
nbr[a]	http://rdf.ncbi.nlm.nih.gov/pubchem/neighbor/
source	http://rdf.ncbi.nlm.nih.gov/pubchem/source/
concept	http://rdf.ncbi.nlm.nih.gov/pubchem/concept/
vocab	http://rdf.ncbi.nlm.nih.gov/pubchem/vocabulary#

PubChemRDF Overview



PubChemRDF URIs

http://rdf.ncbi.nlm.nih.gov/pubchem/compound/CID60823

http://rdf.ncbi.nlm.nih.gov/pubchem/substance/SID103554720

http://rdf.ncbi.nlm.nih.gov/pubchem/bioassay/AID1788

http://rdf.ncbi.nlm.nih.gov/pubchem/measuregroup/AID447528

http://rdf.ncbi.nlm.nih.gov/pubchem/protein/GI124375976

http://rdf.ncbi.nlm.nih.gov/pubchem/conserveddomain/PSSMID132758

http://rdf.ncbi.nlm.nih.gov/pubchem/gene/GID367

http://rdf.ncbi.nlm.nih.gov/pubchem/biosystem/BSID82991

http://rdf.ncbi.nlm.nih.gov/pubchem/reference/PMID10395478

Practice

Pick any one and put in your browser

http://rdf.ncbi.nlm.nih.gov/pubchem/inchikey/XUKUURHRXDUEBC-KAYWLYCHSA-N *Md5 hash calculated based on lower case*

http://rdf.ncbi.nlm.nih.gov/pubchem/synonym/MD5_9a05646d461669f86de312d88ab5748a

http://rdf.ncbi.nlm.nih.gov/pubchem/concept/ATC_L01XE

http://rdf.ncbi.nlm.nih.gov/pubchem/source/ChEMBL

Replace "," with ""
Replace "." with ""
Replace "&" with "-"
Replace "/" with "-"
Replace " " with "_"

Question: How to retrieve all the PubChem depositors?

PubChemRDF Composite URIs

PubChem

Appendix Table 2

http://rdf.ncbi.nlm.nih.gov/pubchem/descriptor/CID60823_Molecular_Weight

http://rdf.ncbi.nlm.nih.gov/pubchem/measuregroup/AID1788_1

http://rdf.ncbi.nlm.nih.gov/pubchem/measuregroup/AID363_PMID16161995

http://rdf.ncbi.nlm.nih.gov/pubchem/endpoint/SID103164874_AID443491

http://rdf.ncbi.nlm.nih.gov/pubchem/endpoint/SID99445338_AID2202_1

http://rdf.ncbi.nlm.nih.gov/pubchem/endpoint/SID8033500_AID363_PMID10395478

http://rdf.ncbi.nlm.nih.gov/pubchem/protein/GI2506129GI254763435

Question: How to retrieve all the protein complexes?

http://rdf.ncbi.nlm.nih.gov/pubchem/neighbor/CID60823_CID68019409_2DSimilarity

http://rdf.ncbi.nlm.nih.gov/pubchem/neighbor/CID60823_CID68019409_2DTanimotoScore

http://rdf.ncbi.nlm.nih.gov/pubchem/neighbor/CID60823_CID11330946_3DSimilarity

http://rdf.ncbi.nlm.nih.gov/pubchem/neighbor/CID60823_CID11330946_3DShapeTanimotoScore

http://rdf.ncbi.nlm.nih.gov/pubchem/neighbor/CID60823_CID11330946_3DFeatureTanimotoScore

