Semantic Annotation of PubChem Database

Gang Fu, Colin Batchelor, Michel Dumontier, Janna Hastings, Hande Küçük, Stephan Schurer, Uma Vempati, Egon Willighagen, Evan Bolton

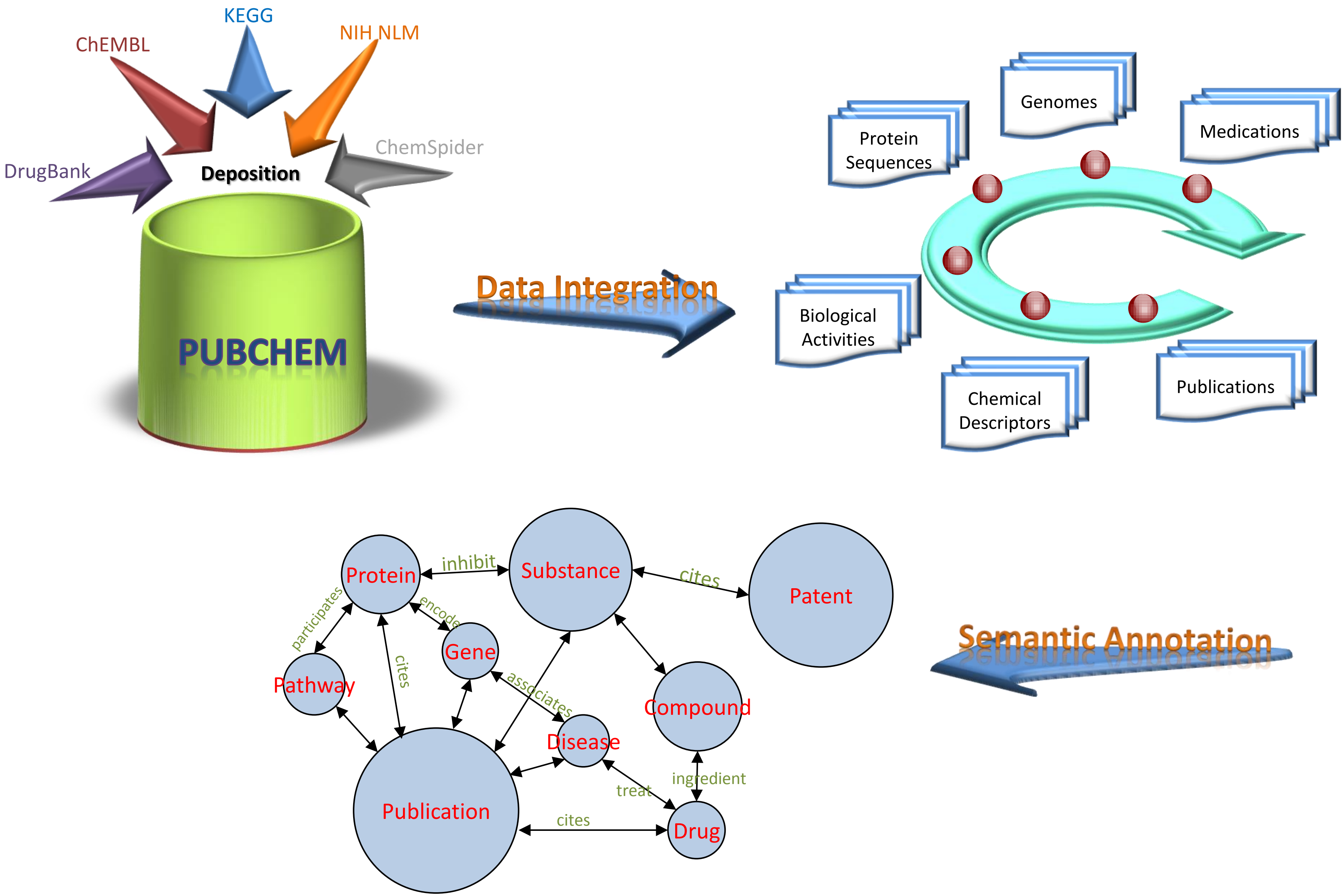
PubChem

OUTLINE

- Background Information
- Ontology-based Data Integration
- PubChem RDF URI References
- PubChem RDF Subdomains
- REST Interface
- Use Cases

Bankground Information

PubChem



Ontology-based Data Integration

	Standardized Ontologies	
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
chebi	http://purl.obolibrary.org/obo/	ChEBI and
uo		UO
sio	http://semanticscience.org/resource/	CHEMINF and
cheminf		SIO
skos	http://www.w3.org/2004/02/skos/core#	SKOS
obo	http://purl.obolibrary.org/obo/	BFO and OBI and IAO
bao	http://www.bioassayontology.org/bao#	BAO
qudt	http://data.nasa.gov/qudt/owl/qudt#	QUDT
cito	http://purl.org/spar/cito/	CiTO
fabio	http://purl.org/spar/fabio/	FaBio
ops	http://www.openphacts.org/units/	Open PHACTS vocabulary
pr	http://purl.obolibrary.org/obo/pr#	PRO
go	http://purl.obolibrary.org/obo/go#	GO
dcterms	http://purl.org/dc/terms/	DCMI Terms
pav	http://purl.org/pav/	PAV
foaf	http://xmlns.com/foaf/0.1/	FOAF vocabulary

PubChem RDF Subdomains

PubChemRDF Subdomains

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/
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/
domain	http://rdf.ncbi.nlm.nih.gov/pubchem/domain/
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	http://rdf.ncbi.nlm.nih.gov/pubchem/neighbor/
source	http://rdf.ncbi.nlm.nih.gov/pubchem/source/
vocab	http://rdf.ncbi.nlm.nih.gov/pubchem/vocabulary#

PubChem RDF URI References

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/descriptor/CID60823_Molecular_Weight

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

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/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/GI124375976

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

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

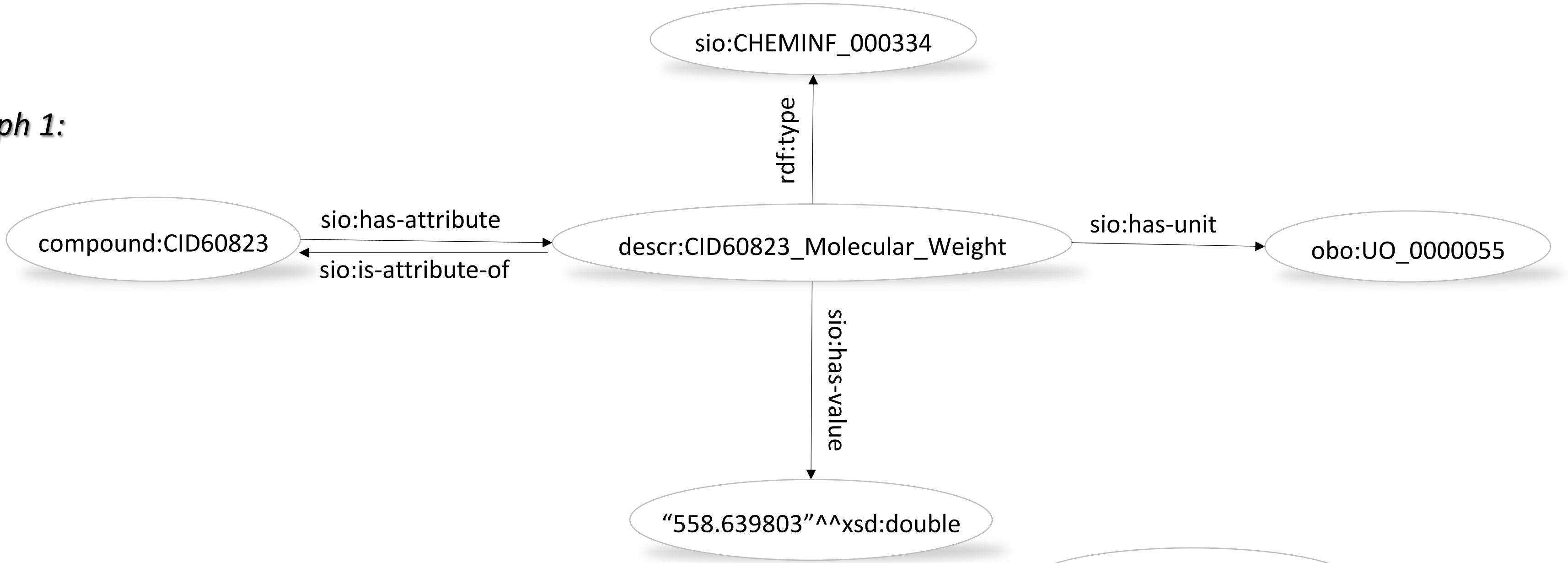
http://rdf.ncbi.nlm.nih.gov/pubchem/neighbor/CID60823_CID68019409_2DSimilarity

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

http://rdf.ncbi.nlm.nih.gov/pubchem/neighbor/GI548481_GI129900_SequenceSimilarity

