Handling of Small Molecules in the Semantic Web

Issaku YAMADA

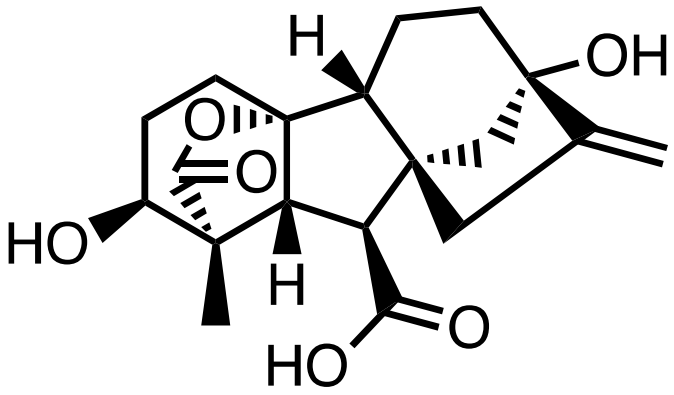
The Noguchi Institute, JAPAN

InChI Workshop @ NIH

Natcher Conference Center, Bethesda MD, USA

August 16, 2017

InChI/InChIkey of Small molecule



InChI=1S/C19H24O6/c1-9-7-17-8-18(9,24)5-3-10(17)19-6-4-11(20)16(2,15(23)25-19)13(19)12(17)14(21)22/h10-13,20,24H,1,3-8H2,2H3,(H,21,22)/t10-,11+,12-,13-,16-,17+,18+,19-/m1/s1

JLJLRLWOEMWYQK-OBDJNFEBSA-N

The **Semantic Web** is an extension of the Web through standards by the World Wide Web Consortium (W3C). The standards promote common data formats and exchange protocols on the Web, most fundamentally the Resource Description Framework (RDF).

from wikipedia

Semantic Web

Resource Description Framework (RDF)

- The RDF data model is based upon the idea of making statements about resources in the form of **subject–predicate–object** expressions, known as **triples**.
- The **subject** denotes the resource, and the **predicate** denotes traits or aspects of the resource, and expresses a relationship between the **subject** and the **object**.

from wikipedia

Layers in InChI Format

- Main Layer (immediately follows the InChI version)
  - /{formula}
  - /c{connections}
  - /h{H_atoms}
- Charge layer
  - /q{charge}
  - /p{protons}
- Stereo layer
  - /b{stereo:dbond}
  - /t{stereo:sp3}
  - /m{stereo:sp3:inverted}
  - /s{stereo:type (1=abs, 2=rel, 3=rac)}
- Isotopic Layer
  - /i{isotopic:atoms}*
  - /h{isotopic:exchangeable_H}
  - /b{isotopic:stereo:dbond}
  - /t{isotopic:stereo:sp3}
  - /m{isotopic:stereo:sp3:inverted}
  - /s{isotopic:stereo:type (1=abs, 2=rel, 3=rac)}