Practice

Pick any one and put in your browser

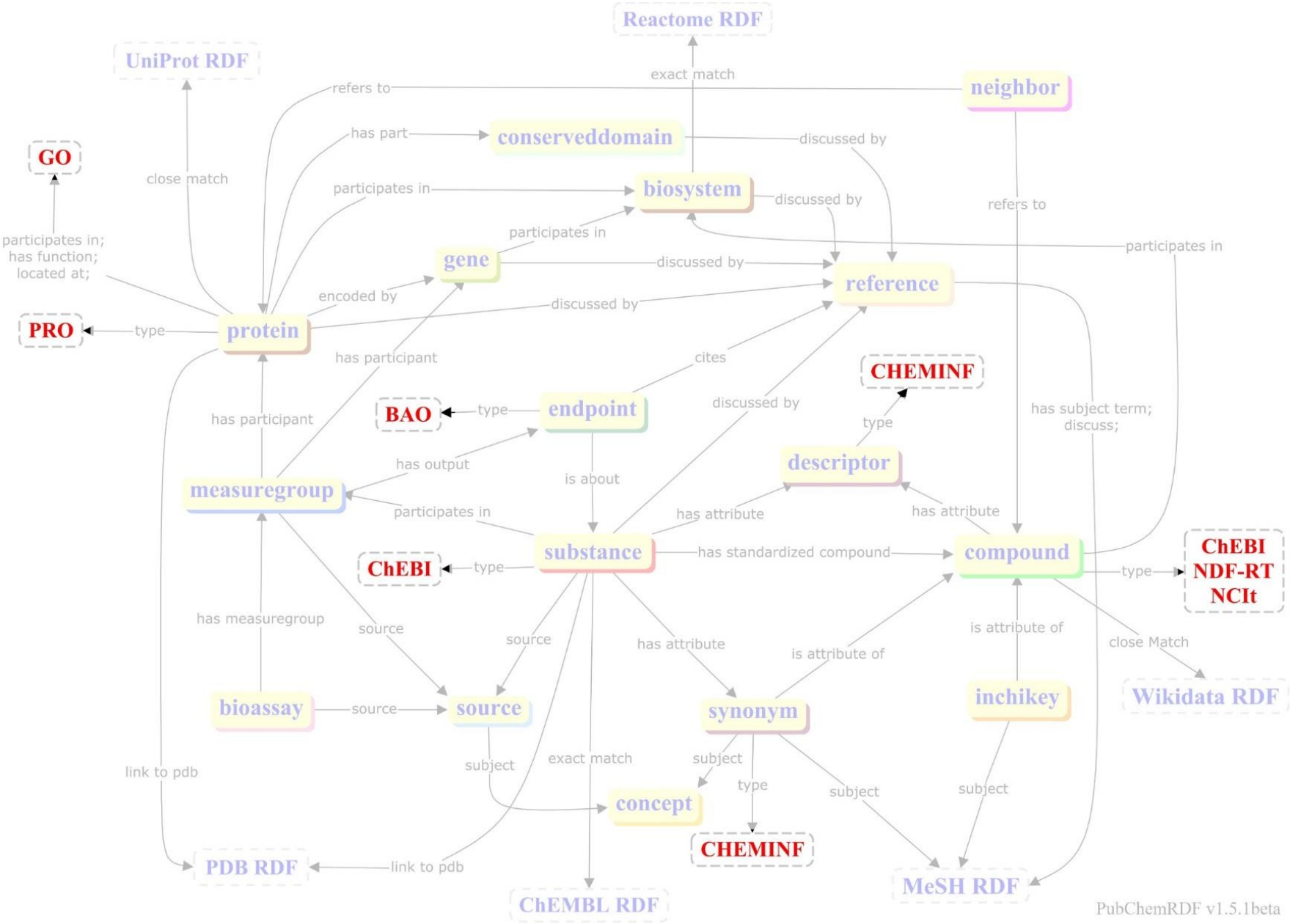
PubChemRDF Ontologies

PubChem

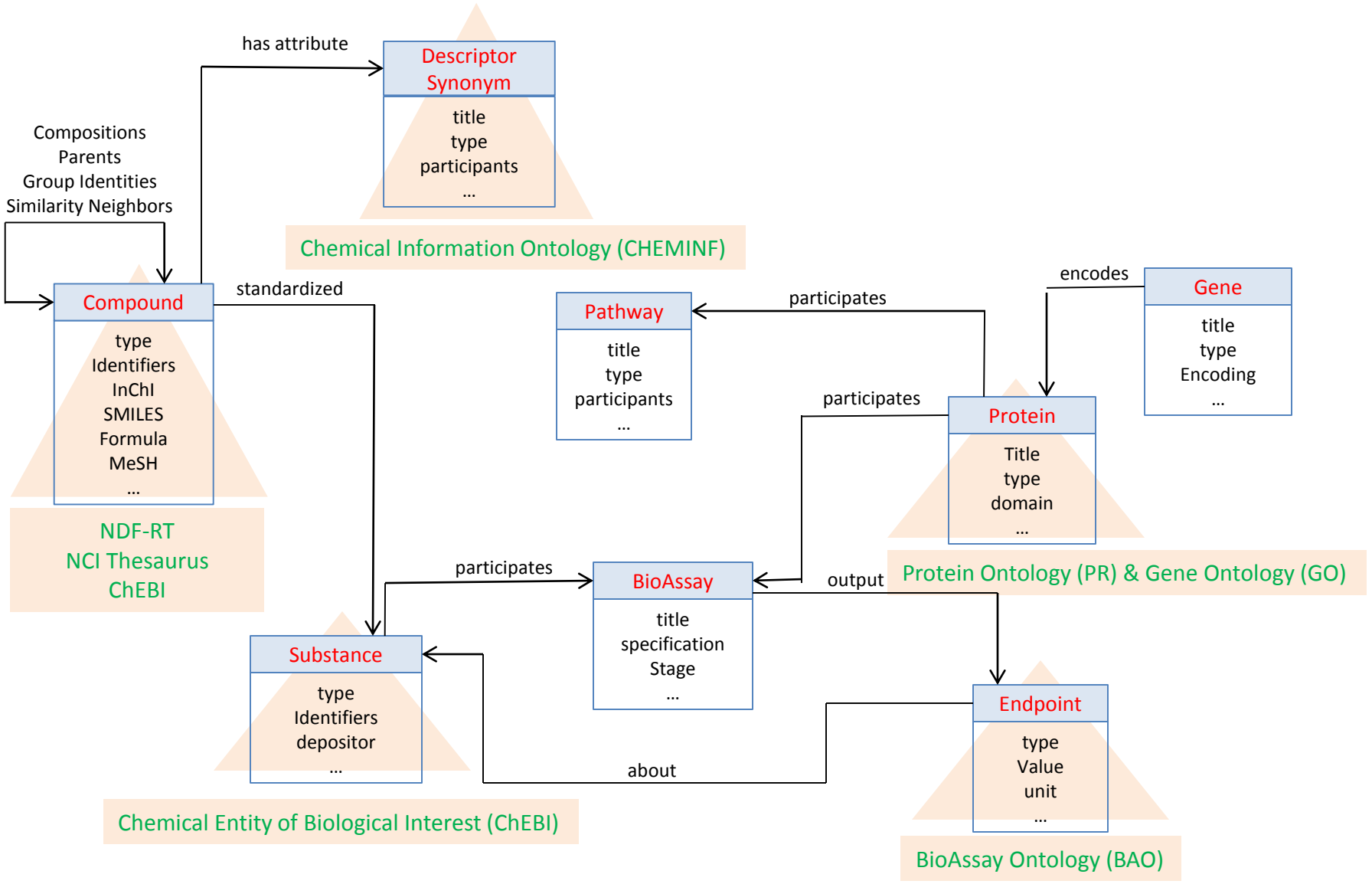
Prefix	Namespace	Vocabularies
rdfs	http://www.w3.org/2000/01/rdf-schema#	RDF Schema
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	RDF
owl	http://www.w3.org/2002/07/owl#	OWL
xsd	http://www.w3.org/2001/XMLSchema#	XML Schema
ndfrt	http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#	NDF-RT
ncit	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#	NCIt
sio[a]	http://semanticscience.org/resource/	SIO
cheminf[a]	http://semanticscience.org/resource/	CHEMINF
skos	http://www.w3.org/2004/02/skos/core#	SKOS
obo	http://purl.obolibrary.org/obo/	BFO, OBI, IAO, UO, ChEBI, PR, GO
bao	http://www.bioassayontology.org/bao#	BAO
bp	http://www.biopax.org/release/biopax-level3.owl#	BioPAX
cito	http://purl.org/spar/cito/	CiTO
fabio	http://purl.org/spar/fabio/	FaBio
pdbo	http://rdf.wwpdb.org/schema/pdbx-v40.owl#	PDBo
dcterms	http://purl.org/dc/terms/	DCMI Terms
pav	http://purl.org/pav/	PAV
foaf	http://xmlns.com/foaf/0.1/	FOAF Vocabulary

Hierarchical Classifications

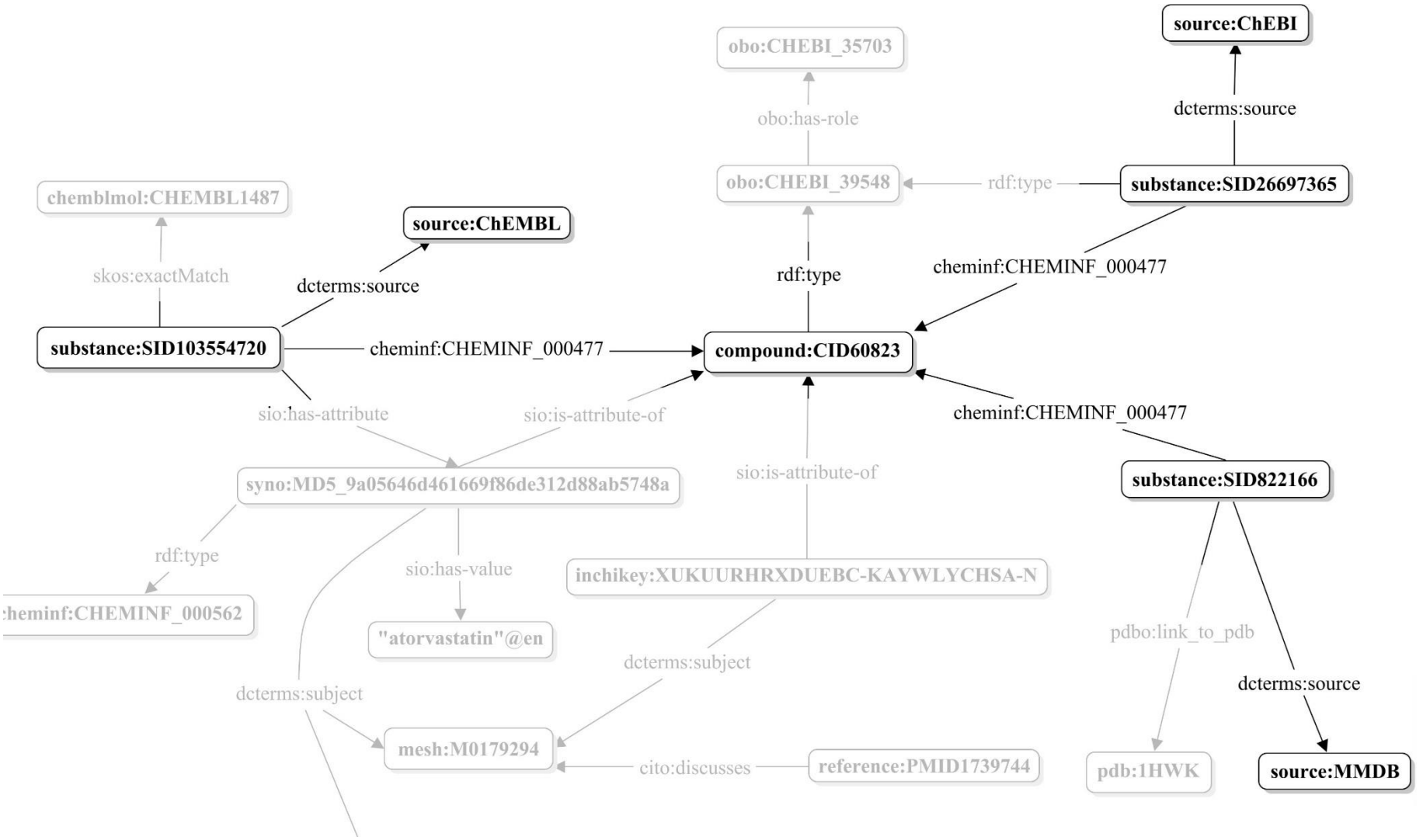
PubChemRDF Overview



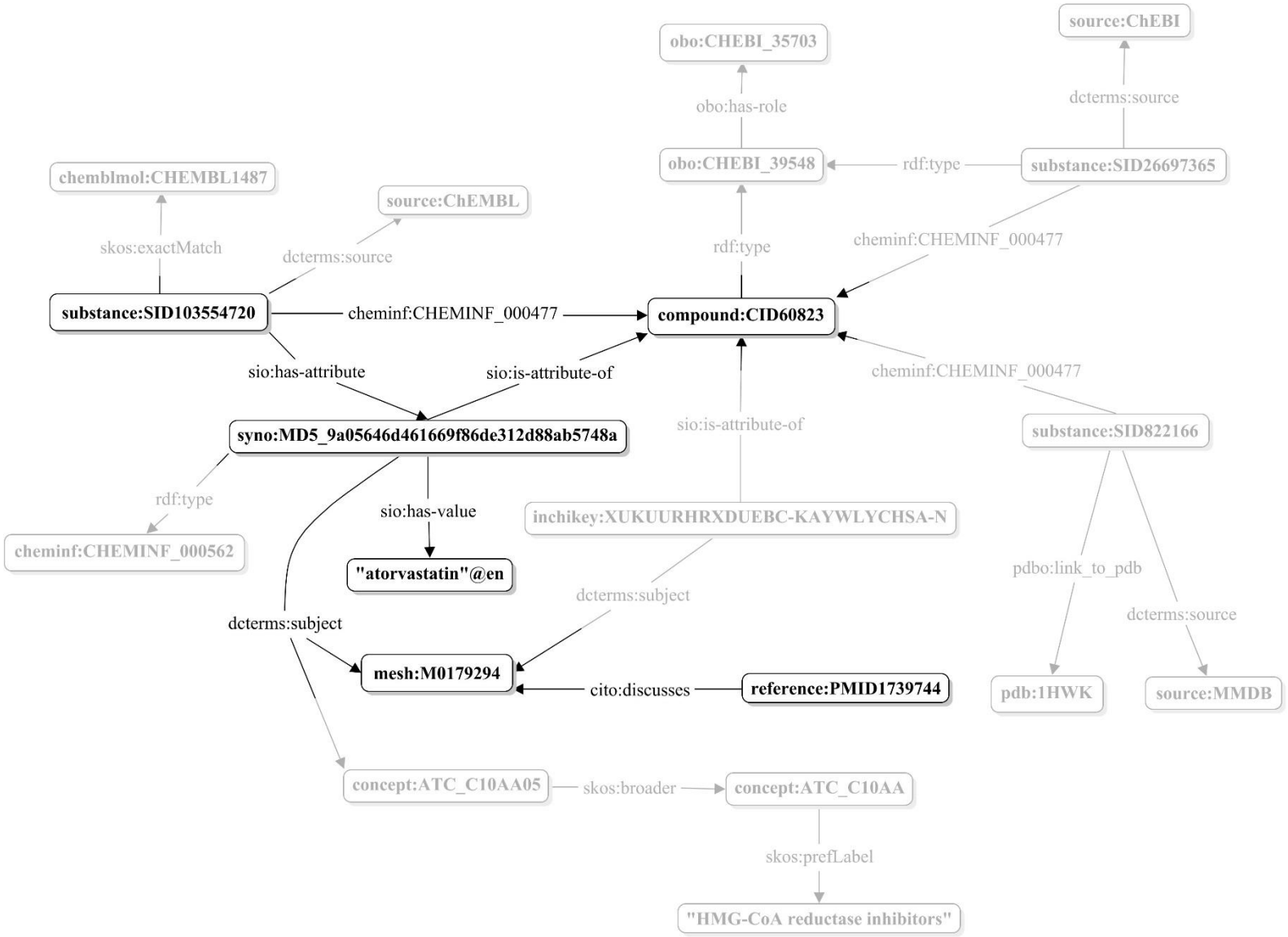
PubChemRDF Overview: ontology-based data integration



