

UniProt SPARQL: A small tutorial



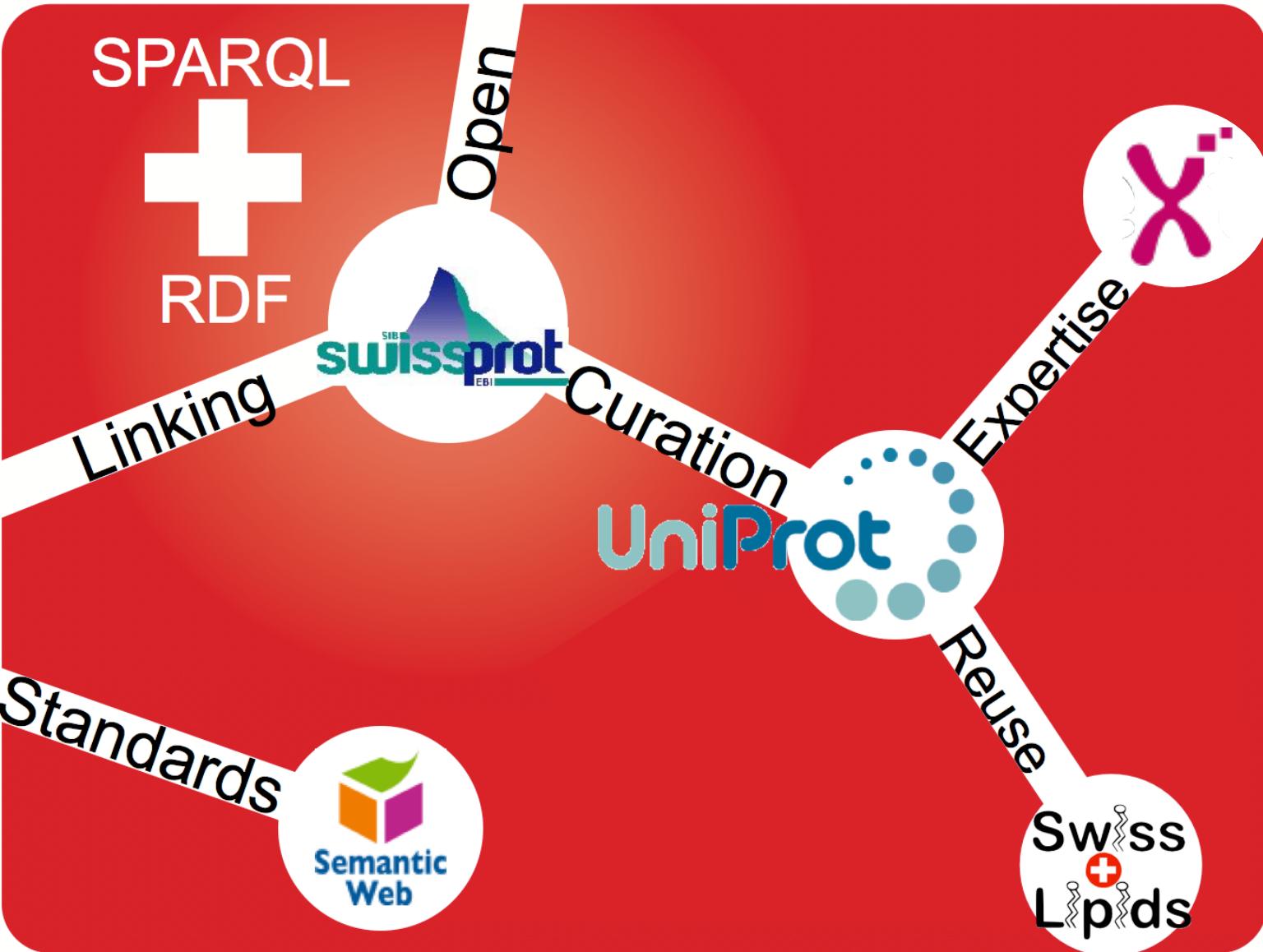
Jerven Bolleman
Swiss-Prot



Swiss Institute of
Bioinformatics

UniProt SPARQL

- Next to www.uniprot.org
 - for analytics
 - answering questions
 - not aimed at keyword search
 - your UniProt Data Base in the cloud
- 14 billion triples
 - in 17 graphs
- help@uniprot.org
 - interested in your queries/questions/comments
 - we want you to use our data!
 - all questions are welcome