from InChI documentation

InChI Ontology

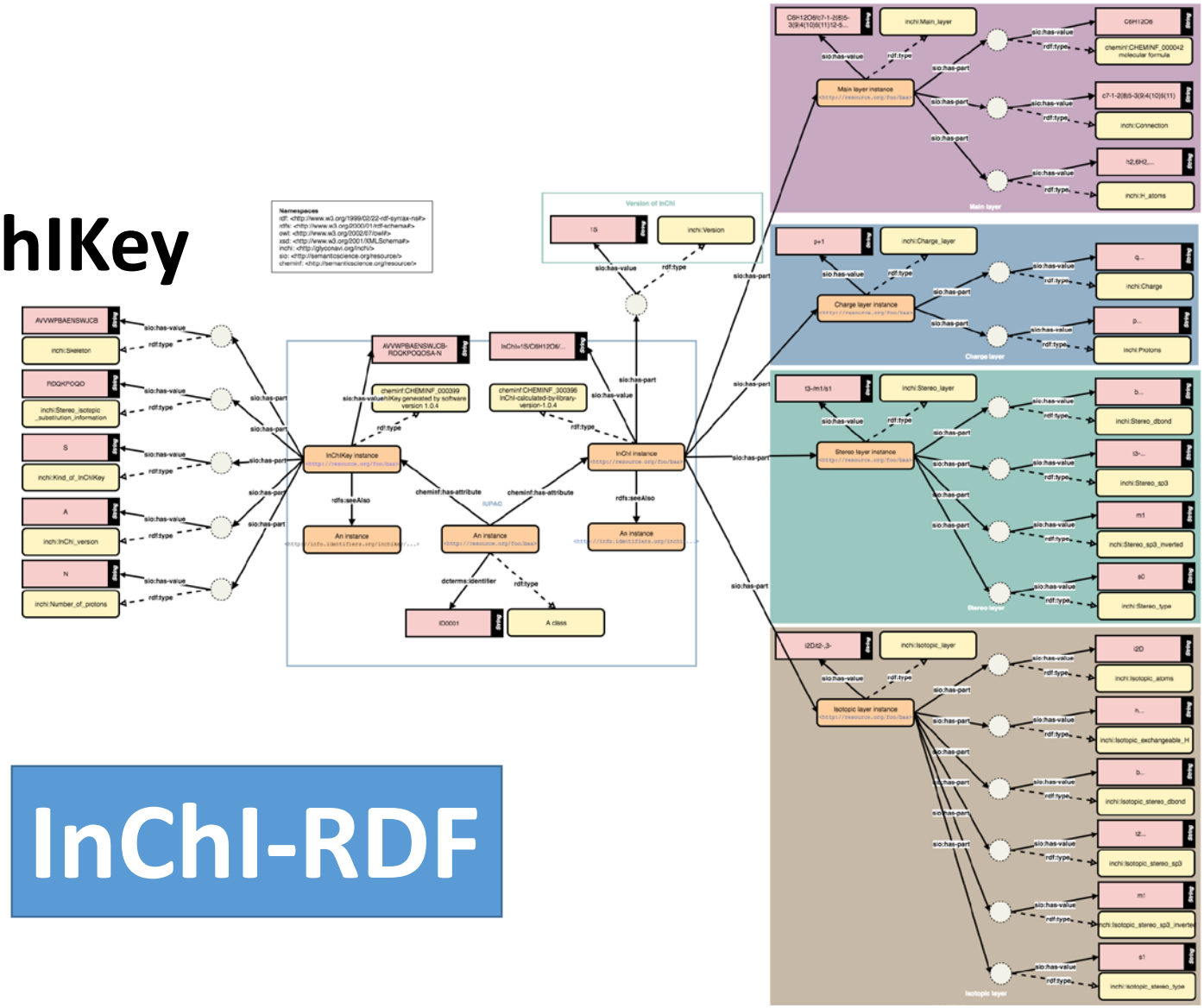
InChI RDF Schema

RDFization

RDF triples

RDF Schema of InChI and InChIKey

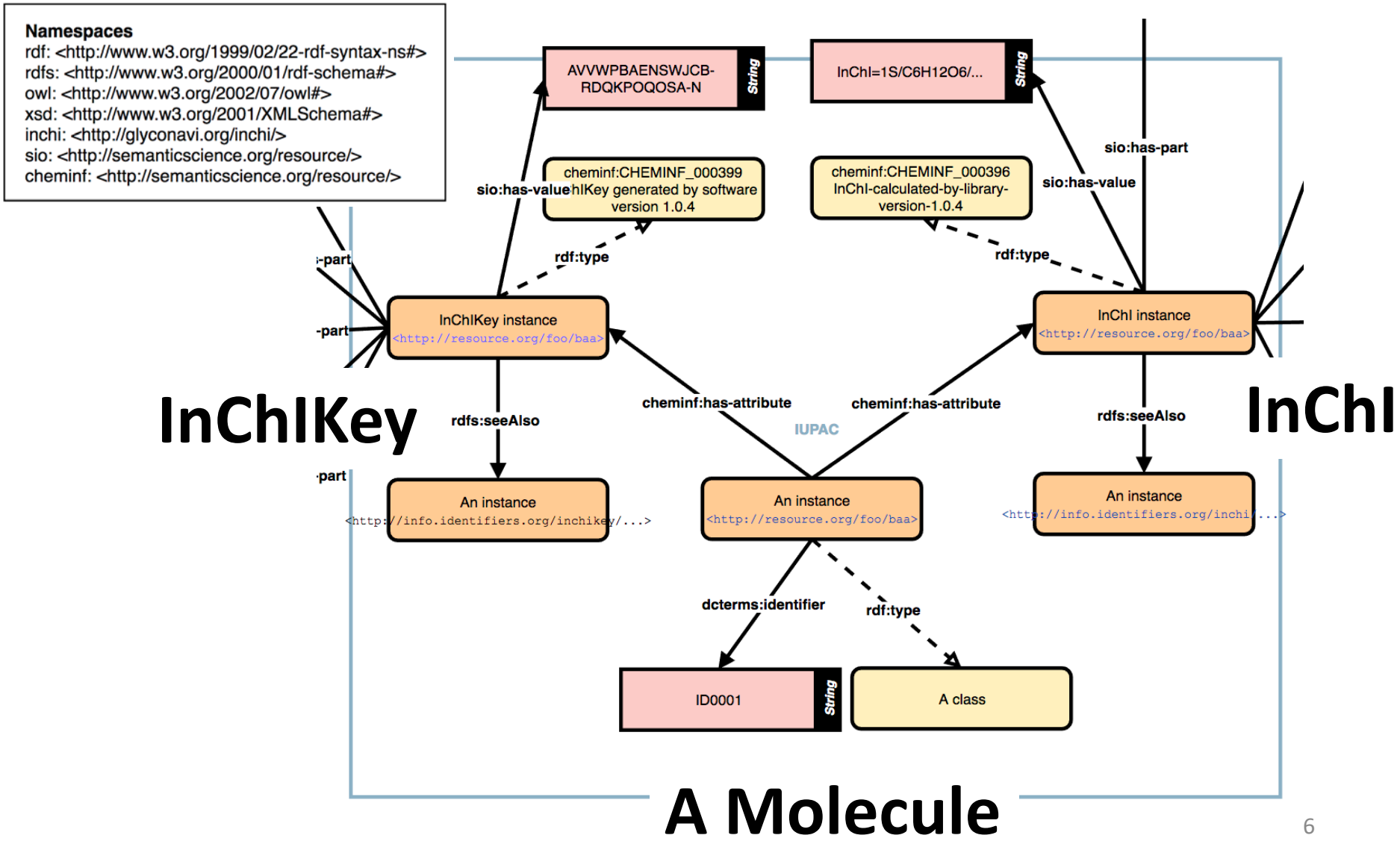
InChIKey

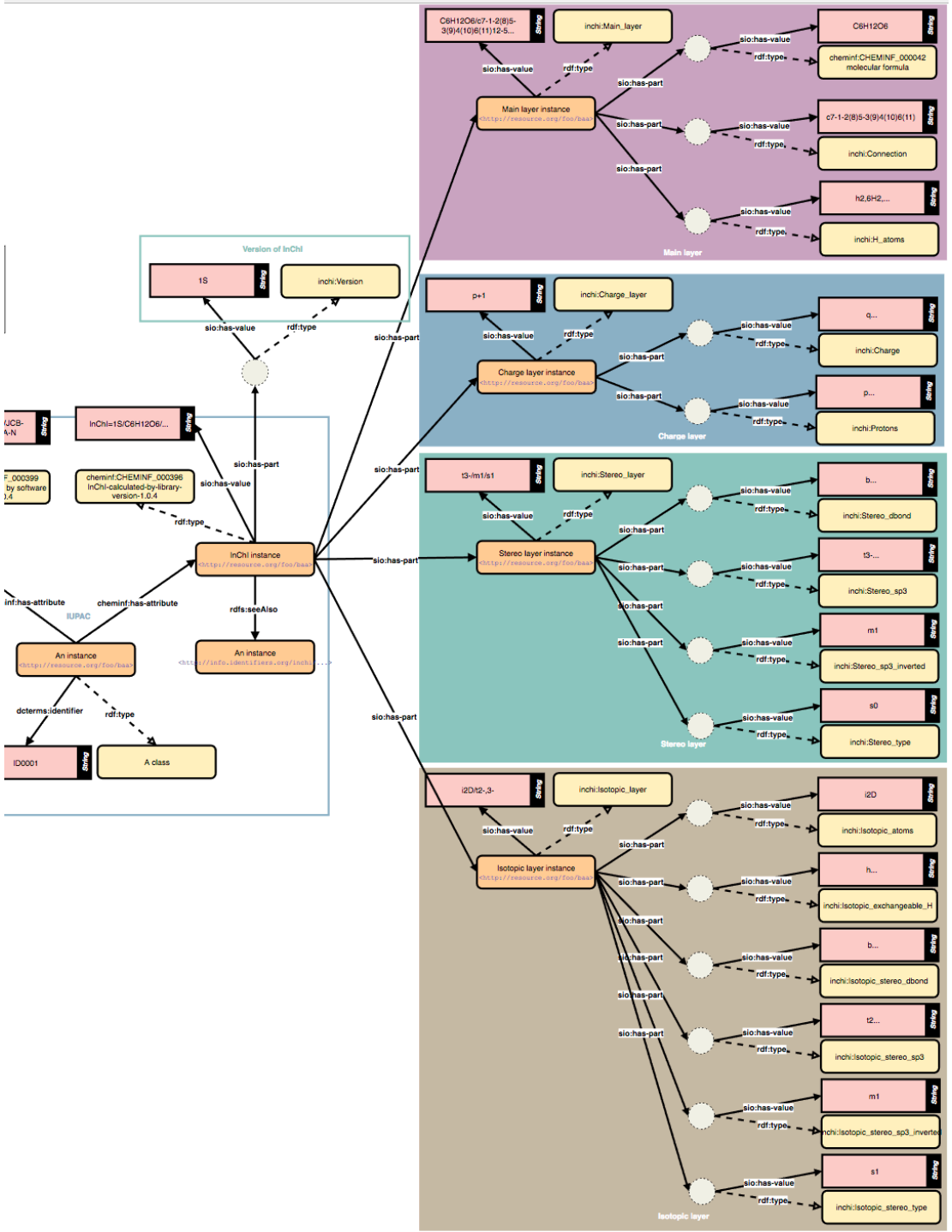


InChI

InChI-RDF

RDF Schema of InChI/InChIKey Strings





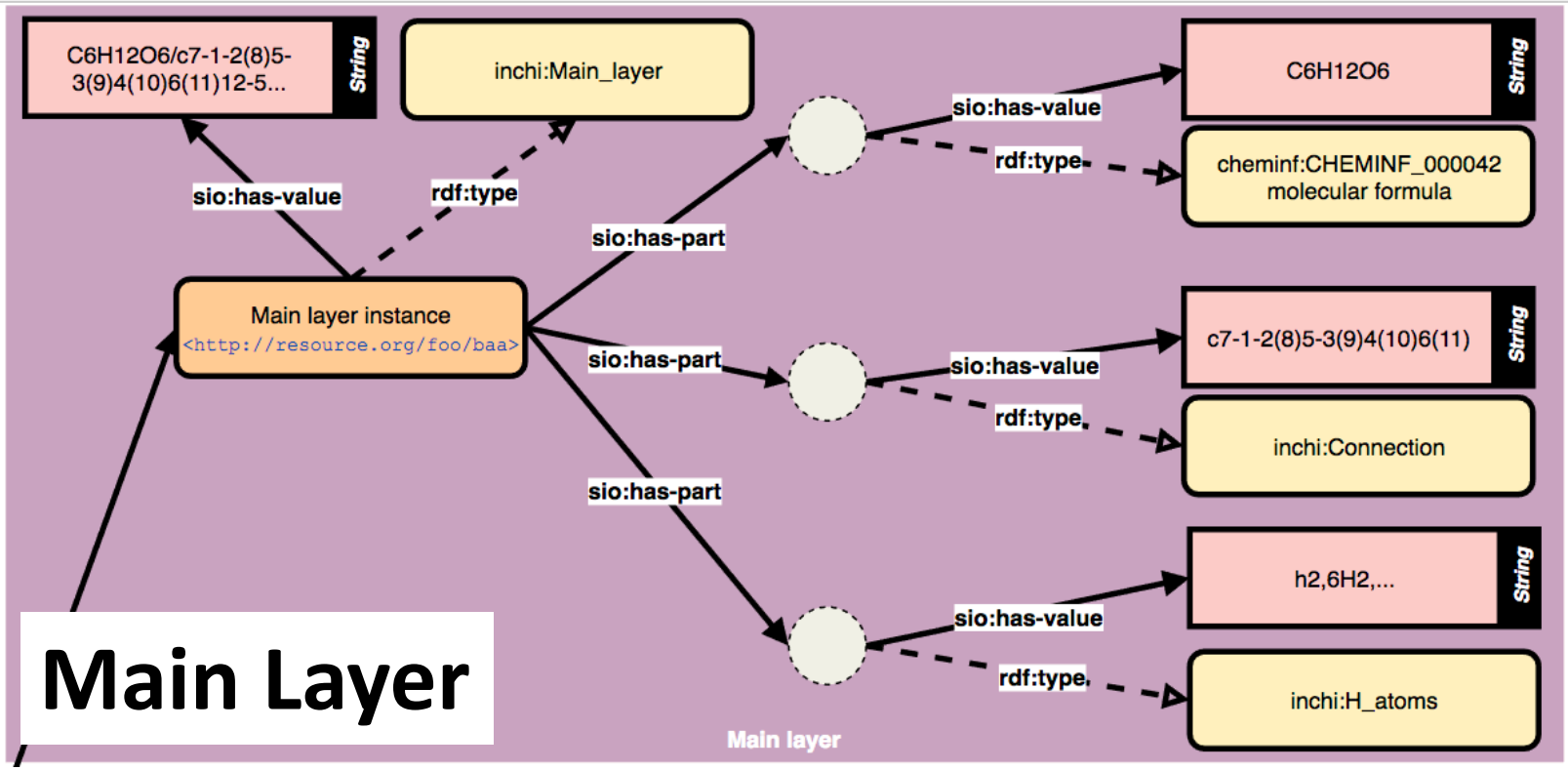
Main Layer

Charge Layer

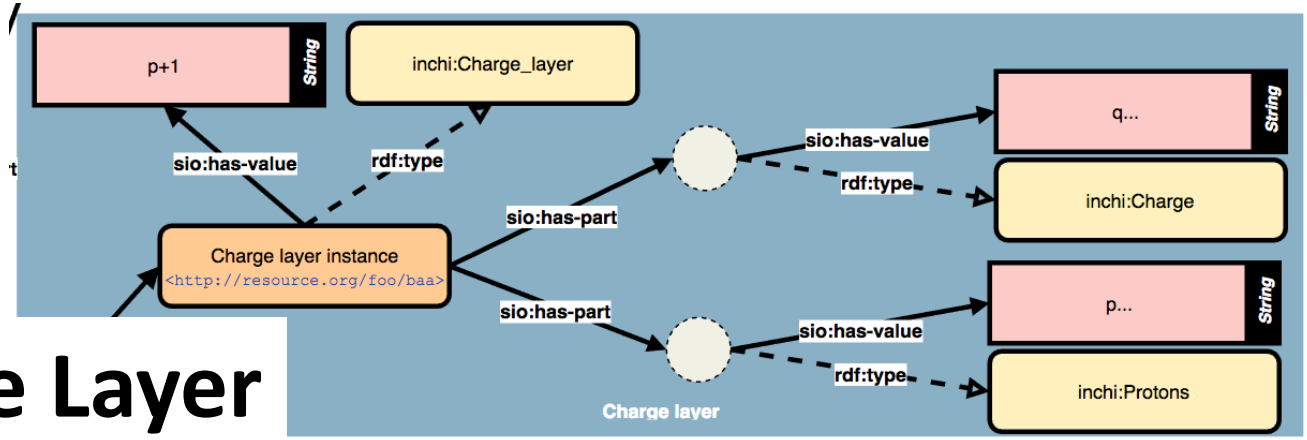
Stereo Layer

Isotopic Layer

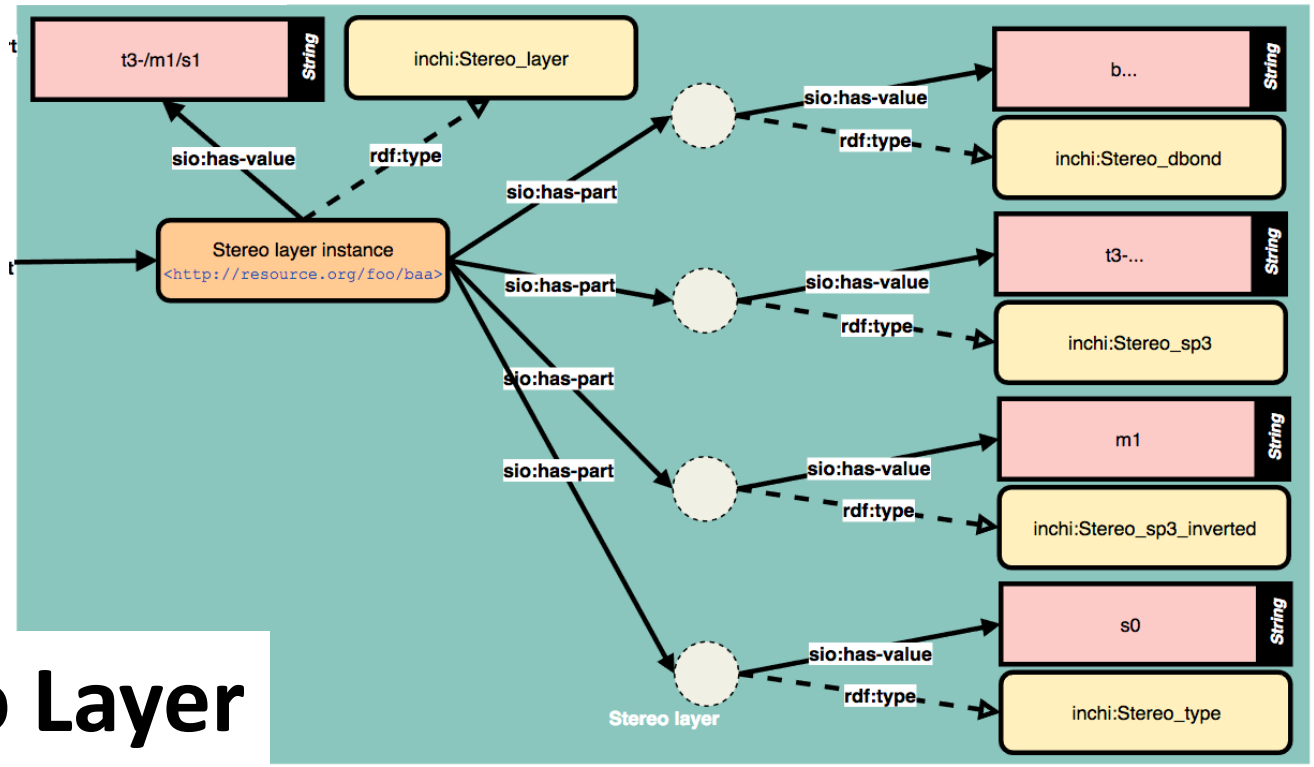
Main Layer in InChI RDF Schema



Charge and Stereo Layer



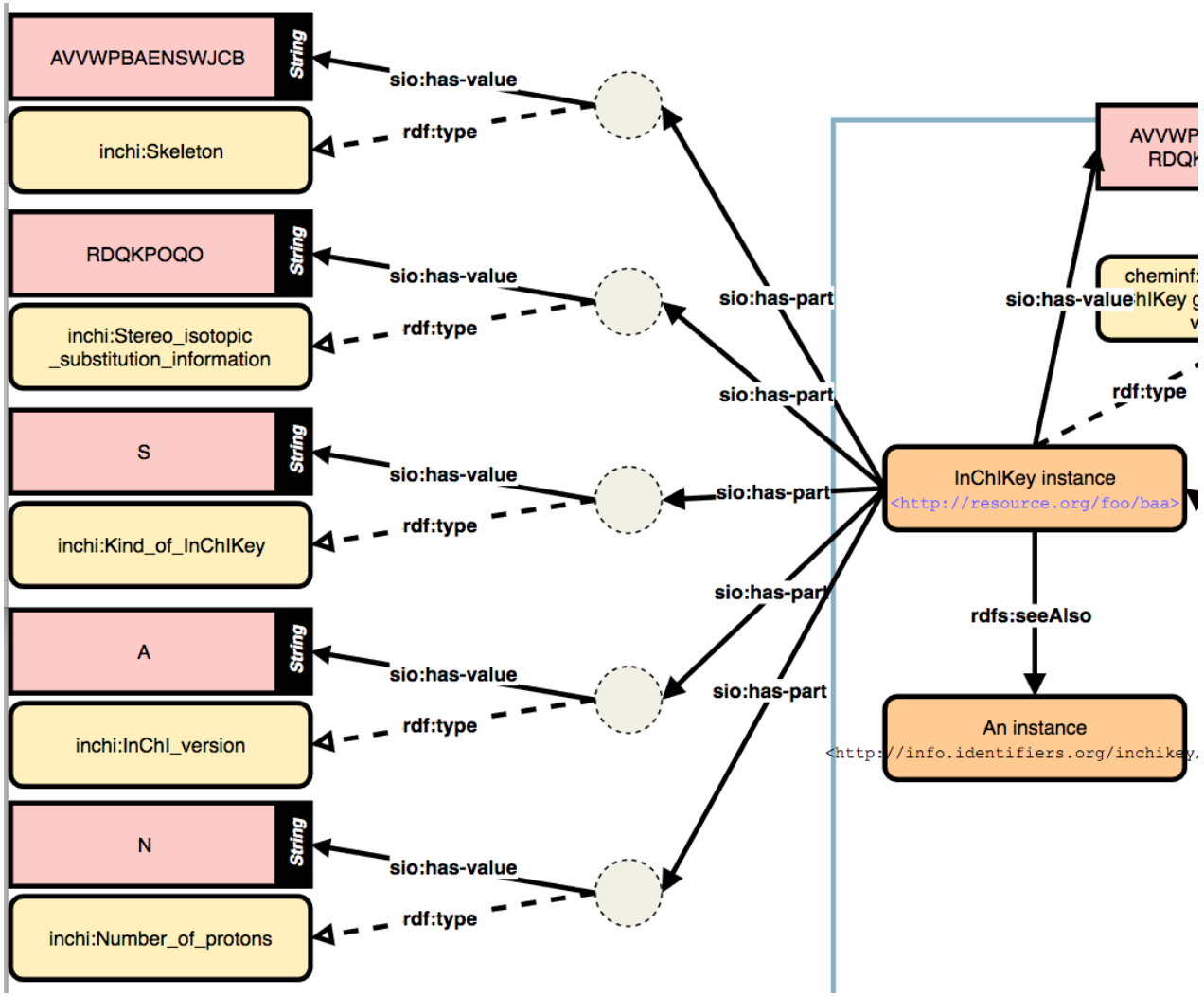
Charge Layer