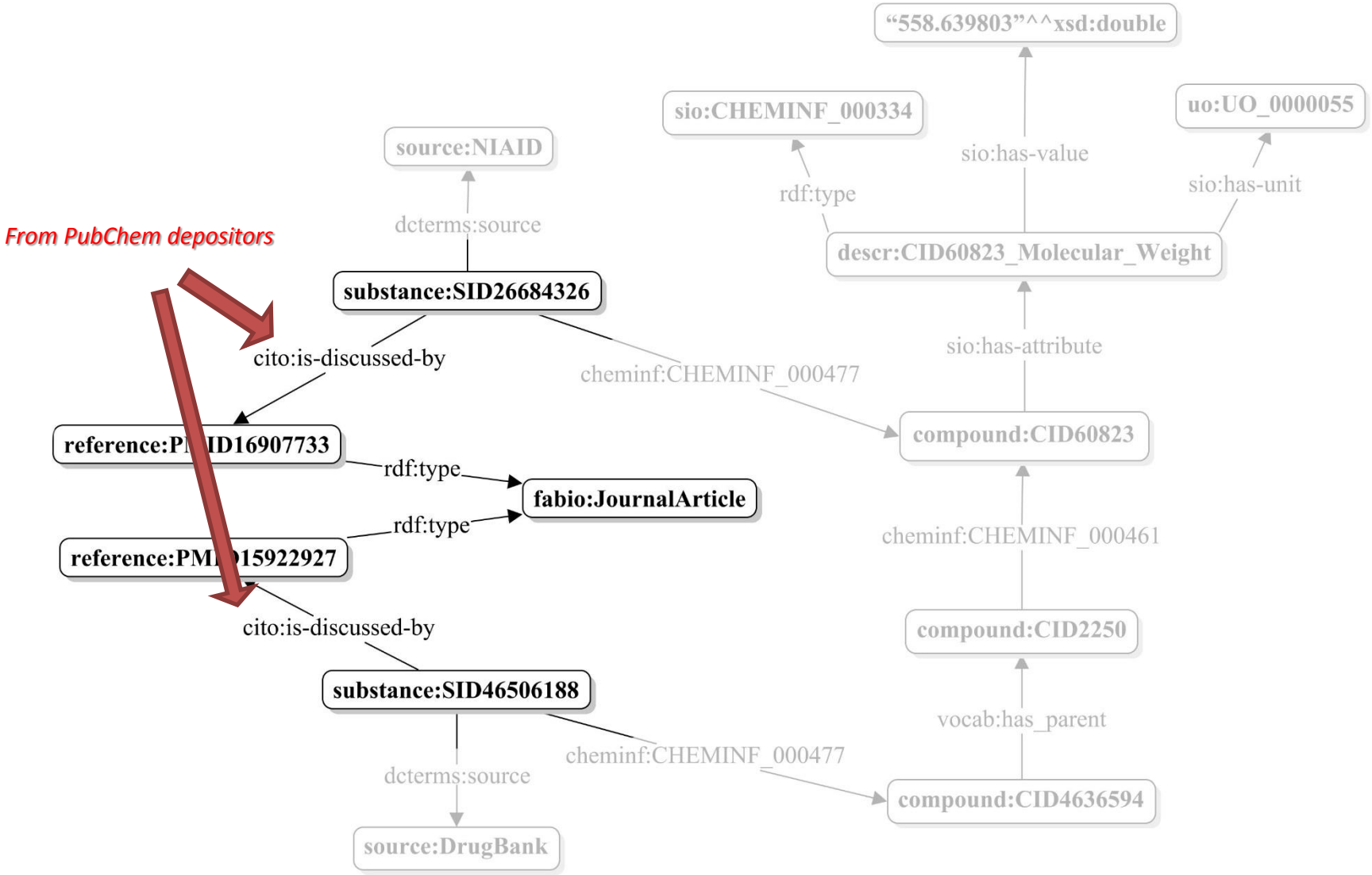
PubChemRDF Graph 1: compound aggregates substances from different sources



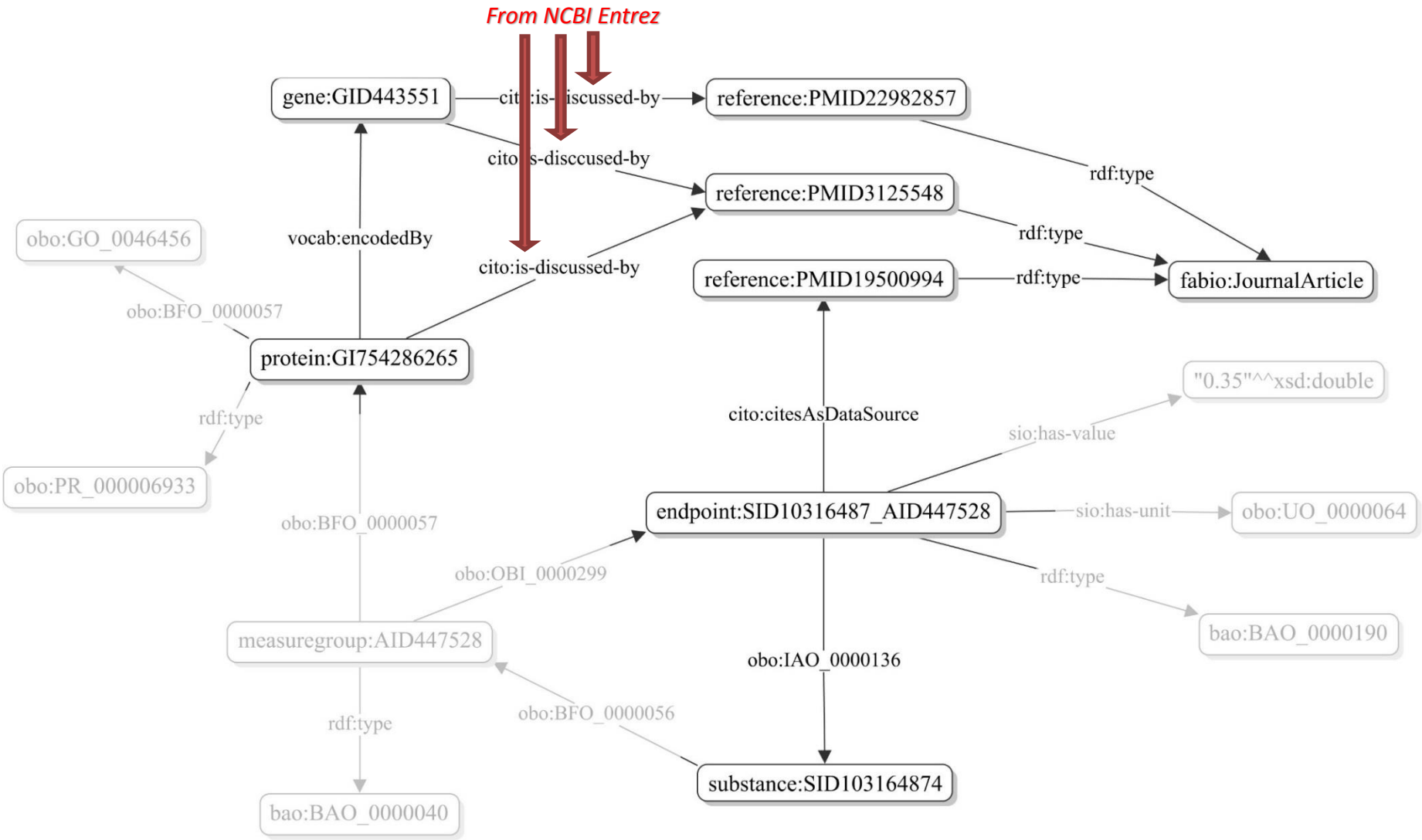
PubChemRDF Graph 2: synonym and MeSH annotations