Your SPARQL query

[Add common prefixes](#)

1

[Submit Query](#)

About

This SPARQL endpoint contains all UniProt data. It is free to access and supports the [SPARQL 1.1 Standard](#).

There are 13,948,325,880 triples in this release (2015_06). All triples are available in the default graph. There are 17 named graphs.

Documentation

The documentation about UniProt RDF is spread into 2 parts

1. [Classes and predicates defined by the UniProt consortium](#)
2. [Statistics and diagrams](#)

News



POLQ, a new target for cancer therapy?

[UniProt release 2015_06](#)

A never-ending race between evolution and genomic integrity | Removal of IPI species proteome data sets from FTP site | UniProtKB XSD change for evidence attribution

[UniProt release 2015_05](#)

Of CAT tails and protein translation by-products | Reducing redundancy in proteomes | Changes to the controlled vocabulary of human diseases

[UniProt release 2015_04](#)

[News archive](#)

Examples

1. Select all taxa from the [UniProt taxonomy](#): ([show](#))
2. Select all bacterial taxa, and their scientific name, from the [UniProt taxonomy](#): ([show](#))
3. Select all E-Coli K12 (including strains) UniProt entries and their amino acid sequence: ([show](#))
4. Select the UniProt entry with the mnemonic 'A4_HUMAN': ([show](#))
5. Select a mapping of UniProt to PDB entries using the UniProt cross-references to the [PDB database](#): ([show](#))
6. Select all cross-references to external databases of the category '3D structure databases' of UniProt entries that are classified with the keyword '3Fe-4S': ([show](#))
7. Select all UniProt entries, and their recommended protein name, that have a preferred gene name that contains the text 'DNA': ([show](#))
8. Select the preferred gene name and disease annotation of all human UniProt entries that are known to be involved in a disease: ([show](#))
9. Select all human UniProt entries with a sequence variant that leads to a 'loss of function': ([show](#))
10. Select all human UniProt entries with a sequence variant that leads to a tyrosine to phenylalanine substitution: ([show](#))
11. Select all UniProt entries with annotated transmembrane regions and the regions' begin and end coordinates on the canonical sequence: ([show](#))
12. Select all UniProt entries that were integrated on the 30th of November 2010: ([show](#))
13. Was any UniProt entry integrated on the 9th of January 2013? ([show](#))
14. Construct new triples of the type 'HumanProtein' from all human UniProt entries: ([show](#))
15. Select all triples that relate to the EMBL CDS entry AA089367.1: ([show](#))
16. Select all triples that relate to the taxon that describes *Homo sapiens*: ([show](#))
17. Select the average number of cross-references to the [PDB database](#) of UniProt entries that have at least one cross-reference to the PDB database: ([show](#))
18. Select the number of UniProt entries for each of the EC (Enzyme Commission) second level categories: ([show](#))

UniProt

SPARQL Downloads Documentation/Help

Your SPARQL query

Add common prefixes

```
1
```

Submit Query

About

This SPARQL endpoint contains all UniProt data. It is free to access and supports the SPARQL 1.1 Standard.

There are 13,948,325,880 triples in this release (2015_06). All triples are available in the default graph. There are 17 named graphs.