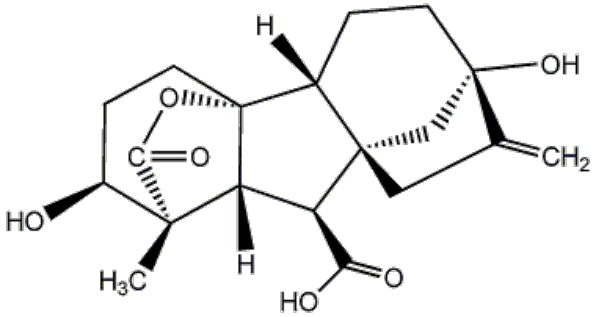
Stereo Layer

RDF Schema of InChIKey

AVVWPBAENSWJCB-RDQKPOQOSA-N



Triples of Gibberellin A1



has attribute

<http://kanaya.naist.jp/knapsack_jsp/information.jsp?word=C00000001>
a sio:SIO_011125 ; cheminf:CHEMINF_000200 <http://kanaya.naist.jp/knapsack_jsp/information.jsp?word=C00000001_IUPAC_InChI>, <http://kanaya.naist.jp/knapsack_jsp/information.jsp?word=C00000001_IUPAC_InChIKey> ; dcterms:identifier "C00000001" .

 # Triples of InChI RDF

<http://kanaya.naist.jp/knapsack_jsp/information.jsp?word=C00000001_IUPAC_InChI>
sio:SIO_000300 "**InChI=1S/C19H24O6/c1-9-7-17-8-18(9,24)5-3-10(17)19-6-4-11(20)16(2,15(23)25-19)13(19)12(17)14(21)22/h10-13,20,24H,1,3-8H2,2H3,(H,21,22)/t10-,11+,12-,13-,16-,17+,18+,19-/m1/s1**" ;
inchi:Version "1S" ;
sio:SIO_000028 <http://kanaya.naist.jp/knapsack_jsp/information.jsp?word=C00000001_IUPAC_InChI_Main_layer> ;
sio:SIO_000028 <http://kanaya.naist.jp/knapsack_jsp/information.jsp?word=C00000001_IUPAC_InChI_Stereo_layer> ;
rdfs:seeAlso <http://info.identifiers.org/inchi/InChI=1S/C19H24O6/c1-9-7-17-8-18(9,24)5-3-10(17)19-6-4-11(20)16(2,15(23)25-19)13(19)12(17)14(21)22/h10-13,20,24H,1,3-8H2,2H3,(H,21,22)/t10-,11+,12-,13-,16-,17+,18+,19-/m1/s1> ;
rdf:type cheminf:CHEMINF_000396 .