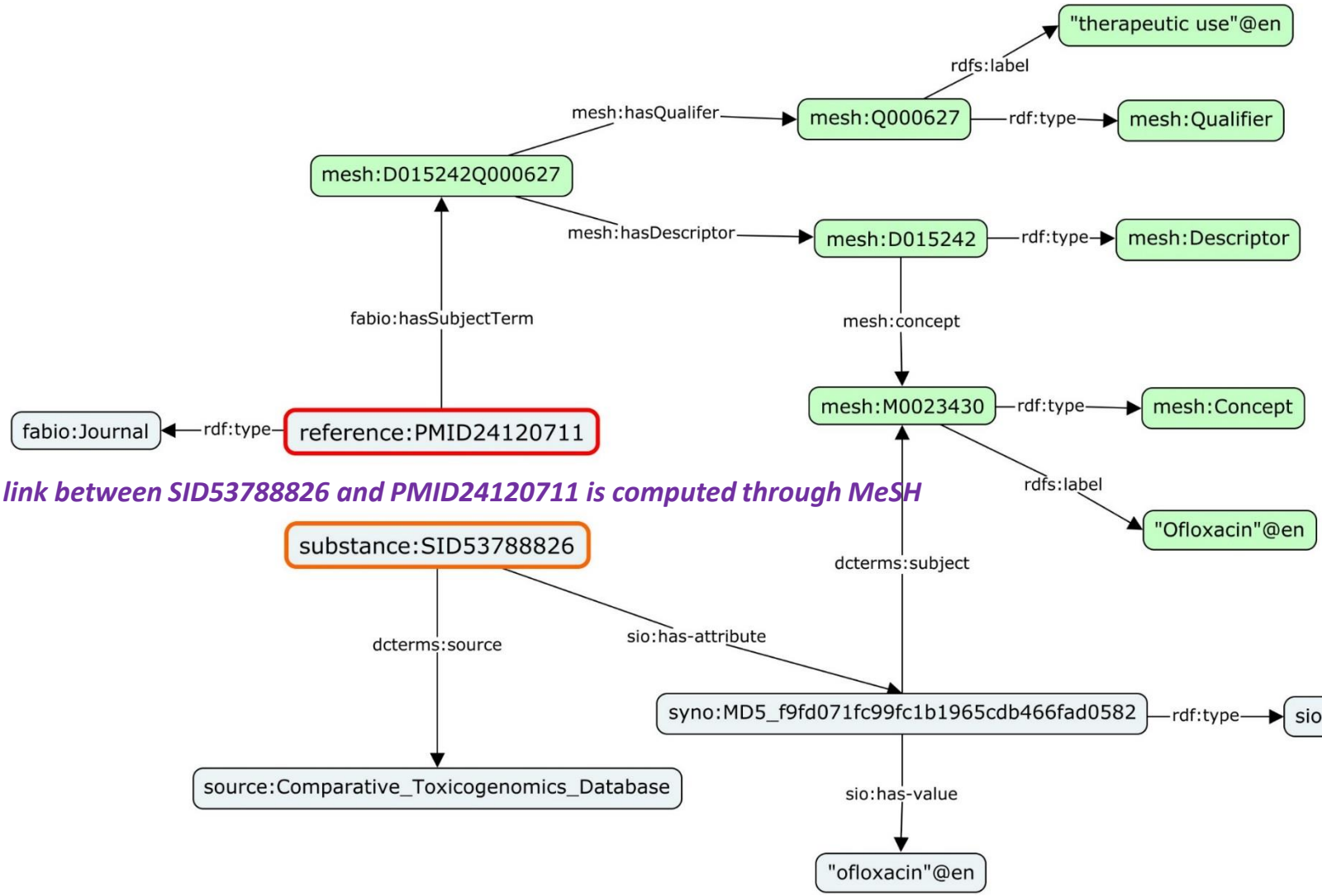
PubChemRDF Graph 3: depositor-provided references for substances



PubChemRDF Graph 4: references for proteins, genes, bioassays, and so on



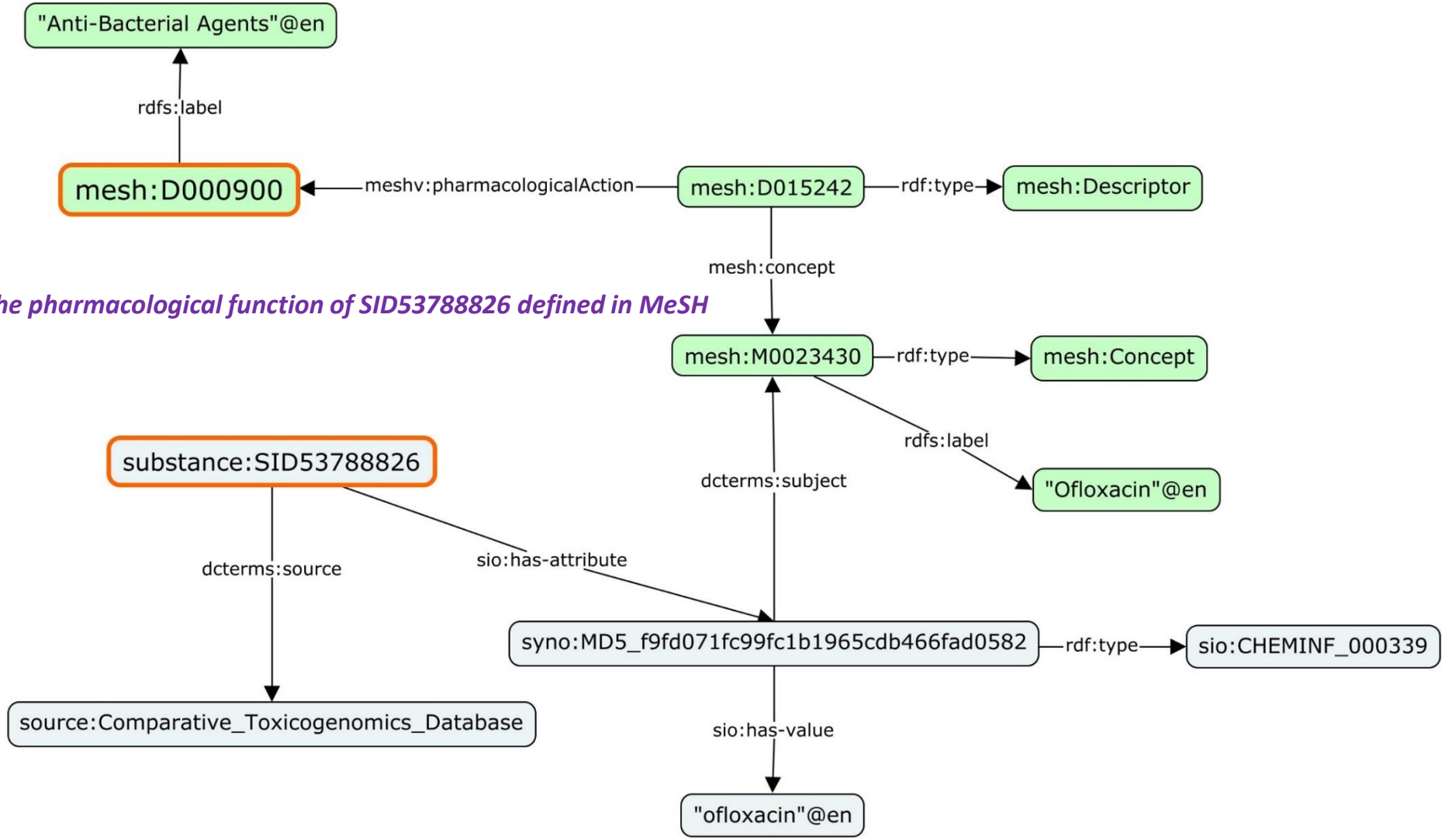
PubChemRDF Graph 5: substance link to reference through MeSH



The link between SID53788826 and PMID24120711 is computed through MeSH

PubChemRDF Graph 5: substance link to pharmacological action through MeSH

The pharmacological function of SID53788826 defined in MeSH



OUTLINE

PubChem

- How the PubChemRDF is formulated?
- **How to Access the Data?**
  - Programmatic access REST interface
  - Bulk download from FTP site layer by layer
- How to answer scientific questions?

PubChemRDF REST API: Content Negotiation

PubChem

MIME Type	HTTP Accept Header	URI Suffix Extension
Abbreviated RDF/XML	application/rdf+xml+abbrev	rdfxml-abbrev
RDF/XML	application/rdf+xml text/rdf	rdfxml rdf xml
HTML	application/xhtml+xml text/html	html htm
TURTLE[a]	application/n3 application/rdf+n3 application/turtle application/x-turtle text/n3 text/turtle text/rdf+n3 text/rdf+turtle	turtle ttl n3
JSON[b]	application/json text/json	json
JSON-LD[c]	application/x-json+ld application/x-json+rdf application/json+ld application/json+rdf application/ld+json application/rdf+json	Jsonld Json-ld ldjson ld-json
N-TRIPLES	text/plain	ntriples (default)

New Format

PubChemRDF REST API: Content Negotiation

- http://rdf.ncbi.nlm.nih.gov/pubchem/compound/CID2244.rdf
- http://rdf.ncbi.nlm.nih.gov/pubchem/compound/CID2244.xml
- http://rdf.ncbi.nlm.nih.gov/pubchem/compound/CID2244.rdfxml
- http://rdf.ncbi.nlm.nih.gov/pubchem/compound/CID2244.html
- http://rdf.ncbi.nlm.nih.gov/pubchem/compound/CID2244.turtle
- http://rdf.ncbi.nlm.nih.gov/pubchem/compound/CID2244.ttl
- http://rdf.ncbi.nlm.nih.gov/pubchem/compound/CID2244.json
- http://rdf.ncbi.nlm.nih.gov/pubchem/compound/CID2244.ntriples

Practice

Try different format in your browser

Follow redirect ***Content negotiation***

curl -L -H "Accept: text/rdf"
http://rdf.ncbi.nlm.nih.gov/pubchem/compound/CID2244

PubChemRDF REST API: query string match

Endpoint: https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?

Parameters:

- Required: graph (or domain), name (or string)
- Optional: contain (or substring), return (or retrieve), format, limit, offset

Practice

Example 1: Retrieve the PubChemRDF synonyms having the value of "aspirin":

https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?graph=synonym&name=aspirin

Example 2: Substring search with the parameter "contain" (or "substring"), which can be either true or false:

https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?graph=synonym&name=aspirin&contain=true

Example 3: the related compounds or substances can be retrieved using parameter "return" (or "retrieve"), which can be either "compound" (or "cid") or "substance" (or "sid"):

https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?graph=synonym&name=aspirin&return=compound

Exmaple 4. The query functions support content negotiation with parameter "format" specified in Table 4. For instance, the following query will return JSON format:

https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?graph=synonym&name=aspirin&format=json

Endpoint: https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?