PubChem RDF Graphs

Graph 1:



sio:CHEMINF_000334

rdf:type

compound:CID60823

sio:has-attribute

sio:is-attribute-of

descr:CID60823_Molecular_Weight

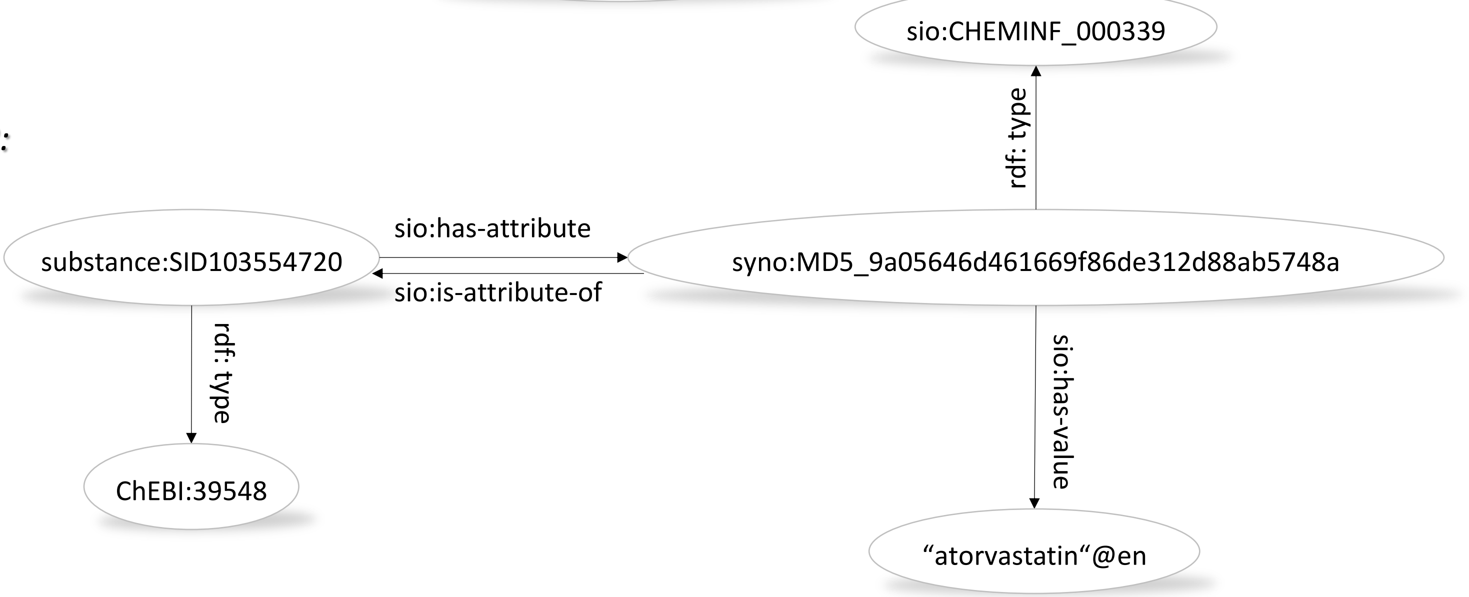
sio:has-unit

obo:UO_0000055

sio:has-value

“558.639803”^^xsd:double

Graph 2:



sio:CHEMINF_000339

rdf: type

sio:has-attribute

substance:SID103554720

sio:is-attribute-of

syno:MD5_9a05646d461669f86de312d88ab5748a

rdf: type

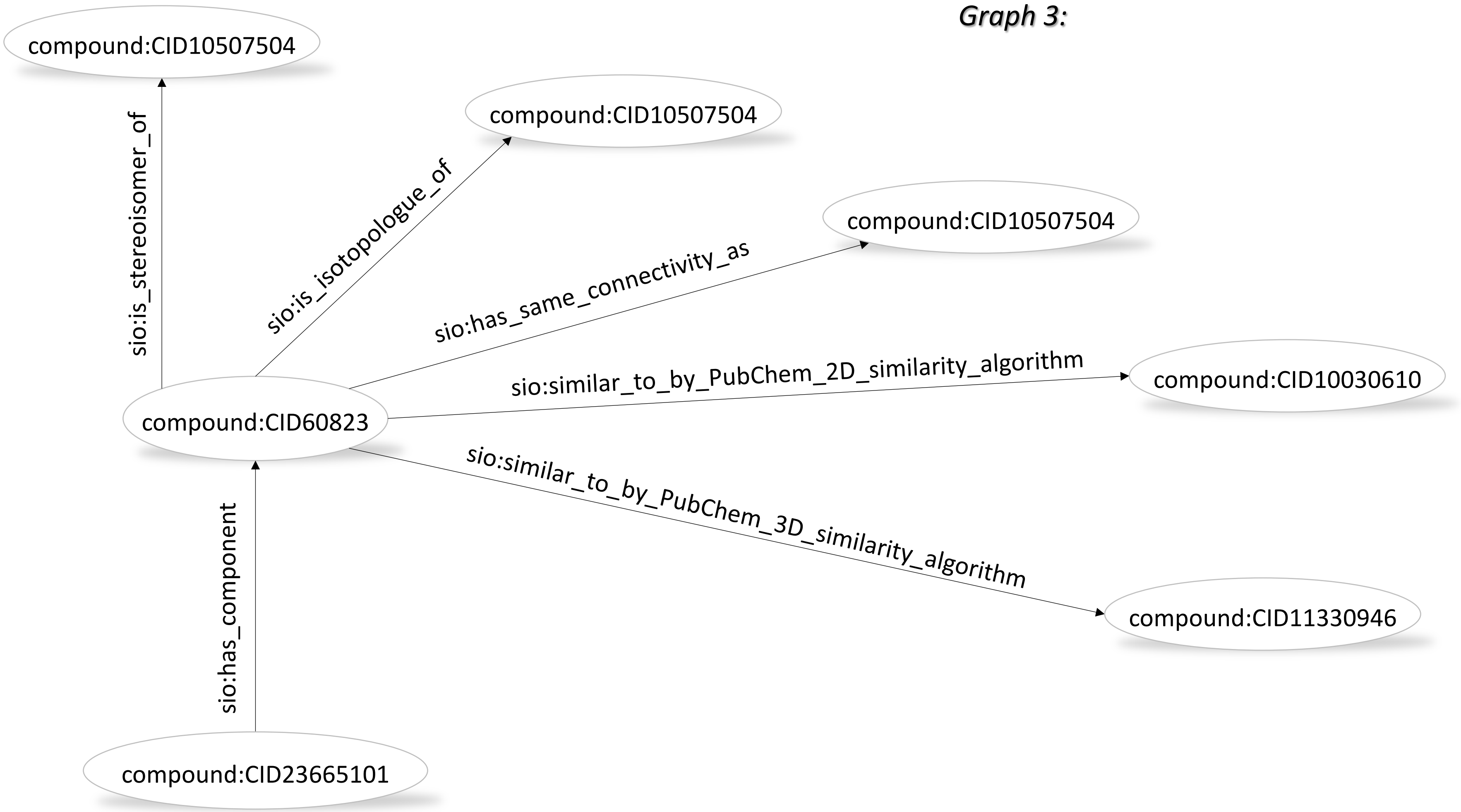
sio:has-value

ChEBI:39548

“atorvastatin“@en

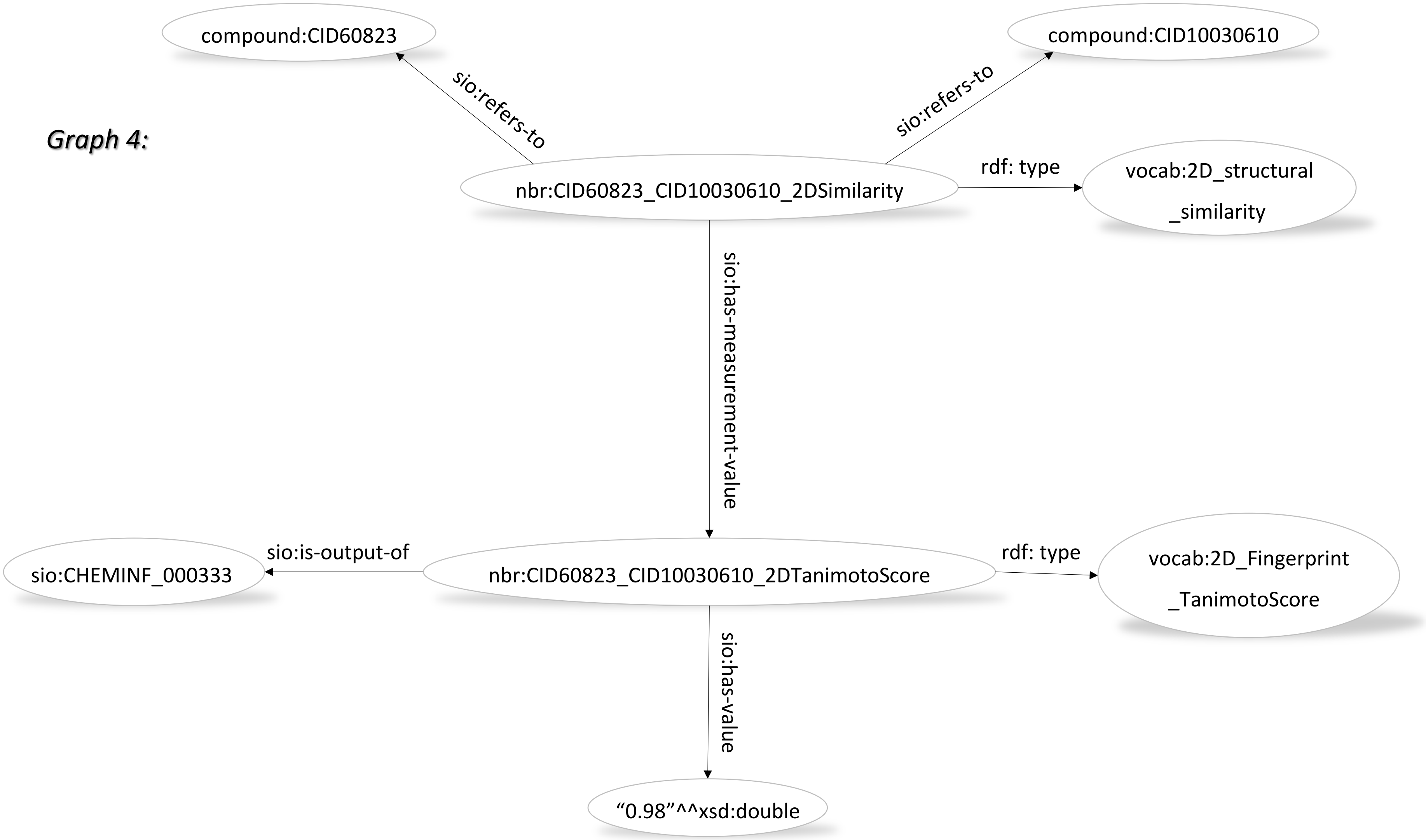
PubChem RDF Graphs

Graph 3:



PubChem RDF Graphs

Graph 4:



compound:CID60823

sio:refers-to

compound:CID10030610

sio:refers-to

nbr:CID60823_CID10030610_2DSimilarity

rdf: type

vocab:2D_structural _similarity

sio:has-measurement-value

sio:CHEMINF_000333

sio:is-output-of

nbr:CID60823_CID10030610_2DTanimotoScore

rdf: type

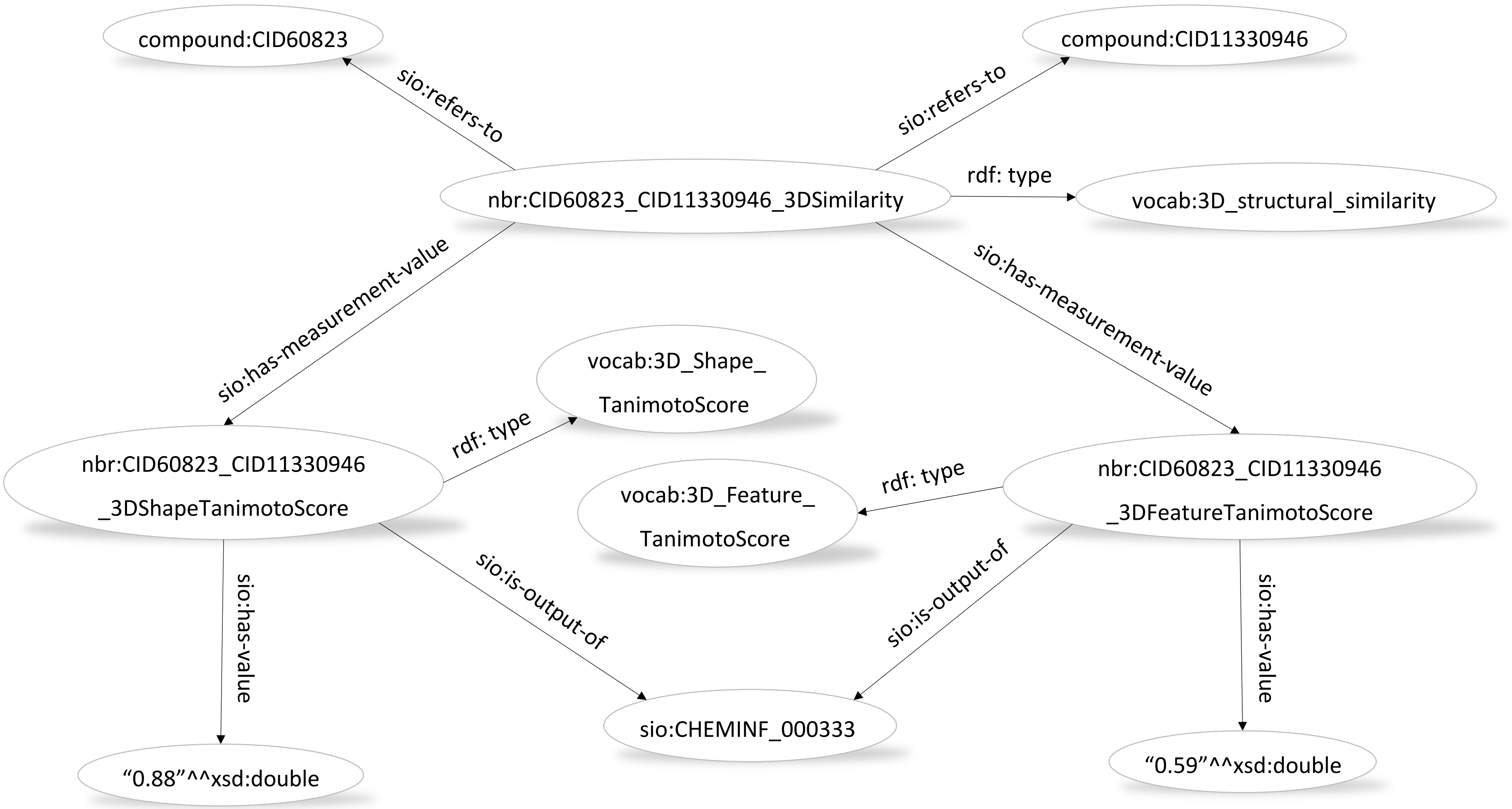
vocab:2D_Fingerprint _TanimotoScore

sio:has-value

“0.98”^^xsd:double

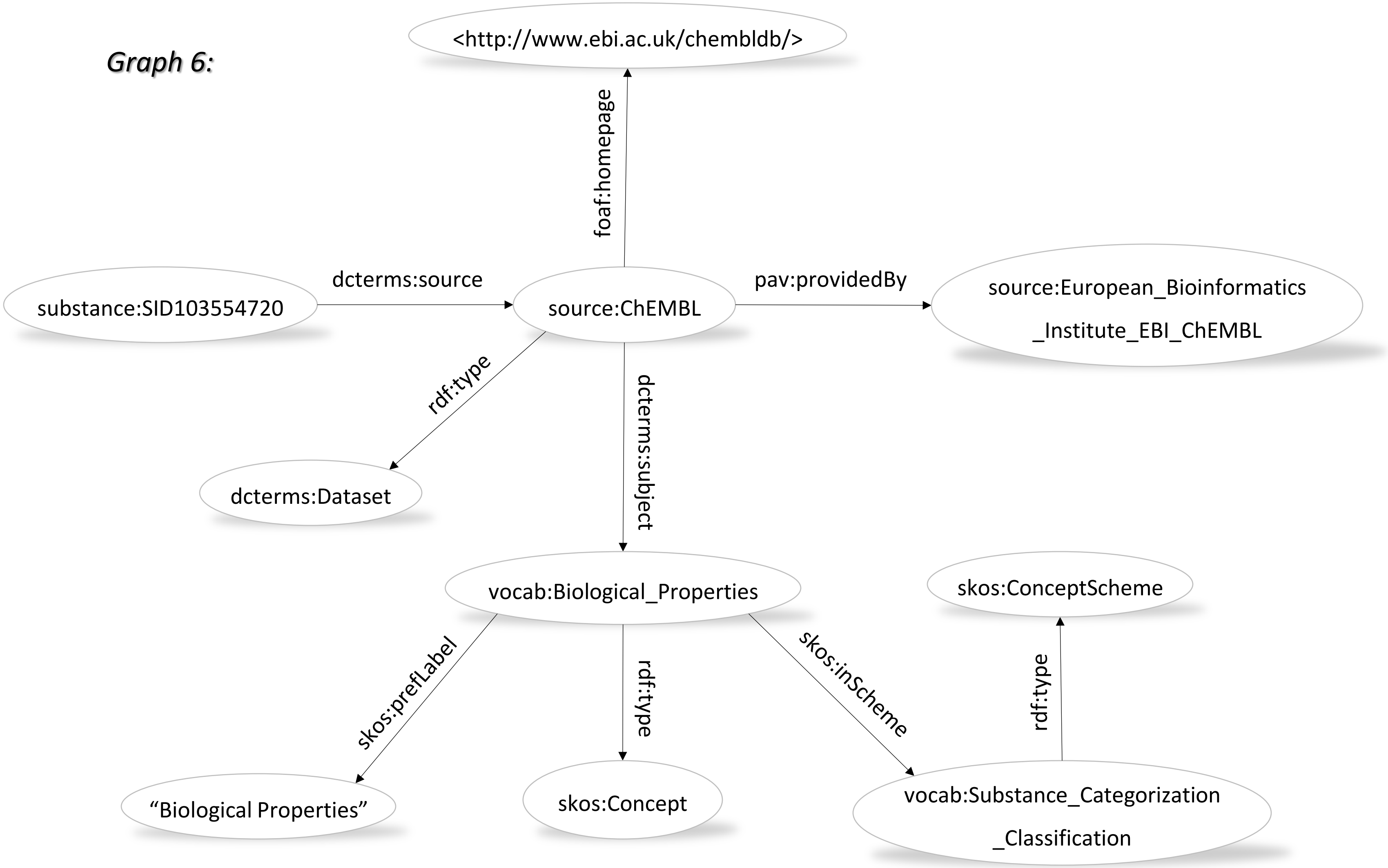
PubChem RDF Graphs

Graph 5:



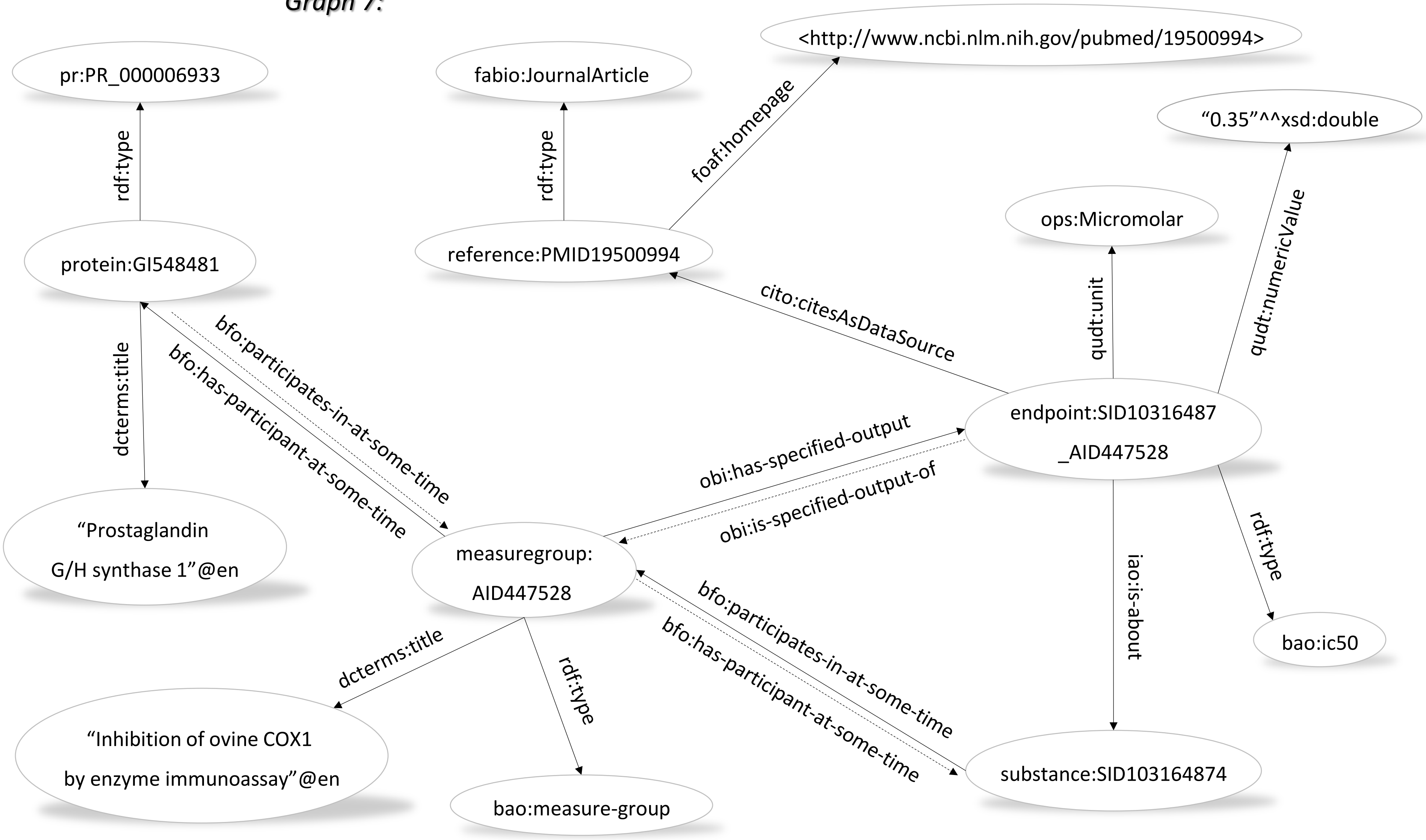
PubChem RDF Graphs

Graph 6:



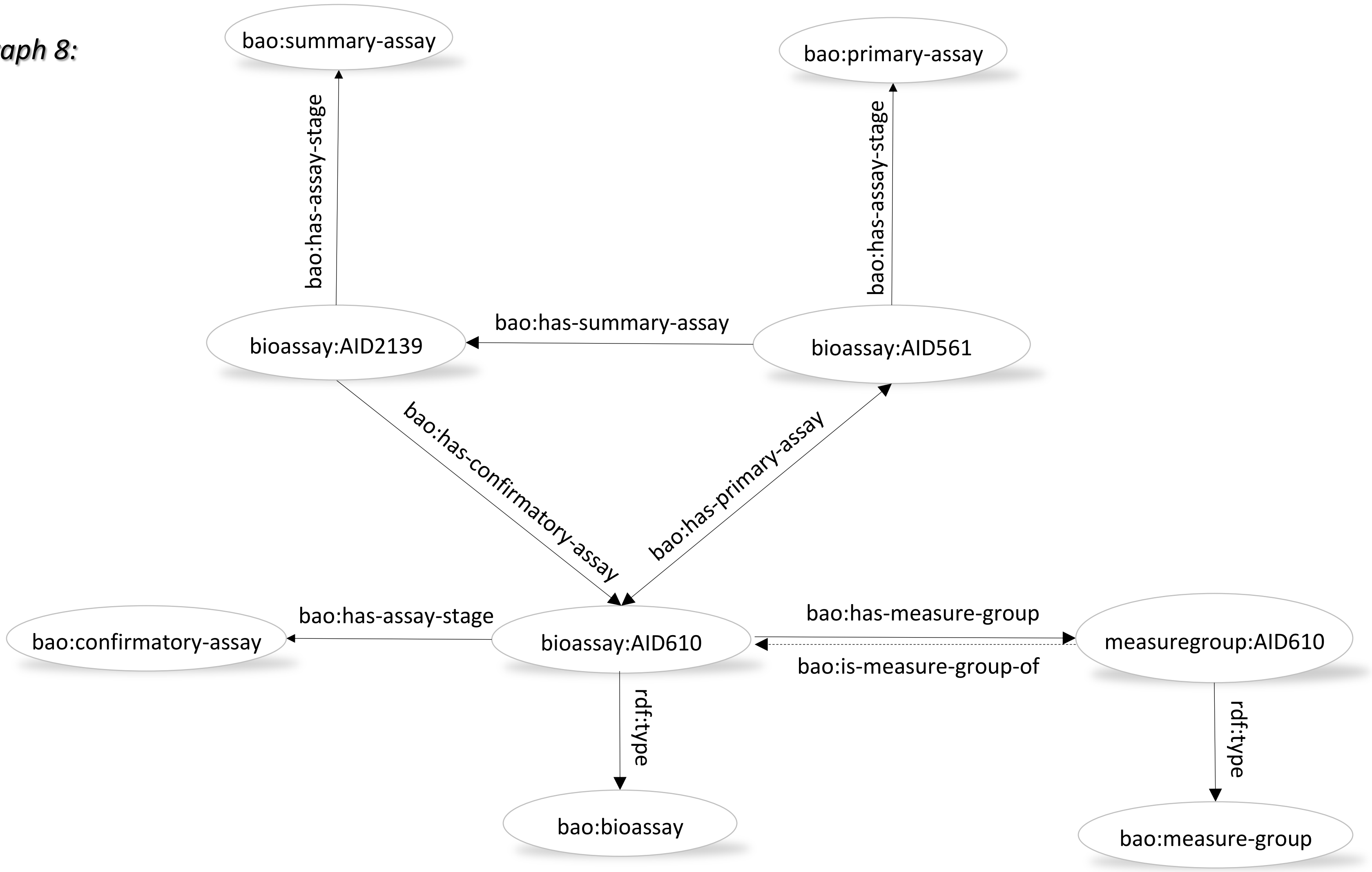
PubChem RDF Graphs

Graph 7:

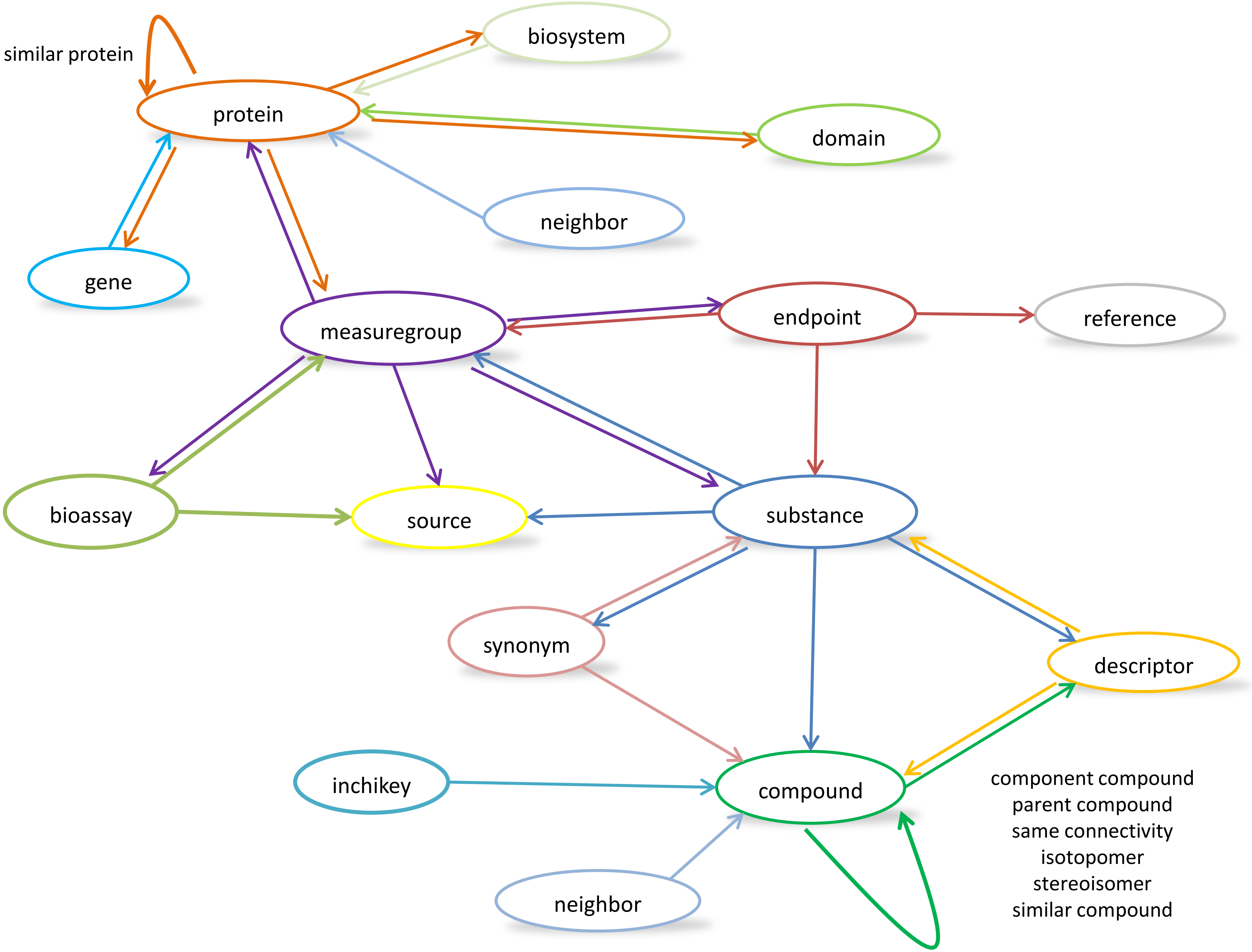


PubChem RDF Graphs

Graph 8:



PubChem RDF Linkage between Subdomains



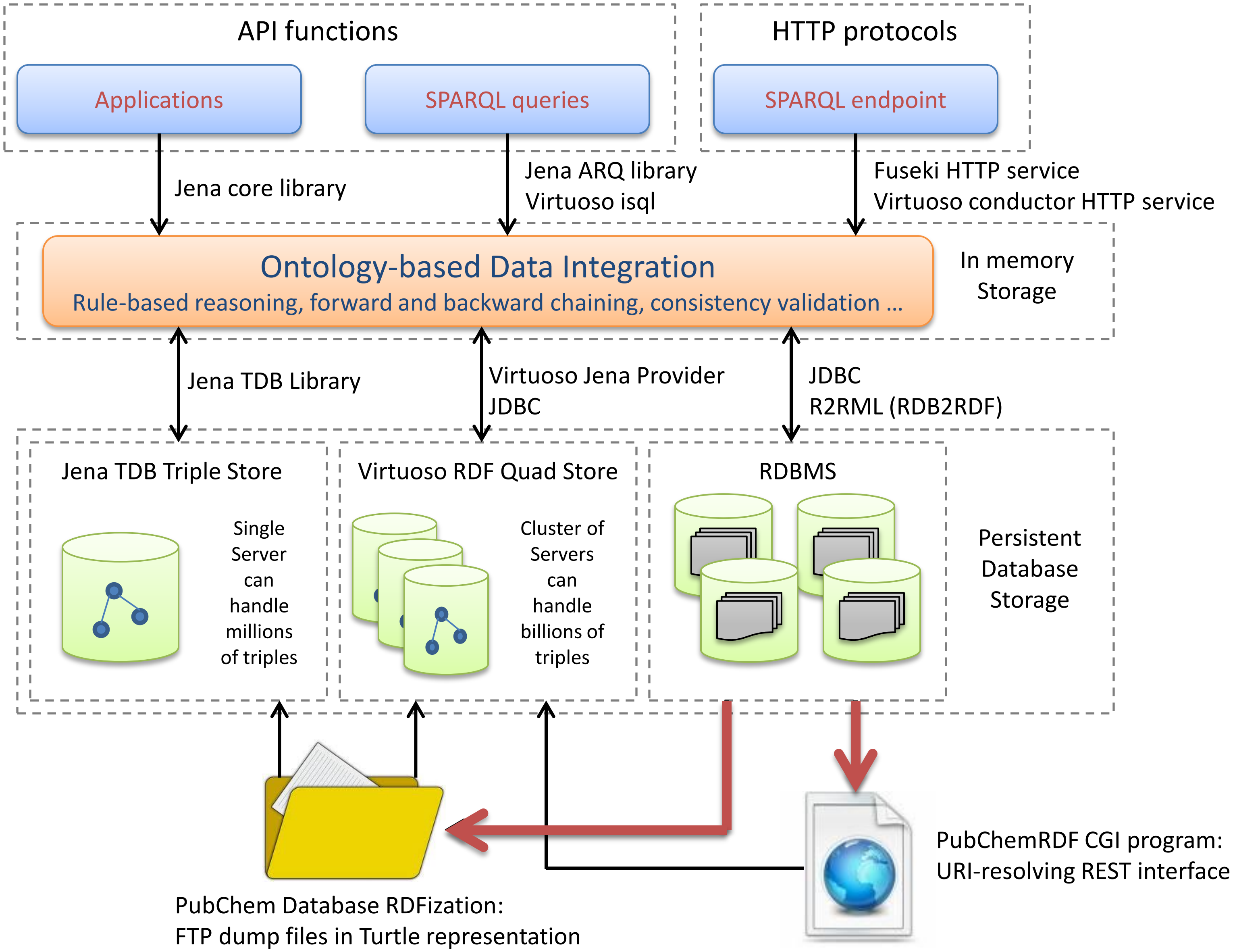
PubChem RDF REST Interface

- 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

HTTP Status	Description
400	Bad Query URL or Request URI
404	Input URI is invalid or cannot be identified in databases
405	MIME output format is unspecified or invalid
500	Some problem on the server side occurs
504	The request timed out (over 28 second)

MIME Type	HTTP Accept Header or URI extension
application/rdf+xml+abbrev	default application/rdf+xml+abbrev rdfxml-abbrev
application/rdf+xml	application/rdf+xml text/rdf rdfxml rdf xml
application/xhtml+xml	application/xhtml+xml text/html html
application/x-turtle[a]	application/rdf+n3 application/turtle application/x-turtle turtle ttl
application/json[b]	application/json text/json json
text/plain	text/plain ntriples
text/rdf+n3	text/n3 text/rdf+n3 Accept: n3

PubChem RDF Utility



API functions

Applications

SPARQL queries

HTTP protocols

SPARQL endpoint

Jena core library

Jena ARQ library
Virtuoso isql

Fuseki HTTP service
Virtuoso conductor HTTP service

Ontology-based Data Integration
Rule-based reasoning, forward and backward chaining, consistency validation ...