has part

molecular formula

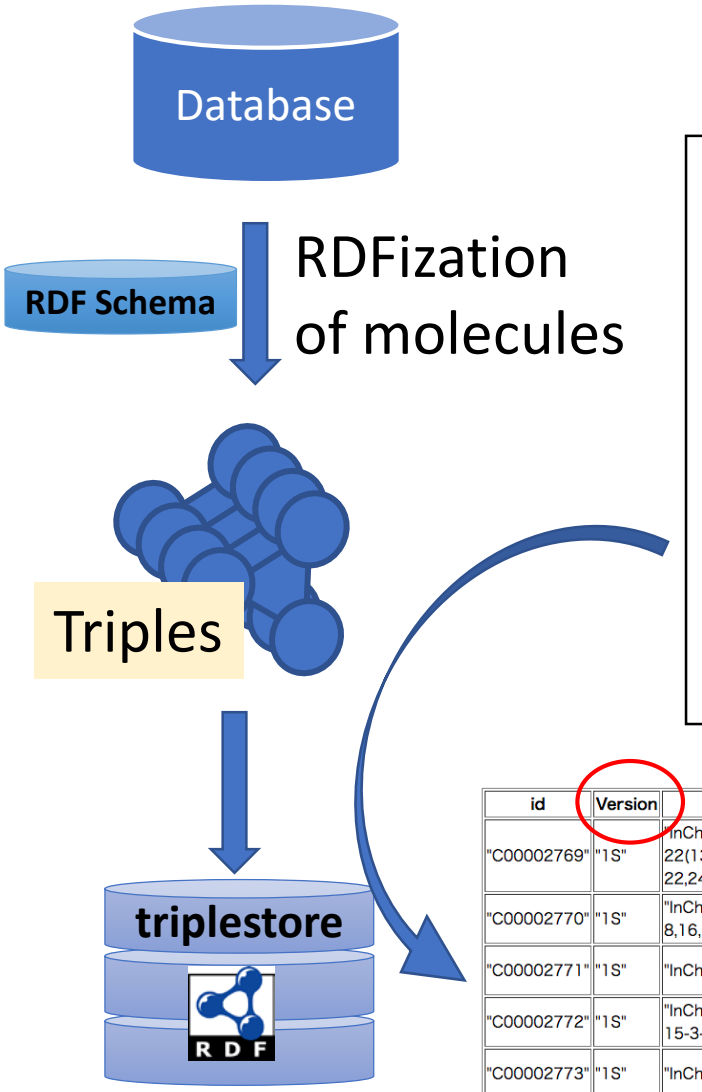
<http://kanaya.naist.jp/knapsack_jsp/information.jsp?word=C00000001_IUPAC_InChI_Main_layer>
sio:SIO_000028 [sio:SIO_000300 "**C19H24O6**" ; rdf:type cheminf:CHEMINF_000042] ;
sio:SIO_000028 [sio:SIO_000300 "**c1-9-7-17-8-18(9,24)5-3-10(17)19-6-4-11(20)16(2,15(23)25-19)13(19)12(17)14(21)22**"; rdf:type inchi:Connections] ;
sio:SIO_000028 [sio:SIO_000300 "**h10-13,20,24H,1,3-8H2,2H3,(H,21,22)"** ; rdf:type inchi:H_atoms] ;
sio:SIO_000028 [sio:SIO_000300 "**C19H24O6/c1-9-7-17-8-18(9,24)5-3-10(17)19-6-4-11(20)16(2,15(23)25-19)13(19)12(17)14(21)22/h10-13,20,24H,1,3-8H2,2H3,(H,21,22)**" ;
rdf:type inchi:Main_layer] .

<http://kanaya.naist.jp/knapsack_jsp/information.jsp?word=C00000001_IUPAC_InChI_Stereo_layer>
sio:SIO_000028 [sio:SIO_000300 "**t10-,11+,12-,13-,16-,17+,18+,19-**" ; rdf:type inchi:Stereo_sp3] ;
sio:SIO_000028 [sio:SIO_000300 "**m1**" ; rdf:type inchi:Stereo_sp3_inverted] ;
sio:SIO_000028 [sio:SIO_000300 "**s1**" ; rdf:type inchi:Stereo_type] ;
sio:SIO_000028 [sio:SIO_000300 "**t10-,11+,12-,13-,16-,17+,18+,19-/m1/s1**" ; rdf:type inchi:Stereo_layer] .

 # Triples of InChIKey RDF

<http://kanaya.naist.jp/knapsack_jsp/information.jsp?word=C00000001_IUPAC_InChIKey>
sio:SIO_000300 "**JLJLRLWOEMWYQK-OBDJNFEBSA-N**";
sio:SIO_000028 [sio:SIO_000300 "**JLJLRLWOEMWYQK**" ; rdf:type inchi:Skeleton] ;
sio:SIO_000028 [sio:SIO_000300 "**OBDJNFEB**" ; rdf:type inchi:Stereo_isotopic_substitution_information] ;
sio:SIO_000028 [sio:SIO_000300 "**S**" ; rdf:type inchi:Kind_of_InChIKey] ;
sio:SIO_000028 [sio:SIO_000300 "**A**" ; rdf:type inchi:InChI_version] ;
sio:SIO_000028 [sio:SIO_000300 "**N**" ; rdf:type inchi:Number_of_protons] ;
rdfs:seeAlso <http://info.identifiers.org/inchikey/JLJLRLWOEMWYQK-OBDJNFEBSA-N> ;
rdf:type cheminf:CHEMINF_000399 .

Search using SPARQL Query

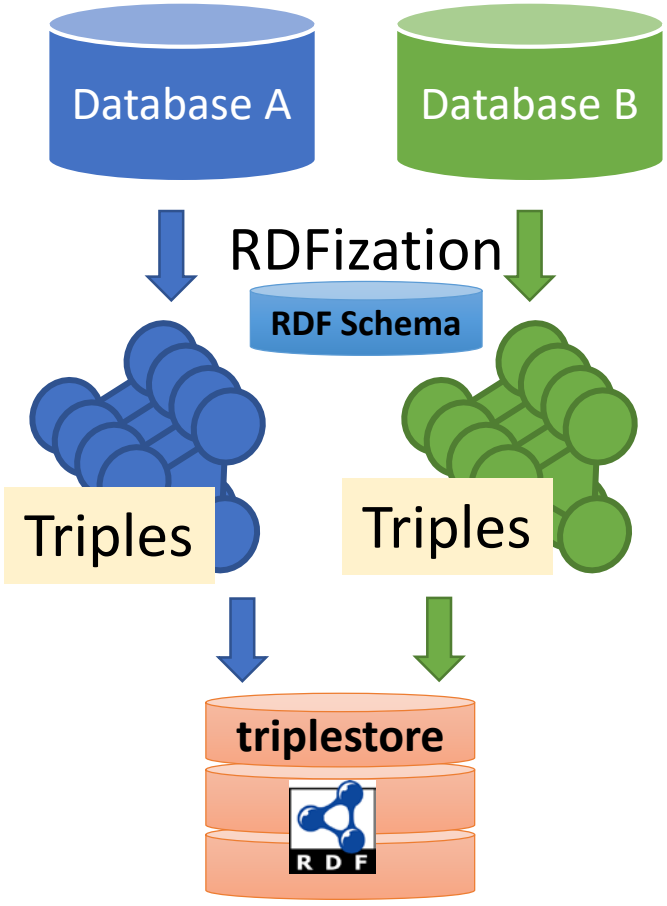


SPARQL Query

```
prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix dcterms: <http://purl.org/dc/terms/>
prefix sio: <http://semanticscience.org/resource/>
prefix inchi: <http://glyconavi.org/inchi/>

SELECT DISTINCT ?id ?Version ?inchi ?inchikey ?key_value
FROM <http://www.glyconavi.org/SPARQLthon40/iupac-inchi/KNApSAcK>
WHERE {
        ?s dcterms:identifier ?id .
        ?s sio:CHEMINF_000200 [ sio:SIO_000300 ?inchi ; rdf:type sio:CHEMINF_000396 ] .
        ?s sio:CHEMINF_000200 [ sio:SIO_000300 ?inchikey ; rdf:type sio:CHEMINF_000399 ] .
        ?s sio:CHEMINF_000200 / inchi:Version ?Version .
        ?s sio:CHEMINF_000200 / sio:SIO_000028 [ rdf:type inchi:Number_of_protons ;
        sio:SIO_000300 ?key_value  ] .
}
LIMIT 5
```