Parameters:

- Required: graph (or domain), name (or string)
- Optional: contain (or substring), return (or retrieve), format, limit, offset

Practice

Example 5: Retrieve the proteins with name containing "glycogen synthase kinase":

https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?graph=protein&name=glycogen%20synthase%20kinase&contain=true

Example 6: Retrieve the genes with symbol containing "GSK3":

https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?graph=gene&name=GSK3&contain=true

Example 7: Retrieve the references with title containing "alzheimer":

https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?graph=reference&name=alzheimer&contain=true

Return to page 27

Endpoint: https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?

Parameters:

- Required: graph (or domain)
- Optional: predicate (or pred), subject (or subj), object(or obj), format, limit, offset
- Multiple values of the given "subject" (or "subj") or "object" can be supplied and queried, which should be delimited by comma (",")

Practice

Example 1: Retrieve all of the unique predicates in substance subdomain:

https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?graph=substance

Example 2: retrieve the ChEBI class assignments for the PubChem substances:

https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?graph=substance&predicate=rdf:type

Example 3: retrieve the first 10 000 synonyms that are drug brand names (trademarks):

https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?graph=synonym&pred=rdf:type&obj=sio:CHEMINF_000561

Exmaple 4. retrieve the synonyms that are either Chemical Abstracts Service registry numbers or European Commission numbers:

https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?graph=synonym&pred=rdf:type&object=sio:CHEMINF_000446,sio:CHEMINF_000447&offset=1275000

PubChemRDF REST API: filter triple pattern

Endpoint: https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?

Parameters:

- Required: graph (or domain)
- Optional: predicate (or pred), subject (or subj), object(or obj), format, limit, offset
- Multiple values of the given “subject” (or “subj”) or “object” can be supplied and queried, which should be delimited by comma (",")

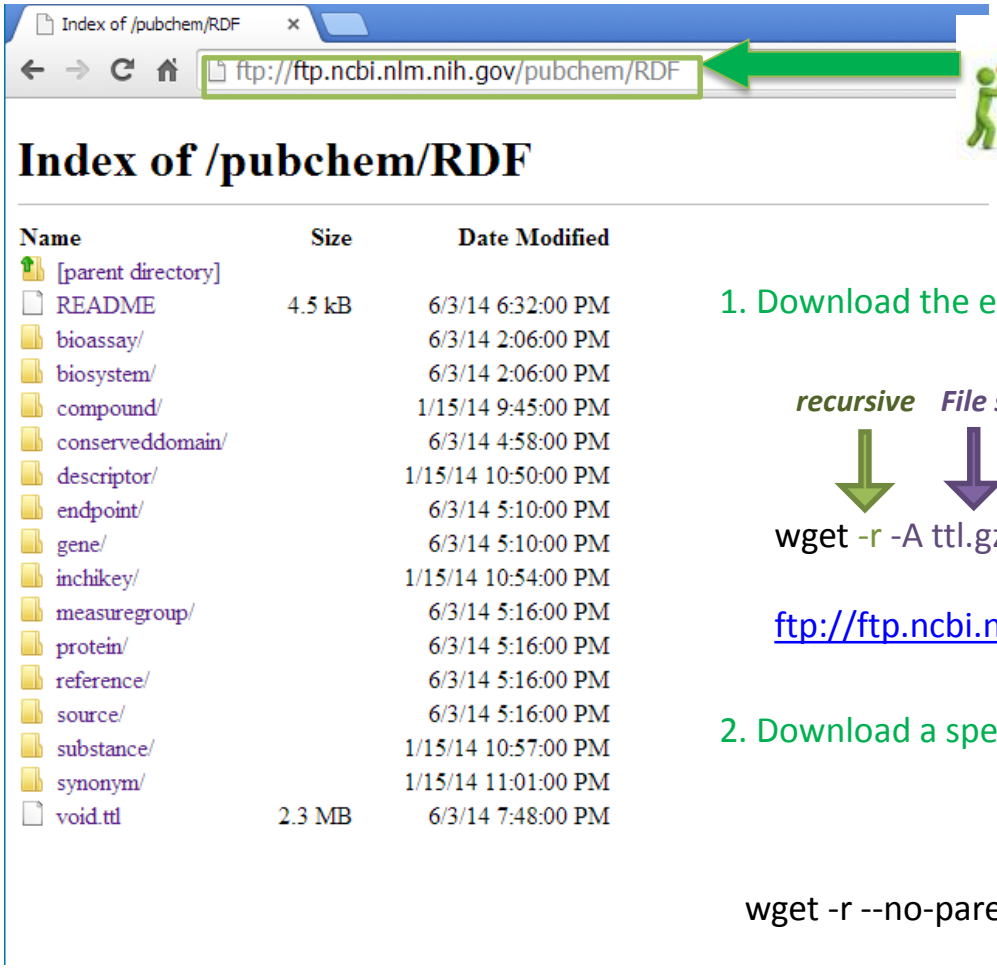
Practice

Exmaple 5. retrieve all of the PubChem depositors:
https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?graph=source&pred=rdf:type&obj=dcterms:Dataset

Exmaple 6. retrieve all of the protein complex:
https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?graph=protein&pred=rdf:type&obj=obo:GO_0043234

Exmaple 7. retrieve protein close match to UniProt P05067:
https://pubchem.ncbi.nlm.nih.gov/rest/rdf/query?graph=protein&pred=skos:closeMatch&obj=<http://purl.uniprot.org/uniprot/P05067>

PubChemRDF FTP Download



Practice

Browse FTP site

1. Download the entire directory of substance subdomain using ***wget***:

recursive *File suffix*

wget -r -A ttl.gz --no-host-directories

ftp://ftp.ncbi.nlm.nih.gov/pubchem/RDF/substance

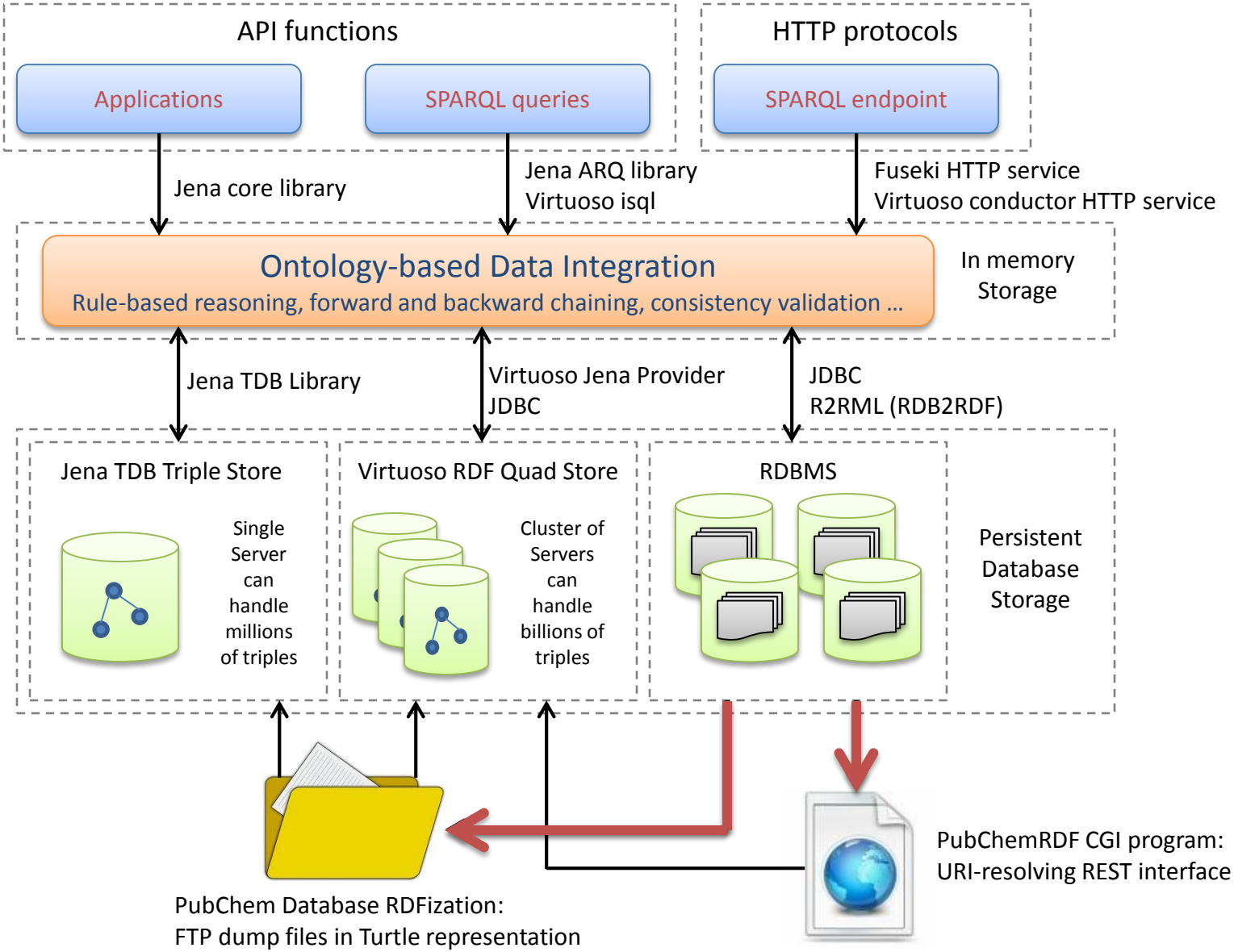
2. Download a specific type of link (substance to compound):

File suffix

wget -r --no-parent -A 'pc_substance2compound_*.ttl.gz'

ftp://ftp.ncbi.nlm.nih.gov/pubchem/RDF/substance

PubChemRDF Utility



OUTLINE

PubChem

- How the PubChemRDF is formulated?
- How to Access the Data?
- **How to answer scientific questions?**
  - SPARQL query use cases
  - http://52.18.71.59/sparql

Q: What are substance against protein GI754286265 with bioactivity less than 10 micromolar?

```
Select distinct ?substance
from <http://rdf.ncbi.nlm.nih.gov/pubchem/substance>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/measuregroup>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/endpoint>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/protein>
WHERE {
  ?substance obo:BFO_0000056 ?measuregroup .
  ?measuregroup obo:BFO_0000057 protein:GI754286265 .
  ?measuregroup obo:OBI_0000299 ?endpoint .
  ?endpoint obo:IAO_0000136 ?substance .
  ?endpoint rdf:type bao:BAO_0000190 .
  ?endpoint sio:has-value ?value .
  filter ( ?value < 10 )
}
```

Replace using Property Path:

```
?substance obo:BFO_0000056/obo:BFO_0000057 protein:GI754286265 .
?substance obo:BFO_0000056/obo:OBI_0000299 ?endpoint .
```

Appendix Table 4

Q: What are substance against protein GI754286265 with bioactivity less than 10 micromolar?