In memory
Storage

Jena TDB Library

Virtuoso Jena Provider
JDBC

JDBC
R2RML (RDB2RDF)

Jena TDB Triple Store

Single Server can handle millions of triples

Virtuoso RDF Quad Store

Cluster of Servers can handle billions of triples

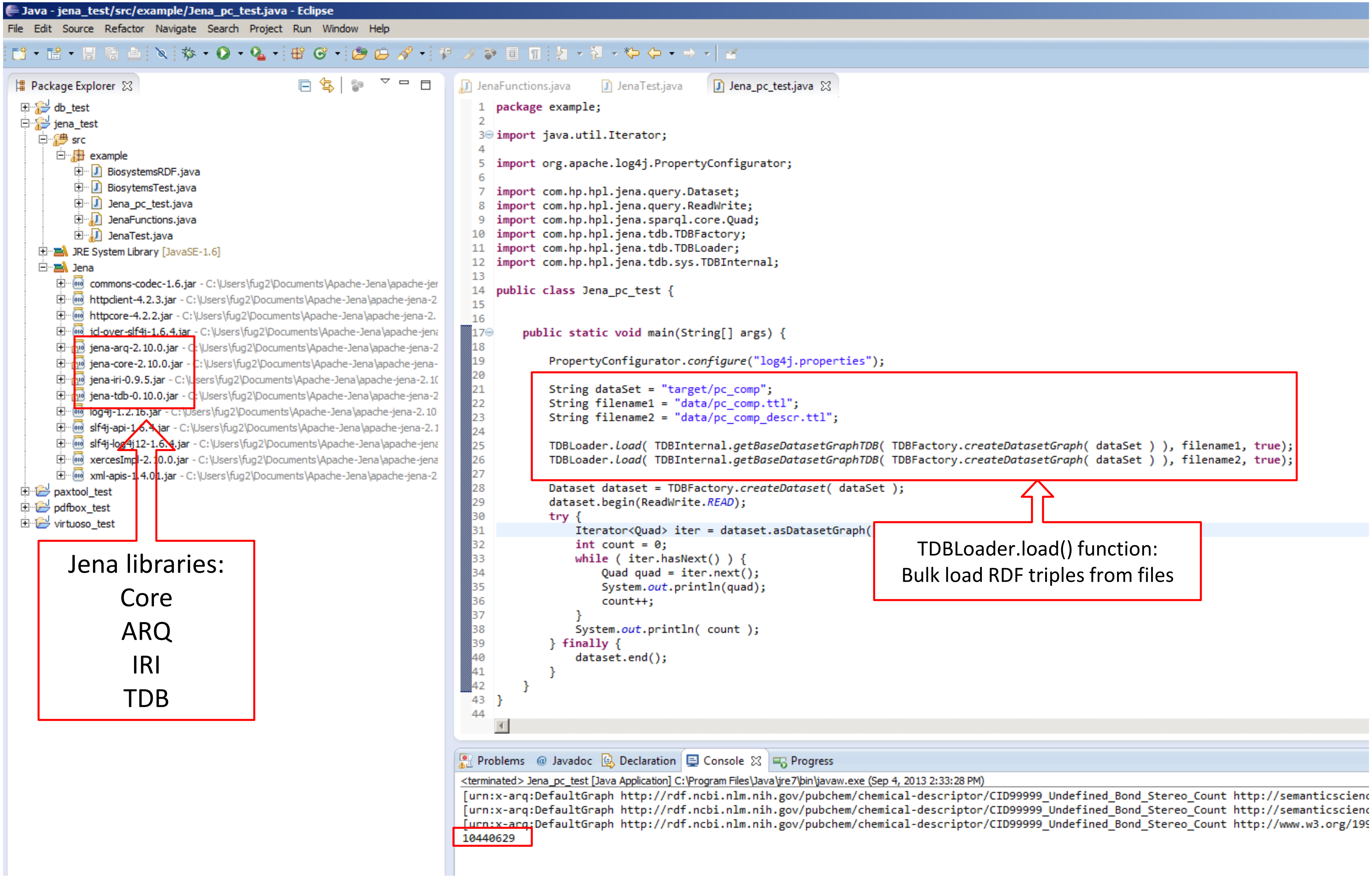
RDBMS

Persistent
Database
Storage

PubChem Database RDFization:
FTP dump files in Turtle representation

PubChemRDF CGI program:
URI-resolving REST interface

PubChem RDF Utility



```
db_test
jena_test
  src
    example
      BiosystemsRDF.java
      BiosytemsTest.java
      Jena_pc_test.java
      JenaFunctions.java
      JenaTest.java
  JRE System Library [JavaSE-1.6]
  Jena
    commons-codec-1.6.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jer
    httpclient-4.2.3.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-2
    httpcore-4.2.2.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-2.
    jcl-over-slf4j-1.6.4.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena
    jena-arq-2.10.0.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-2
    jena-core-2.10.0.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-
    jena-iri-0.9.5.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-2.10
    jena-tdb-0.10.0.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-2
    log4j-1.2.16.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-2.10
    slf4j-api-1.6.4.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-2.1
    slf4j-log4j12-1.6.4.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena
    xercesImpl-2.10.0.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena
    xml-apis-1.4.01.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-2
paxtool_test
pdfbox_test
virtuoso_test
```

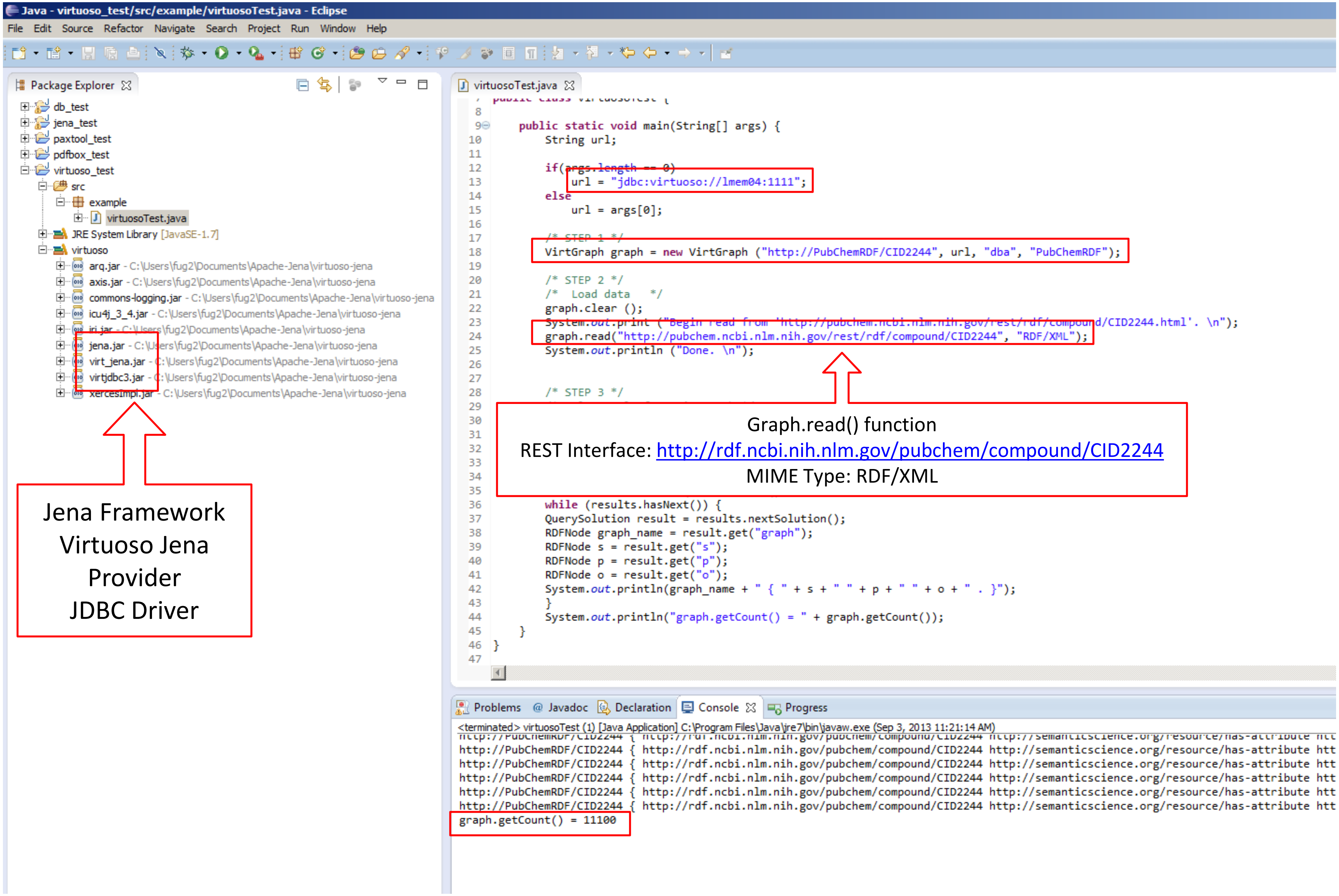
```java
package example;

import java.util.Iterator;

import org.apache.log4j.PropertyConfigurator;

import com.hp.hpl.jena.query.Dataset;
import com.hp.hpl.jena.query.ReadWrite;
import com.hp.hpl.jena.sparql.core.Quad;
import com.hp.hpl.jena.tdb.TDBFactory;
import com.hp.hpl.jena.tdb.TDBLoader;
import com.hp.hpl.jena.tdb.sys.TDBInternal;

public class Jena_pc_test {


    public static void main(String[] args) {

        PropertyConfigurator.configure("log4j.properties");

        String dataSet = "target/pc_comp";
        String filename1 = "data/pc_comp.ttl";
        String filename2 = "data/pc_comp_descr.ttl";

        TDBLoader.load( TDBInternal.getBaseDatasetGraphTDB( TDBFactory.createDatasetGraph( dataSet ) ), filename1, true);
        TDBLoader.load( TDBInternal.getBaseDatasetGraphTDB( TDBFactory.createDatasetGraph( dataSet ) ), filename2, true);

        Dataset dataset = TDBFactory.createDataset( dataSet );
        dataset.begin(ReadWrite.READ);
        try {
            Iterator<Quad> iter = dataset.asDatasetGraph(
            int count = 0;
            while ( iter.hasNext() ) {
                Quad quad = iter.next();
                System.out.println(quad);
                count++;
            }
            System.out.println( count );
        } finally {
            dataset.end();
        }
    }
}