id	Version	inchi	inchikey	key_value
"C00002769"	"1S"	"InChI=1S/C35H46O20/c1-14-24(42)26(44)29(47)35(51-14)55-32-30(48)34(49-9-8-16-3-6-18(38)20(40)11-16)53-22(13-50-33-28(46)27(45)25(43)21(12-36)52-33)31(32)54-23(41)7-4-15-2-5-17(37)19(39)10-15/h2-7,10-11,14,21-22,24-40,42-48H,8-9,12-13H2,1H3/b7-4+/t14?,21?,22?,24-,25-,26-,27?,28?,29?,30?,31+,32+,33+,34+,35-/m0/s1"	"FSBUXLDOLNLABB-MQAZSWENSA-N"	"N"
"C00002770"	"1S"	"InChI=1S/C18H16O8/c19-12-4-1-10(7-14(12)21)3-6-17(23)26-16(18(24)25)9-11-2-5-13(20)15(22)8-11/h1-8,16,19-22H,9H2,(H,24,25)/b6-3+/t16-/m1/s1"	"DOUMFZQKYFQNTF-WUTVXBCWSA-N"	"N"
"C00002771"	"1S"	"InChI=1S/C10H10O2/c1-2-3-8-4-5-9-10(6-8)12-7-11-9/h2,4-6H,1,3,7H2"	"ZMQAAUBTXCXRIC-UHFFFAOYSA-N"	"N"
"C00002772"	"1S"	"InChI=1S/C26H22O10/c27-18-7-2-14(11-21(18)30)1-6-17-16(4-9-20(29)25(17)33)5-10-24(32)36-23(26(34)35)13-15-3-8-19(28)22(31)12-15/h1-12,23,27-31,33H,13H2,(H,34,35)/b6-1+,10-5+/t23-/m1/s1"	"YMGFTDKNIWPMGF-UCPJVGPRSA-N"	"N"
"C00002773"	"1S"	"InChI=1S/C11H12O3/c1-3-4-8-5-10-11(14-7-13-10)6-9(8)12-2/h3,5-6H,1,4,7H2,2H3"	"FYRHTIWFKXZWAD-UHFFFAOYSA-N"	"N"

Skeleton Match using a Single Query



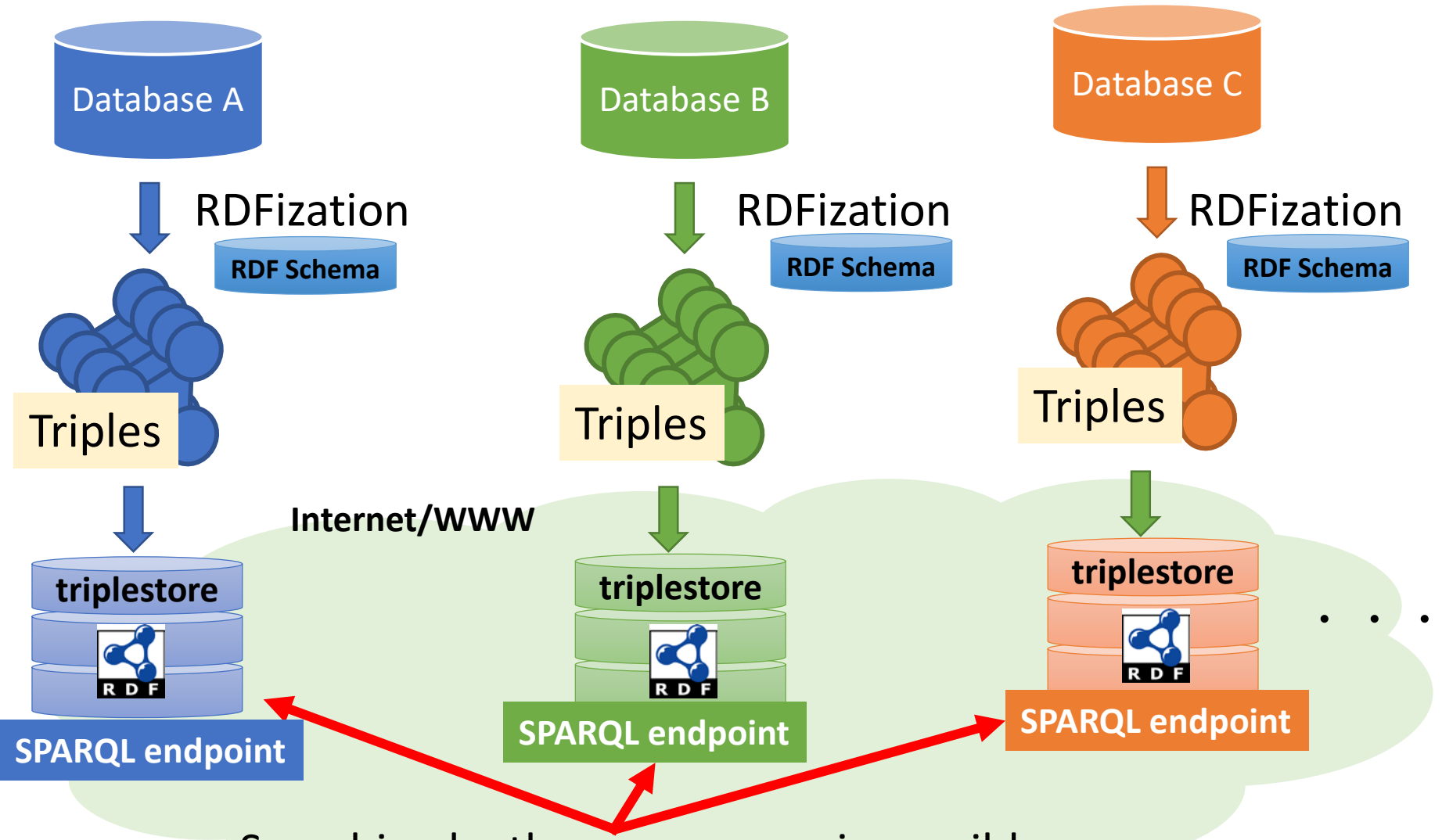
SPARQL Query

```
prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix dcterms: <http://purl.org/dc/terms/>
prefix sio: <http://semanticscience.org/resource/>
prefix inchi: <http://glyconavi.org/inchi/>
prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix dcterms: <http://purl.org/dc/terms/>

SELECT
DISTINCT ?knapsack_id  ?knapsack_inchi ?knapsack_inchikey ?Skeleton ?nikkaji_inchikey ?nikkaji_inchi ?nikkaji_id
WHERE {

GRAPH <http://www.glyconavi.org/SPARQLthon40/iupac-inchi/KNApSAcK> {
 ?s sio:CHEMINF_000200 [ sio:SIO_000300 ?knapsack_inchikey ; rdf:type sio:CHEMINF_000399 ] .
 ?s sio:CHEMINF_000200 / sio:SIO_000028 [ rdf:type inchi:Skeleton ; sio:SIO_000300 ?Skeleton  ] .}

GRAPH <http://www.glyconavi.org/iupac-inchi/NIKKAJI> {
 ?nikkaji_s <http://vocab.jst.go.jp/terms/sti#InChIKey> ?nikkaji_inchikey .
 FILTER ( ?knapsack_inchikey != ?nikkaji_inchikey &&  contains(str(?nikkaji_inchikey), ?Skeleton ))
 ?nikkaji_s <http://purl.org/dc/terms/identifier> ?nikkaji_id .
 ?nikkaji_s <http://vocab.jst.go.jp/terms/sti#InChI> ?nikkaji_inchi .}

?s sio:CHEMINF_000200 [ sio:SIO_000300 ?knapsack_inchi ; rdf:type sio:CHEMINF_000396 ] .
?s sio:CHEMINF_000200 [ sio:SIO_000300 ?knapsack_inchikey ; rdf:type sio:CHEMINF_000399 ] .
?s dcterms:identifier ?knapsack_id .
}
Limit 1
```