How to query by protein names containing a substring: "glycogen synthase kinase"

Step 1): query REST interface with substring search

Step 2): pick up a GI number and replace it in the SPARQL query

```
Select distinct ?substance
from <http://rdf.ncbi.nlm.nih.gov/pubchem/substance>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/measuregroup>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/endpoint>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/protein>
WHERE {
  ?substance obo:BFO_0000056 ?measuregroup .
  ?measuregroup obo:BFO_0000057 protein:GI11133187 .
  ?measuregroup obo:OBI_0000299 ?endpoint .
  ?endpoint obo:IAO_0000136 ?substance .
  ?endpoint rdf:type bao:BAO_0000190 .
  ?endpoint sio:has-value ?value .
  filter ( ?value < 10 )
}
```

Q: What protein targets are inhibited by substances with IC_{50} less than 10 µM and have the same standardized chemical structure (CID3152)?

```
Select distinct ?sub ?protein ?title
from <http://rdf.ncbi.nlm.nih.gov/pubchem/protein>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/measuregroup>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/endpoint>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/substance>
where {
  ?sub sio:CHEMINF_000477 compound:CID3152 ; obo:BFO_0000056 ?mg .
  ?mg obo:BFO_0000057 ?protein ; obo:OBI_0000299 ?ep .
  ?protein rdf:type bp:Protein ; dcterms:title ?title .
  ?ep rdf:type bao:BAO_0000190 ; obo:IAO_0000136 ?sub ; sio:has-value ?value .
  filter (?value < 10 )
}
```

Q: What protein targets does donepezil (CHEBI_53289) inhibit with an IC50 less than 10 microMolar?

```
SELECT distinct ?protein ?title
from <http://rdf.ncbi.nlm.nih.gov/pubchem/protein>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/measuregroup>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/endpoint>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/substance>
WHERE {
  ?sub rdf:type obo:CHEBI_53289 ; obo:BFO_0000056 ?mg .
  ?mg obo:BFO_0000057 ?protein ; obo:OBI_0000299 ?ep .
  ?protein rdf:type bp:Protein ; dcterms:title ?title .
  ?ep rdf:type bao:BAO_0000190 ; obo:IAO_0000136 ?sub ; sio:has-value ?value .
  filter (?value < 10 )
}
```

Q: What are the protein target for "acetylcholinesterase inhibitor" (ChEBI_37733) with IC50 < 10 µM?

```
Select distinct ?protein ?title
from <http://rdf.ncbi.nlm.nih.gov/pubchem/ruleset>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/substance>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/measuregroup>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/endpoint>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/protein>
where {
  ?chebi rdfs:subClassOf _:I .
  _:I a owl:Restriction .
  _:I owl:onProperty obov:has_role .
  _:I owl:someValuesFrom obo:CHEBI_37733 .
  ?sub rdf:type ?chebi ; obo:BFO_0000056 ?mg .
  ?mg obo:BFO_0000057 ?protein ; obo:OBI_0000299 ?ep .
  ?protein rdf:type bp:Protein ; dcterms:title ?title .
  ?ep rdf:type bao:BAO_0000190 ; obo:IAO_0000136 ?sub ; sio:has-value ?value .
  filter ( ?value < 10 )
}
```

```
?chebi rdfs:subClassOf [ a
owl:Restriction ; owl:onProperty
obov:has_role ; owl:someValuesFrom
obo:CHEBI_37733 ] .
```

Q: What substances inhibit the proteins involved in the same biological pathway: prostaglandin biosynthetic process (GO:0001516), with an IC_{50} less than 10 µM?

```
select distinct ?substance ?protein
from <http://rdf.ncbi.nlm.nih.gov/pubchem/substance>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/measuregroup>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/endpoint>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/protein>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/biosystem>
where {
  ?substance obo:BFO_0000056 ?measuregroup .
  ?measuregroup obo:BFO_0000057 ?protein .
  ?protein rdf:type bp:Protein .
  ?protein obo:BFO_0000056 obo:GO_0001516 .
  ?measuregroup obo:OBI_0000299 ?endpoint .
  ?endpoint obo:IAO_0000136 ?substance .
  ?endpoint rdf:type bao:BAO_0000190 .
  ?endpoint sio:has-value ?value .
  filter (?value < 10)
}
```

PubChemRDF Use Cases: SPARQL query

Q: What the pharmacological roles defined by CHEBI are for the substances that inhibit protein target GI754286265 with an IC_{50} less than 10 µM?

```
select distinct ?rolelabel
from <http://rdf.ncbi.nlm.nih.gov/pubchem/measuregroup>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/endpoint>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/substance>
from <http://rdf.ncbi.nlm.nih.gov/pubchem/ruleset>
from <http://purl.obolibrary.org/obo>
where {
  ?sub obo:BFO_0000056 ?mg .
  ?mg obo:BFO_0000057 protein:GI754286265 ;
  obo:OBI_0000299 ?ep .
  ?sub rdf:type ?chebi .
  ?chebi rdfs:subClassOf _:I .
  _:I a owl:Restriction .
  _:I owl:onProperty obov:has_role .
  _:I owl:someValuesFrom ?role .
  ?role rdfs:label ?rolelabel .
  ?ep obo:IAO_0000136 ?sub ; rdf:type bao:BAO_0000190 ; sio:has-value ?value .
  filter (?value < 10 )
}
```

PubChemRDF Use Cases: SPARQL query

Q: What are the protein target for “anti-bacterial agent” (mesh:D000900) with IC50 < 10 µM?

```
prefix mesh: <http://id.nlm.nih.gov/mesh/>
prefix meshv: <http://id.nlm.nih.gov/mesh/vocab#>
Select distinct ?title
where {
  service <http://id.nlm.nih.gov/mesh/sparql>
  {
    graph <http://id.nlm.nih.gov/mesh> {
      ?descr meshv:pharmacologicalAction mesh:D0000900 .
      ?descr meshv:concept ?concept .
     }
  }
  ?synonym dcterms:subject ?concept .
  ?substance sio:has-attribute ?synonym .
  ?mg obo:BFO_0000057 ?protein ; obo:OBI_0000299 ?ep .
  ?protein rdf:type bp:Protein ; dcterms:title ?title .
  ?ep rdf:type bao:BAO_0000190 ; obo:IAO_0000136 ?sub ; sio:has-value?value .
  filter ( ?value < 10 )
}
```

Federated SPARQL query over MeSH RDF

Conclusion

PubChem

- PubChem RDF is intended for ontology-based data integration
- PubChem databases have been semantically exposed to linked open data
- REST interface can be accessed to resolve URI references
- FTP dump files can be bulk-loaded into open source triples stores
- PubChemRDF can be queried using semantic web technologies: SPARQL + Inference

Acknowledgement

NCBI Structure Group:

Steve Bryant
Evan Bolton
Yanli Wang
Yu Bo
Paul Thiessen
Siqian He
Tiejun Cheng
Lianyi Han
Jeff Zhang
Jane He
Jiyao Wang
Sunghwan Kim
Other PubChem fellows

External Collaborators:

Colin Batchelor
Michel Dumontier
Janna Hastings
Hande Küçük
Stephan Schurer
Uma Vempati
Egon Willighagen
Christopher Maloney

Thank you and Questions!

Appendix Table 1: compound interrelations

CHEMINF Term ID	Label	Definition
CHEMINF_000477	has PubChem normalized counterpart	Non-symmetric predicate between substance as domain and compound as range[c]
CHEMINF_000480	has component with uncharged counterpart	Non-symmetric predicate between a mixture compound as domain and its component as range
CHEMINF_000455	is isotopologue of	Symmetric predicate between two compounds (isotopomers)
CHEMINF_000461	is stereoisomer of	Symmetric predicate between two compounds (stereoisomers)
CHEMINF_000462	has same connectivity as	Symmetric predicate between two compounds with same connectivity
CHEMINF_000482	similar to by PubChem 2-D similarity algorithm	Symmetric predicate between two similar compounds according to 2-D Tanimoto score
CHEMINF_000483	similar to by PubChem 3-D similarity algorithm	Symmetric predicate between two similar compound according to 3-D Shape and Color Tanimoto scores

Appendix Table 2: chemical descriptors

Property Name	Term ID	Software Library
Molecular Weight	CHEMINF_000334	PubChem
Molecular Formula	CHEMINF_000335	
Total Formal Charge	CHEMINF_000336	
Mono Isotopic Weight	CHEMINF_000337	
Exact Mass	CHEMINF_000338	
Compound Identifier	CHEMINF_000140	
Covalent Unit Count	CHEMINF_000369	
Defined Atom Stereocenter Count	CHEMINF_000370	
Defined Bond Stereocenter Count	CHEMINF_000371	
Isotope Atom Count	CHEMINF_000372	
Heavy Atome Count	CHEMINF_000373	
Undefined Atom Stereocenter Count	CHEMINF_000374	
Undefined Bond Stereocenter Count	CHEMINF_000375	
Canonical SMILES	CHEMINF_000376	OEChem
Isomeric SMILES	CHEMINF_000379	
Preferred IUPAC Name	CHEMINF_000382	LexiChem
Hydrogen Bond Donor Count	CHEMINF_000387	Cactvs
Hydrogen Bond Acceptor Count	CHEMINF_000388	
Rotatable Bond Count	CHEMINF_000389	
Structure Complexity	CHEMINF_000390	
Tautomer Count	CHEMINF_000391	
TPSA	CHEMINF_000392	
XLogP3	CHEMINF_000395	XLogP3
IUPAC InChI	CHEMINF_000396	InChI
IUPAC InChIKey	CHEMINF_000399	