```

```
<terminated> Jena_pc_test [Java Application] C:\Program Files\Java\jre7\bin\javaw.exe (Sep 4, 2013 2:33:28 PM)
[urn:x-arq:DefaultGraph http://rdf.ncbi.nlm.nih.gov/pubchem/chemical-descriptor/CID99999_Undefined_Bond_Stereo_Count http://semanticscienc
[urn:x-arq:DefaultGraph http://rdf.ncbi.nlm.nih.gov/pubchem/chemical-descriptor/CID99999_Undefined_Bond_Stereo_Count http://semanticscienc
[urn:x-arq:DefaultGraph http://rdf.ncbi.nlm.nih.gov/pubchem/chemical-descriptor/CID99999_Undefined_Bond_Stereo_Count http://www.w3.org/199
10440629
```

Jena libraries:
Core
ARQ
IRI
TDB

TDBLoader.load() function:
Bulk load RDF triples from files

PubChem RDF Utility



Java - virtuoso_test/src/example/virtuosoTest.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Package Explorer
db_test
jena_test
paxtool_test
pdfbox_test
virtuoso_test
src
example
virtuosoTest.java
JRE System Library [JavaSE-1.7]
virtuoso
arq.jar - C:\Users\fug2\Documents\Apache-Jena\virtuoso-jena
axis.jar - C:\Users\fug2\Documents\Apache-Jena\virtuoso-jena
commons-logging.jar - C:\Users\fug2\Documents\Apache-Jena\virtuoso-jena
icu4j_3_4.jar - C:\Users\fug2\Documents\Apache-Jena\virtuoso-jena
iri.jar - C:\Users\fug2\Documents\Apache-Jena\virtuoso-jena
jena.jar - C:\Users\fug2\Documents\Apache-Jena\virtuoso-jena
virt_jena.jar - C:\Users\fug2\Documents\Apache-Jena\virtuoso-jena
virtjdbc3.jar - C:\Users\fug2\Documents\Apache-Jena\virtuoso-jena
xercesImpl.jar - C:\Users\fug2\Documents\Apache-Jena\virtuoso-jena
virtuosoTest.java
Problems
Javadoc
Declaration
Console
Progress

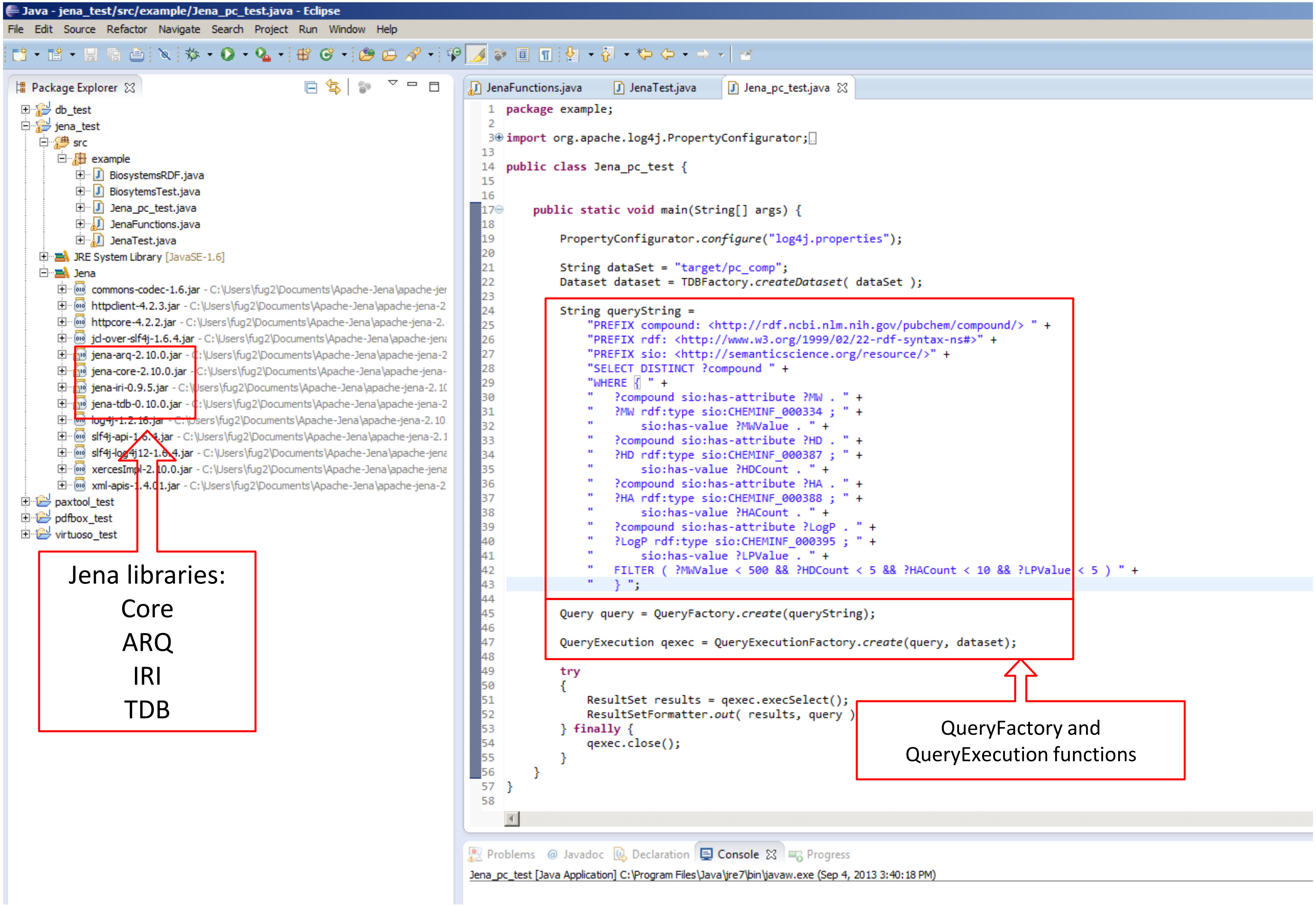
```java
public class virtuosoTest {

    public static void main(String[] args) {
        String url;

        if(args.length == 0)
            url = "jdbc:virtuoso://lmem04:1111";
        else
            url = args[0];

        /* STEP 1 */
        VirtGraph graph = new VirtGraph ("http://PubChemRDF/CID2244", url, "dba", "PubChemRDF");

        /* STEP 2 */
        /*  Load data    */
        graph.clear ();
        System.out.print ("Begin read from 'http://pubchem.ncbi.nlm.nih.gov/rest/rdf/compound/CID2244.html'. \n");
        graph.read("http://pubchem.ncbi.nlm.nih.gov/rest/rdf/compound/CID2244", "RDF/XML");
        System.out.println ("Done. \n");


        /* STEP 3 */

        while (results.hasNext()) {
        QuerySolution result = results.nextSolution();
        RDFNode graph_name = result.get("graph");
        RDFNode s = result.get("s");
        RDFNode p = result.get("p");
        RDFNode o = result.get("o");
        System.out.println(graph_name + " { " + s + " " + p + " " + o + " . }");
        }
        System.out.println("graph.getCount() = " + graph.getCount());
    }
}
```

Graph.read() function
REST Interface: http://rdf.ncbi.nih.nlm.gov/pubchem/compound/CID2244
MIME Type: RDF/XML