C00002769	FSBUXLDOLNLABB-**MQAZSWEN**SA-N
J362.655E	FSBUXLDOLNLABB-**ISAKITKM**SA-N

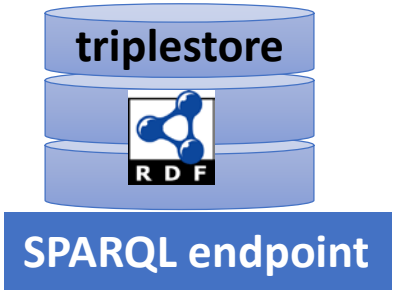
knapsack_id	knapsack_inchi	knapsack_inchikey	Skeleton	nikkaji_inchikey	nikkaji_inchi	nikkaji_id
"C00002769"	"InChI=1S/C35H46O20/c1-14-24(42)26(44)29(47)35(51-14)55-32-30(48)34(49-9-8-16-3-6-18(38)20(40)11-16)53-22(13-50-33-28(46)27(45)25(43)21(12-36)52-33)31(32)54-23(41)7-4-15-2-5-17(37)19(39)10-15/h2-7,10-11,14,21-22,24-40,42-48H,8-9,12-13H2,1H3/b7-4+/t14?,21?,22?,24-,25-,26-,27?,28?,29?,30?,31+,32+,33+,34+,35-/m0/s1"	"FSBUXLDOLNLABB-MQAZSWENSA-N"	"FSBUXLDOLNLABB"	"FSBUXLDOLNLABB-ISAKITKMSA-N"	"InChI=1S/C35H46O20/c1-14-24(42)26(44)29(47)35(51-14)55-32-30(48)34(49-9-8-16-3-6-18(38)20(40)11-16)53-22(13-50-33-28(46)27(45)25(43)21(12-36)52-33)31(32)54-23(41)7-4-15-2-5-17(37)19(39)10-15/h2-7,10-11,14,21-22,24-40,42-48H,8-9,12-13H2,1H3/b7-4+/t14-,21+,22+,24-,25+,26+,27-,28+,29+,30+,31+,32+,33+,34+,35-/m0/s1"	"J362.655E"

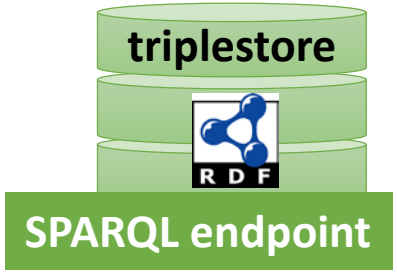
Database Integration



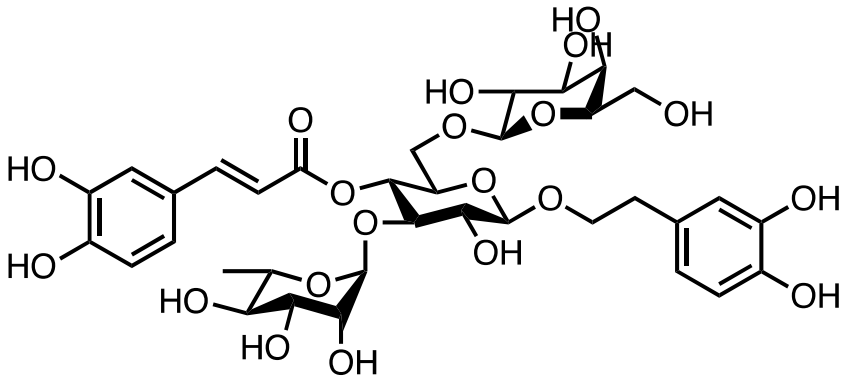
Searching by the same query is possible.
You can search all databases with a single query.

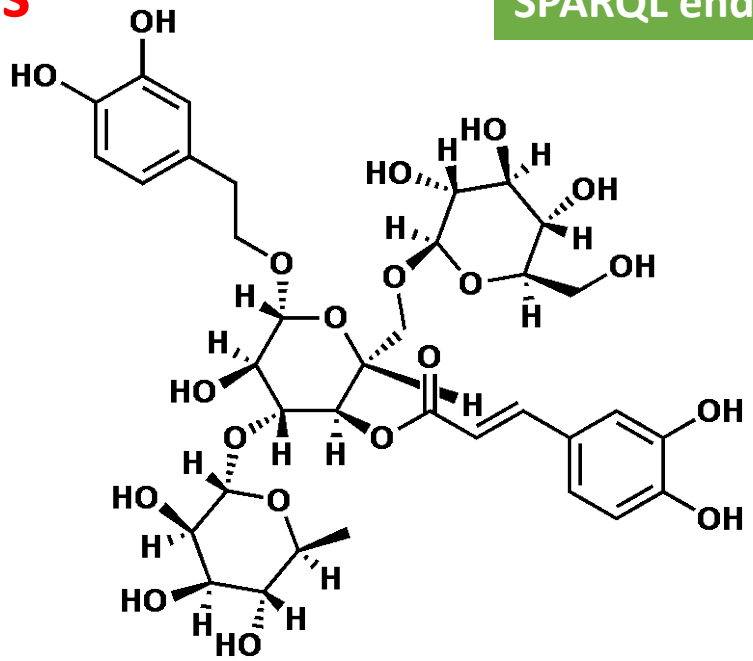
Skeleton Matching in Different Databases





Same molecules

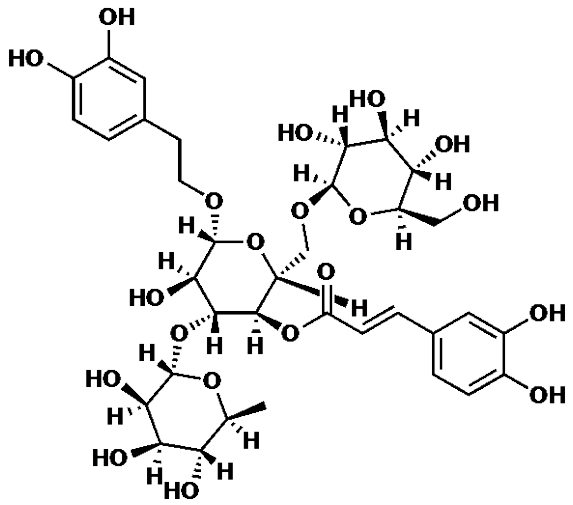




Skeleton	Stereo
FSBUXLDOLNLABB-	**MQAZSWEN**SA-N

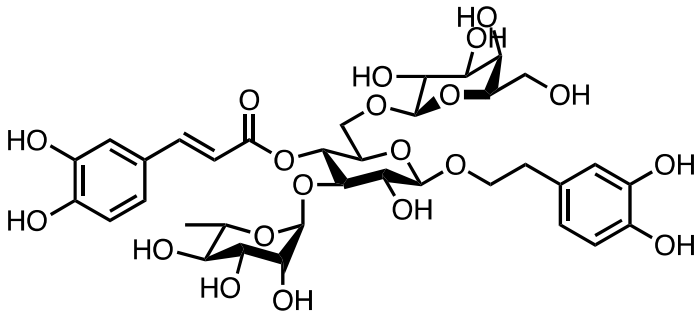
Skeleton	Stereo
FSBUXLDOLNLABB-	**ISAKITKM**SA-N

Information of Stereochemistry



InChI=1S/C35H46O20/c1-14-24(42)26(44)29(47)35(51-14)55-32-30(48)34(49-9-8-16-3-6-18(38)20(40)11-16)53-22(13-50-33-28(46)27(45)25(43)21(12-36)52-33)31(32)54-23(41)7-4-15-2-5-17(37)19(39)10-15/h2-7,10-11,14,21-22,24-40,42-48H,8-9,12-13H2,1H3/b7-4+/t14-,21+,22+,24-,25+,26+,27-,28+,29+,30+,31+,32+,33+,34+,35-/m0/s1

Confirmation of stereochemistry is difficult.