Return to page 5

Appendix Table 3: synonyms

Database identifier	CHEMINF Term ID
ChEMBL identifier	CHEMINF_000412
KEGG identifier	CHEMINF_000409
Human Metabolome Database identifier	CHEMINF_000408
ChemSpider identifier	CHEMINF_000405
ChEBI identifier	CHEMINF_000407
DrugBank identifier	CHEMINF_000406
CAS registry number	CHEMINF_000446
EC number	CHEMINF_000447
LipidMaps identifier	CHEMINF_000564
National service center number	CHEMINF_000565
Unique ingredient identifier	CHEMINF_000563
Validated chemical database identifier	CHEMINF_000467
Drug trade name	CHEMINF_000561
International nonproprietary name	CHEMINF_000562
PubChem depositor-supplied name	CHEMINF_000339

Appendix Table 4: bioassay

BAO: *bioassay ontology*

SO: *sequence ontology*

BFO: *basic formal ontology*

OBI: *ontology for biomedical investigations*

IAO: *information artifact ontology*

Identifier	Label	OWL Type
BAO_0000015	bioassay	class
BAO_0000040	measure group	class
BAO_0000030	confirmatory assay	class
BAO_0000031	primary assay	class
BAO_0000517	summary assay	class
BAO_0002162	concentration response endpoint	class
SO_0000417	polypeptide domain	class
BFO_0000034	function	class
BAO_0000210	has assay stage	object property
BAO_0000809	has confirmatory assay	object property
BAO_0000812	has summary assay	object property
BAO_0000808	has primary assay	object property
BAO_0000209	has measure group	object property
BFO_0000057	has participant at some time	object property
BFO_0000056	participates in at some time	object property
OBI_0000299	has specified output	object property
IAO_0000136	is about	object property
BFO_0000110	has continuant part at all times	object property
BFO_0000171	located in at all times	object property
BFO_0000160	has function at all times	object property

Return to page 25

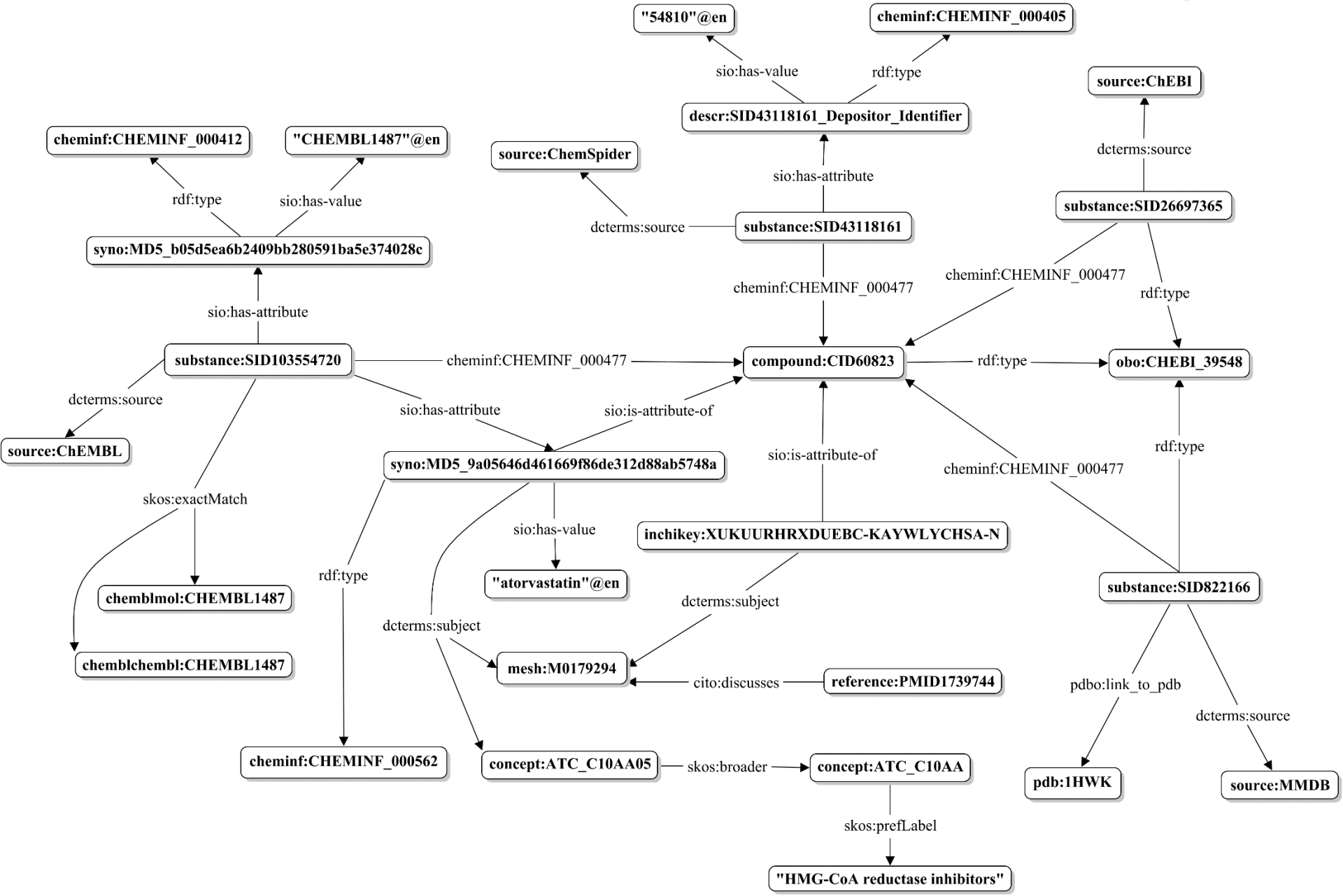
Appendix Table 5: biological endpoints

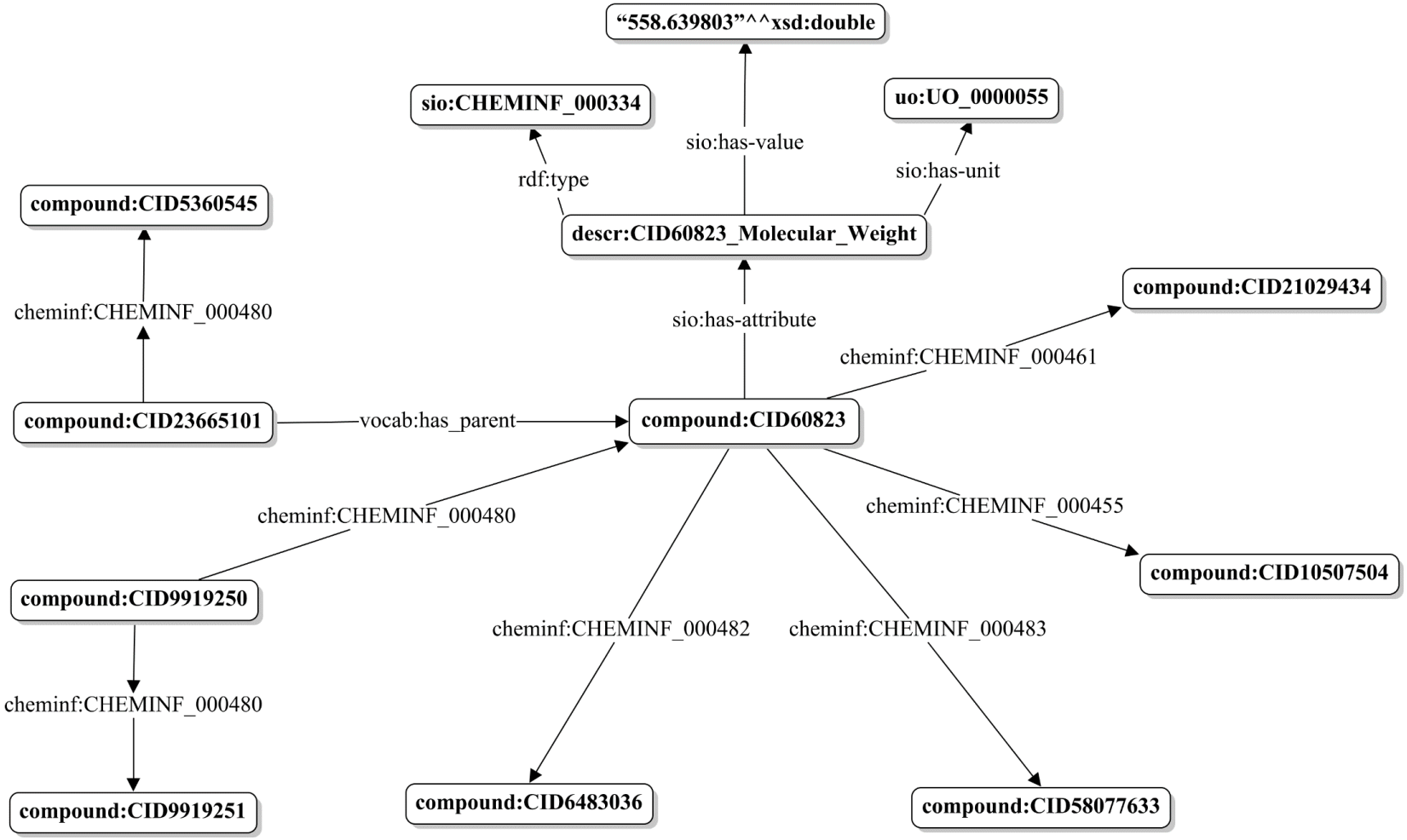
BAO Label	BAO Identifier
EC 5 hour	BAO_0002862
AC50	BAO_0000186
AC1000 absolute	BAO_0002877
AC10 absolute	BAO_0002878
AC26 absolute	BAO_0002879
AC35 absolute	BAO_0002880
AC40 absolute	BAO_0002881
AC500 absolute	BAO_0002882
IC90	BAO_0002144
Ki	BAO_0000192
CC50	BAO_0000187
ECMax_fold increase	BAO_0002886
ECMax_percent inhibition	BAO_0002887
EC50	BAO_0000188
ECMax	BAO_0002883
ED50	BAO_0003036