Jena Framework
Virtuoso Jena Provider
JDBC Driver

```
<terminated> virtuosoTest (1) [Java Application] C:\Program Files\Java\jre7\bin\javaw.exe (Sep 3, 2013 11:21:14 AM)
http://PubChemRDF/CID2244 { http://rdf.ncbi.nlm.nih.gov/pubchem/compound/CID2244 http://semanticscience.org/resource/has-attribute htt
http://PubChemRDF/CID2244 { http://rdf.ncbi.nlm.nih.gov/pubchem/compound/CID2244 http://semanticscience.org/resource/has-attribute htt
http://PubChemRDF/CID2244 { http://rdf.ncbi.nlm.nih.gov/pubchem/compound/CID2244 http://semanticscience.org/resource/has-attribute htt
http://PubChemRDF/CID2244 { http://rdf.ncbi.nlm.nih.gov/pubchem/compound/CID2244 http://semanticscience.org/resource/has-attribute htt
http://PubChemRDF/CID2244 { http://rdf.ncbi.nlm.nih.gov/pubchem/compound/CID2244 http://semanticscience.org/resource/has-attribute htt
http://PubChemRDF/CID2244 { http://rdf.ncbi.nlm.nih.gov/pubchem/compound/CID2244 http://semanticscience.org/resource/has-attribute htt
graph.getCount() = 11100
```

PubChem RDF Utility


Java - jena_test/src/example/Jena_pc_test.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Package Explorer
JenaFunctions.java
JenaTest.java
Jena_pc_test.java
Problems
Javadoc
Declaration
Console
Progress

```
db_test
jena_test
  src
    example
      BiosystemsRDF.java
      BiosytemsTest.java
      Jena_pc_test.java
      JenaFunctions.java
      JenaTest.java
  JRE System Library [JavaSE-1.6]
  Jena
    commons-codec-1.6.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jer
    httpclient-4.2.3.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-2
    httpcore-4.2.2.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-2.
    jcl-over-slf4j-1.6.4.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena
    jena-arq-2.10.0.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-2
    jena-core-2.10.0.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-
    jena-iri-0.9.5.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-2.10
    jena-tdb-0.10.0.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-2
    log4j-1.2.16.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-2.10
    slf4j-api-1.6.4.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-2.1
    slf4j-log4j12-1.6.4.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena
    xercesImpl-2.10.0.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena
    xml-apis-1.4.01.jar - C:\Users\fug2\Documents\Apache-Jena\apache-jena-2
paxtool_test
pdfbox_test
virtuoso_test
```

```java
package example;

import org.apache.log4j.PropertyConfigurator;

public class Jena_pc_test {


    public static void main(String[] args) {

        PropertyConfigurator.configure("log4j.properties");

        String dataSet = "target/pc_comp";
        Dataset dataset = TDBFactory.createDataset( dataSet );

        String queryString =
            "PREFIX compound: <http://rdf.ncbi.nlm.nih.gov/pubchem/compound/> " +
            "PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>" +
            "PREFIX sio: <http://semanticscience.org/resource/>" +
            "SELECT DISTINCT ?compound " +
            "WHERE { " +
            "   ?compound sio:has-attribute ?MW . " +
            "   ?MW rdf:type sio:CHEMINF_000334 ; " +
            "       sio:has-value ?MWValue . " +
            "   ?compound sio:has-attribute ?HD . " +
            "   ?HD rdf:type sio:CHEMINF_000387 ; " +
            "       sio:has-value ?HDCount . " +
            "   ?compound sio:has-attribute ?HA . " +
            "   ?HA rdf:type sio:CHEMINF_000388 ; " +
            "       sio:has-value ?HACount . " +
            "   ?compound sio:has-attribute ?LogP . " +
            "   ?LogP rdf:type sio:CHEMINF_000395 ; " +
            "       sio:has-value ?LPValue . " +
            "   FILTER ( ?MWValue < 500 && ?HDCount < 5 && ?HACount < 10 && ?LPValue < 5 ) " +
            "   } ";

        Query query = QueryFactory.create(queryString);

        QueryExecution qexec = QueryExecutionFactory.create(query, dataset);

        try
        {
            ResultSet results = qexec.execSelect();
            ResultSetFormatter.out( results, query )
        } finally {
            qexec.close();
        }
    }
}
```

```
Jena_pc_test [Java Application] C:\Program Files\Java\jre7\bin\javaw.exe (Sep 4, 2013 3:40:18 PM)
```

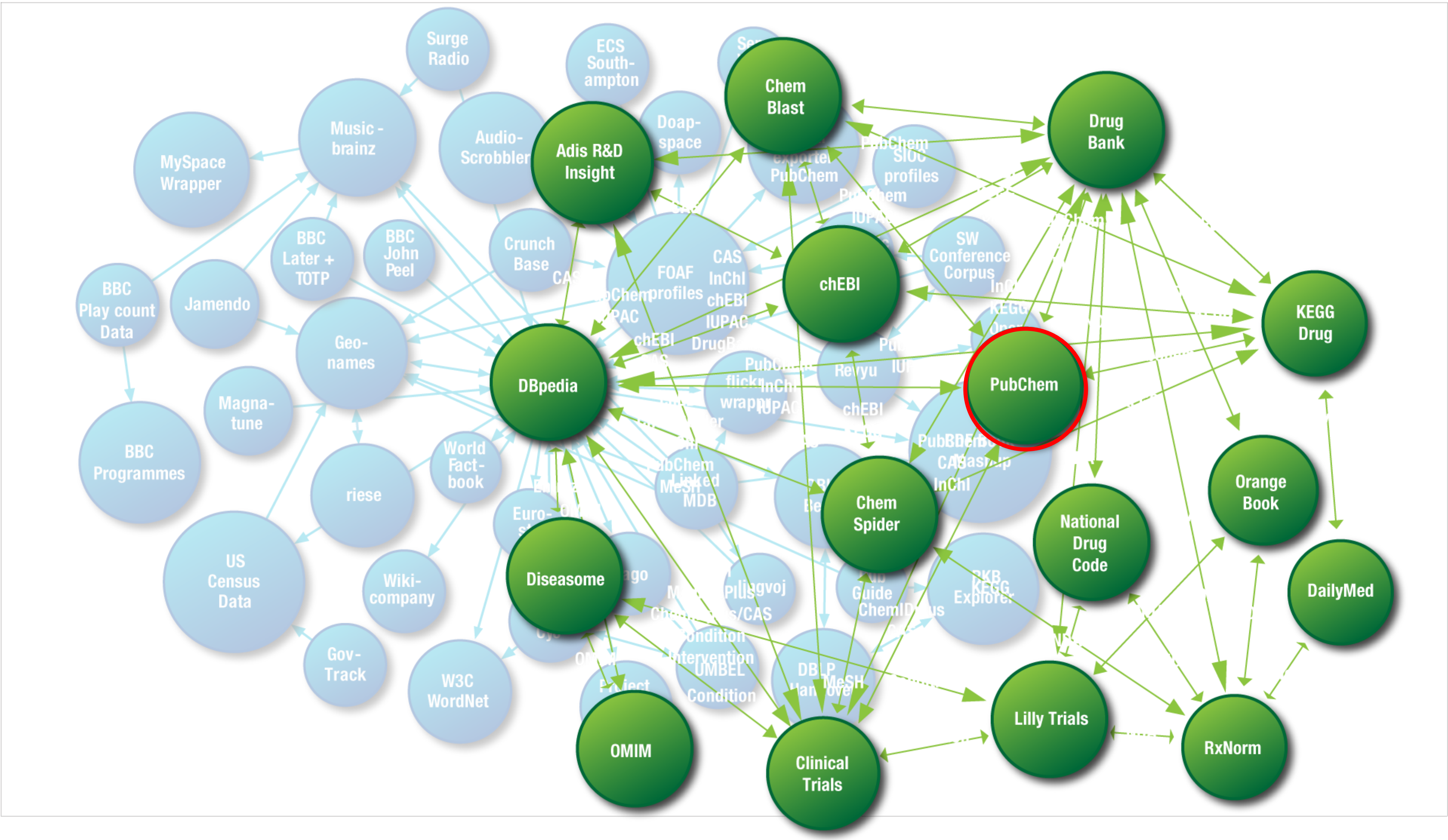
Jena libraries:
Core
ARQ
IRI
TDB

QueryFactory and
QueryExecution functions

Conclusion

- 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

Conclusion



Surge Radio
ECS South-ampton
Chem Blast
Drug Bank
MySpace Wrapper
Music-brainz
Audio-Scrobbler
Adis R&D Insight
Doap-space
PubChem SIOC profiles
BBC Play count Data
Jamendo
BBC Later + TOTP
BBC John Peel
Crunch Base
FOAF profiles
CAS
InChI
chEBI
IUPAC
SW Conference Corpus
KEGG Drug
Geo-names
DBpedia
Revyu
flickr wrappr
PubChem
Magna-tune
BBC Programmes
riese
World Fact-book
Linked MDB
Chem Spider
Orange Book
National Drug Code
US Census Data
Euro-stat
Diseasome
KEGG Explorer
DailyMed
Wiki-company
Gov-Track
W3C WordNet
UMBEL
Condition
DBLP
MeSH
OMIM
Clinical Trials
Lilly Trials
RxNorm

Acknowledgement

NCBI Structure Group:

Steve Bryant
Evan Bolton
Yanli Wang
Paul Thiessen
Siqian He
Jeff Zhang
Tugba Suzek
Lianyi Han
Jane He
Jiyao Wang
Tiejun Cheng

External Collaborators:

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

Thank you and Questions!