InChI=1S/C35H46O20/c1-14-24(42)26(44)29(47)35(51-14)55-32-30(48)34(49-9-8-16-3-6-18(38)20(40)11-16)53-22(13-50-33-28(46)27(45)25(43)21(12-36)52-33)31(32)54-23(41)7-4-15-2-5-17(37)19(39)10-15/h2-7,10-11,14,21-22,24-40,42-48H,8-9,12-13H2,1H3/b7-4+/t14-,21+,22+,24-,25-,26+,27-,28?,29+,30+,31+,32+,33+,34+,35-/m0/s1

WURCS and GlyTouCan

- **WURCS is the Web3 Unique Representation of Carbohydrate Structures.**
  - We have developed tools for generation of WURCS string from Molfile/SDFile.
    - *J Chem Inf Model.* 2014 Jun 23;54(6):1558-66. PMID: 24897372.
    - *J Chem Inf Model.* 2017 Apr 24;57(4):632-637. PMID: 28263066.

- **GlyTouCan is an International Glycan Structure Repository** (https://glytoucan.org)**.**
  - It stored many glycan structures.
  - It can be used to verify glycan structures.
  - It has a SPARQL endpoint.
    - *Nucleic Acids Res.* 2016 Jan 4;44(D1):D1237-42. PMID: 26476458.

Workflow of Glycan Validation

Glycans

WURCS

WURCS-RDF

Triples

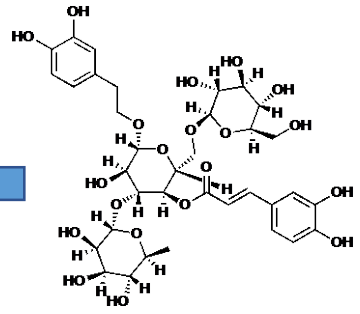
GlyTouCan

SPARQL endpoint

GlycoNAVI

Support System for Carbohydrate Research

molToWURCS Software



WURCS=2.0/3,3,2/[a2122h-1b_1-5_4*OCC=^EC(CC^ECC^ECC$6)/9O/8O/3=O][a2211m-1a_1-5][a2122h-1b_1-5]/1-2-3/a3-b1_a6-c1

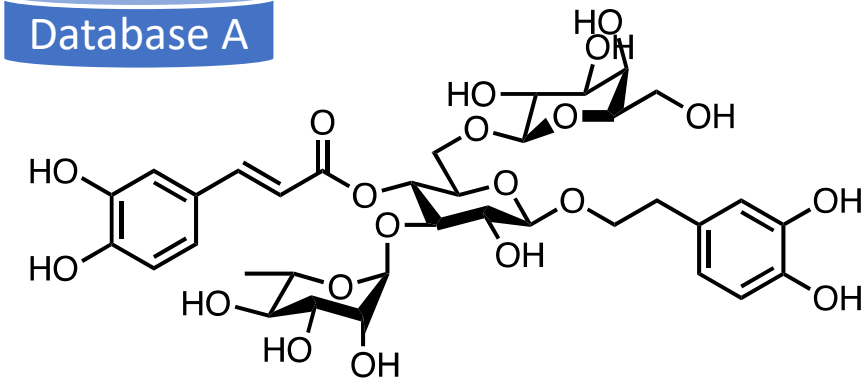
SPARQL Query

WURCS=2.0/1,1,0/[a2122h-1b_1-5]/1/
WURCS=2.0/1,1,0/[a2122h-1b_1-5]/1/
WURCS=2.0/1,1,0/[a2122m-1a_1-5]/1/

Accession number

Validation of molecules using GlyTouCan and WURCS

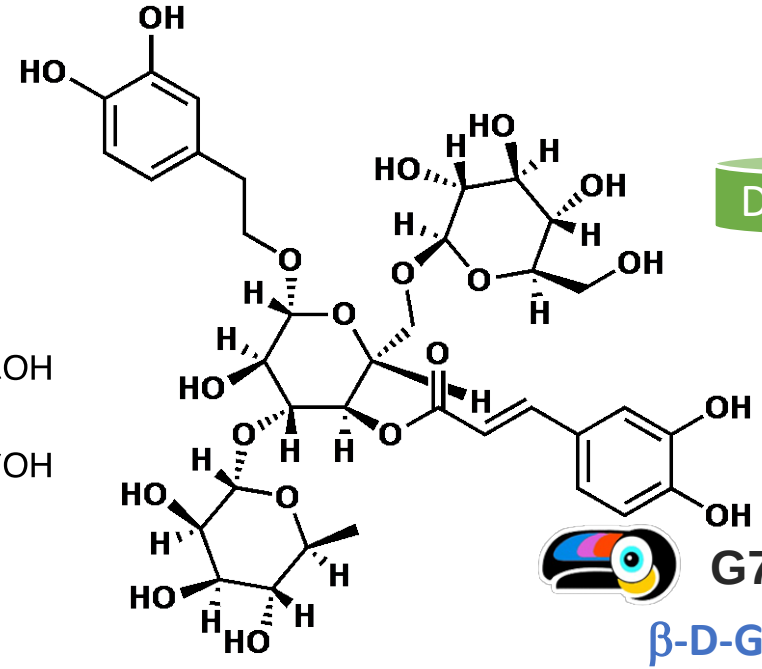
Database A



not registered

WURCS=2.0/3,3,2/[ax12xh-1x_1-5_4*OCC=^EC(C^ECC^ZCC^ZC$6)/9O/8O/3=O][ax11xm-1x_1-5][axx12h-1b_1-5]/1-2-3/a3-b1_a6-c1

Database B



G71142DF

β-D-Glcp

WURCS=2.0/3,3,2/[a2122h-1b_1-5_4*OCC=^EC(CC^ECC^ECC$6)/9O/8O/3=O][a2211m-1a_1-5][a2122h-1b_1-5]/1-2-3/a3-b1_a6-c1

α-L-Rhap

G38999IM

GlyTouCan and WURCS can be used as to whether the monosaccharide structure has usual stereochemistry.

Conclusion

- We have developed InChI/InChIKey triples and RDF Schema.
- InChI-RDF format is based upon InChI/InChIKey Layers in InChI documentation.
- We can obtain information from multiple databases with single SPARQL query.
- By using WURCS and GlyTouCan, we presented the possibility of glycan structure verification.

Acknowledgements

- We thank Toshiaki TOKIMATSU (DDBJ) and Tatsuya KUSHIDA (NBDC) for helpful and constructive comments and discussions .
- This research was supported by JSPS KAKENHI Grant Number JP16K00412.
- This research was supported by Japan Science and Technology Agency (JST), National Bioscience Database Center (NBDC).