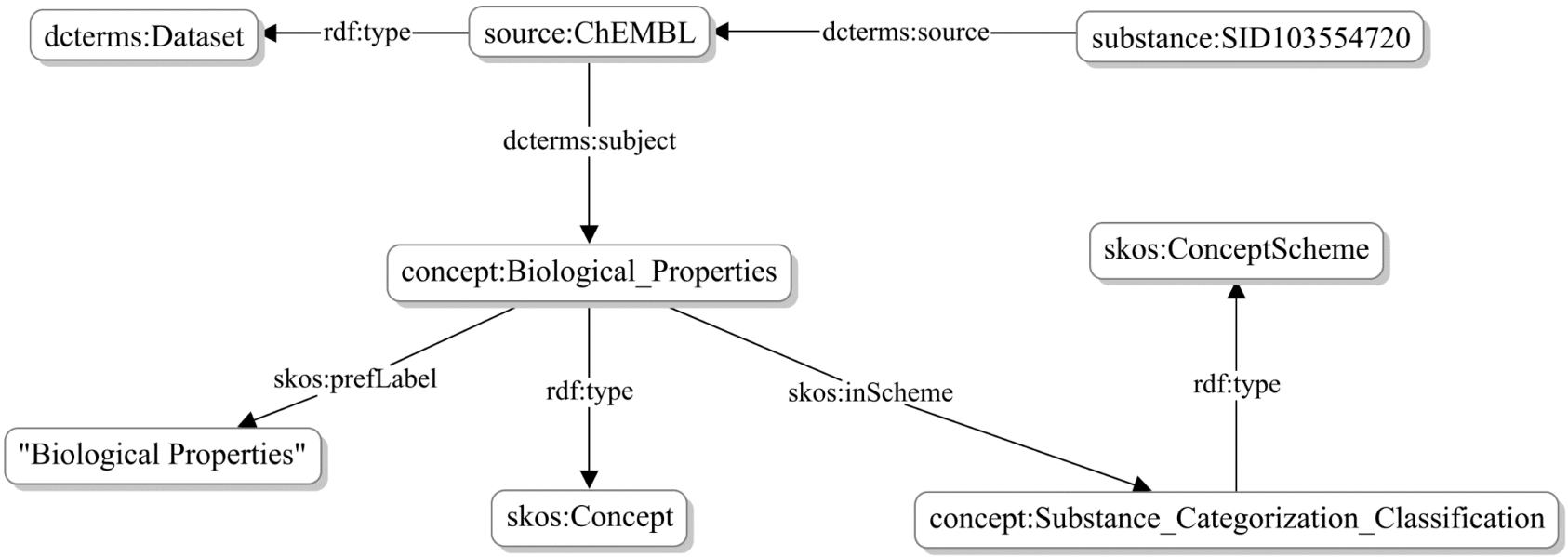
BAO Label	BAO ID
GI50	BAO_0000189
IC50	BAO_0000190
Kd	BAO_0000034
Km	BAO_0000477
LC50	BAO_0002145
LD50	BAO_0002117
MIC	BAO_0002146
ECMax_Tm	BAO_0002884
50 percent cell viability	BAO_0000349
TGI	BAO_0000194

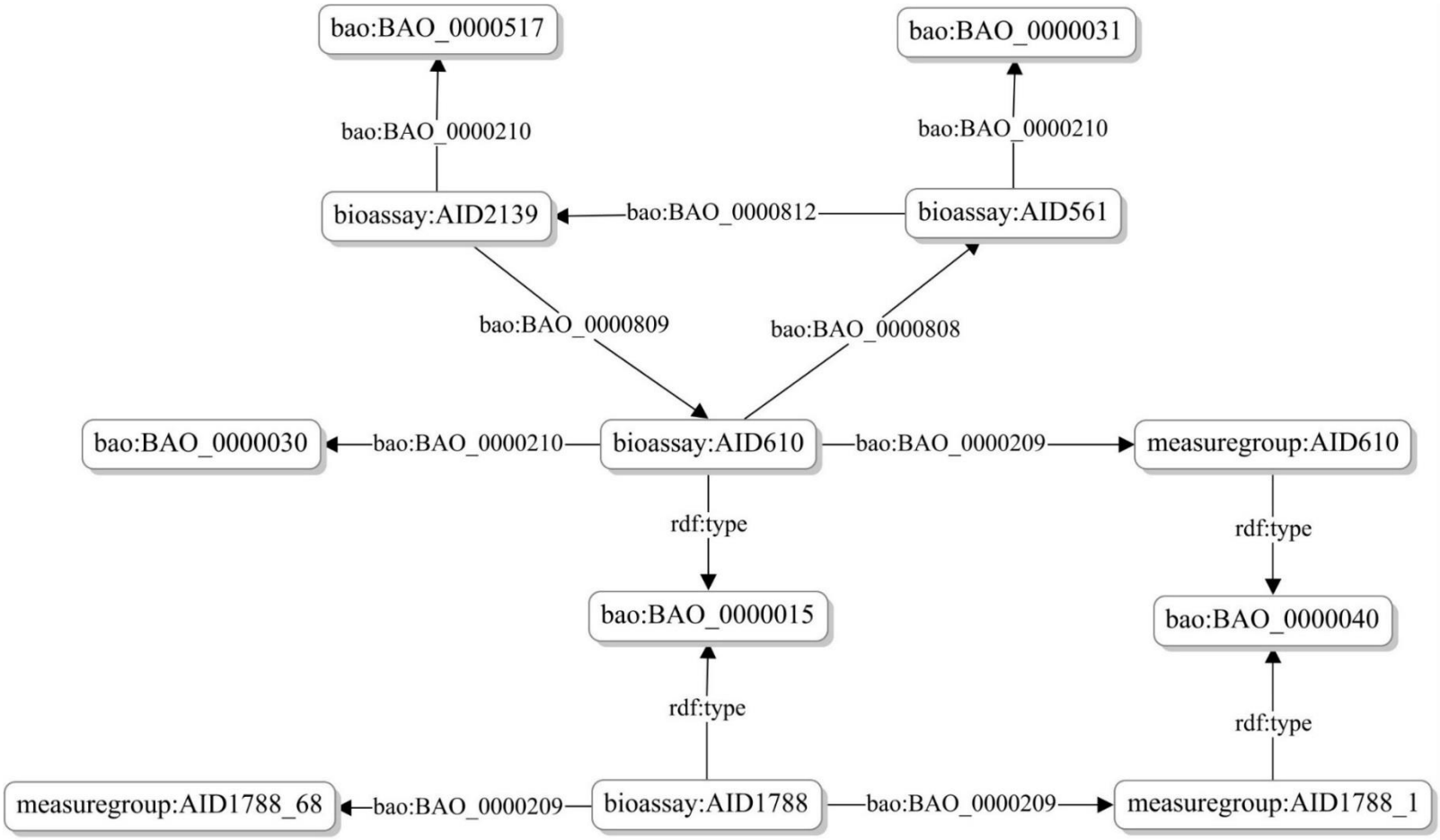
Appendix Figure 1: compound, substance, synonym, et al.

PubChem




"558.639803"^^xsd:double
sio:CHEMINF_000334
uo:UO_0000055
sio:has-value
rdf:type
sio:has-unit
compound:CID5360545
descr:CID60823_Molecular_Weight
compound:CID21029434
cheminf:CHEMINF_000480
sio:has-attribute
cheminf:CHEMINF_000461
compound:CID23665101
vocab:has_parent
compound:CID60823
cheminf:CHEMINF_000480
cheminf:CHEMINF_000455
compound:CID10507504
compound:CID9919250
cheminf:CHEMINF_000482
cheminf:CHEMINF_000483
cheminf:CHEMINF_000480
compound:CID9919251
compound:CID6483036
compound:CID58077633


dcterms:Dataset
rdf:type
source:ChEMBL
dcterms:source
substance:SID103554720
dcterms:subject
concept:Biological_Properties
skos:ConceptScheme
skos:prefLabel
rdf:type
skos:inScheme
rdf:type
"Biological Properties"
skos:Concept
concept:Substance_Categorization_Classification


bao:BAO_0000517
bao:BAO_0000031
bao:BAO_0000210
bao:BAO_0000210
bioassay:AID2139
bao:BAO_0000812
bioassay:AID561
bao:BAO_0000809
bao:BAO_0000808
bao:BAO_0000030
bao:BAO_0000210
bioassay:AID610
bao:BAO_0000209
measuregroup:AID610
rdf:type
rdf:type
bao:BAO_0000015
bao:BAO_0000040
rdf:type
rdf:type
measuregroup:AID1788_68
bao:BAO_0000209
bioassay:AID1788
bao:BAO_0000209
measuregroup:AID1788_1

Appendix Figure 5: measuregroup, endpoint, substance, protein, et al.