Documentation

The documentation about UniProt RDF is spread into 2 parts

1. Classes and predicates defined by the UniProt consortium
2. Statistics and diagrams

News

POLQ, a new target for cancer therapy? [UniProt release 2015_06](#)

A never-ending race between evolution and genomic integrity | Removal of IPI species proteome data sets from FTP site | UniProtKB XSD change for evidence attribution [UniProt release 2015_05](#)

Of CAT tails and protein translation by-products | Reducing redundancy in proteomes | Changes to the controlled vocabulary of human diseases [UniProt release 2015_04](#)

[News archive](#)

Examples

1. Select all taxa from the UniProt taxonomy: ([show](#))
2. Select all bacterial taxa, and their scientific name, from the UniProt taxonomy: ([show](#))
3. Select all E-Coli K12 (including strains) UniProt entries and their amino acid sequence: ([show](#))
4. Select the UniProt entry with the mnemonic 'A4_HUMAN': ([show](#))
5. Select a mapping of UniProt to PDB entries using the UniProt cross-references to the PDB database: ([show](#))
6. Select all cross-references to external databases of the category '3D structure databases' of UniProt entries that are classified with the keyword '3Fe-4S': ([show](#))
7. Select all UniProt entries, and their recommended protein name, that have a preferred gene name that contains the text 'DNA': ([show](#))
8. Select the preferred gene name and disease annotation of all human UniProt entries that are known to be involved in a disease: ([show](#))
9. Select all human UniProt entries with a sequence variant that leads to a 'loss of function': ([show](#))
10. Select all human UniProt entries with a sequence variant that leads to a tyrosine to phenylalanine substitution: ([show](#))
11. Select all UniProt entries with annotated transmembrane regions and the regions' begin and end coordinates on the canonical sequence: ([show](#))
12. Select all UniProt entries that were integrated on the 30th of November 2010: ([show](#))
13. Was any UniProt entry integrated on the 9th of January 2013? ([show](#))
14. Construct new triples of the type 'HumanProtein' from all human UniProt entries: ([show](#))
15. Select all triples that relate to the EMBL CDS entry AA089367.1: ([show](#))
16. Select all triples that relate to the taxon that describes *Homo sapiens*: ([show](#))
17. Select the average number of cross-references to the PDB database of UniProt entries that have at least one cross-reference to the PDB database: ([show](#))
18. Select the number of UniProt entries for each of the EC (Enzyme Commission) second level categories: ([show](#))

Your SPARQL query

[Add common prefixes](#)

1

[Submit Query](#)

About

This SPARQL endpoint contains all UniProt data. It is free to access and supports the SPARQL 1.1 Standard.

There are 13,948,325,880 triples in this release (2015_06). All triples are available in the default graph. There are 17 named graphs.

Documentation

The documentation about UniProt RDF is spread into 2 parts

1. [Classes and predicates defined by the UniProt consortium](#)
2. [Statistics and diagrams](#)

News



POLQ, a new target for cancer therapy?

[UniProt release 2015_06](#)

A never-ending race between evolution and genomic integrity | Removal of IPI species proteome data sets from FTP site | UniProtKB XSD change for evidence attribution

[UniProt release 2015_05](#)

Of CAT tails and protein translation by-products | Reducing redundancy in proteomes | Changes to the controlled vocabulary of human diseases

[UniProt release 2015_04](#)

[News archive](#)

Examples

1. Select all taxa from the UniProt taxonomy: ([show](#))
2. Select all bacterial taxa, and their scientific name, from the UniProt taxonomy: ([show](#))
3. Select all E-Coli K12 (including strains) UniProt entries and their amino acid sequence: ([show](#))
4. Select the UniProt entry with the mnemonic 'A4_HUMAN': ([show](#))
5. Select a mapping of UniProt to PDB entries using the UniProt cross-references to the PDB database: ([show](#))
6. Select all cross-references to external databases of the category '3D structure databases' of UniProt entries that are classified with the keyword '3Fe-4S': ([show](#))
7. Select all UniProt entries, and their recommended protein name, that have a preferred gene name that contains the text 'DNA': ([show](#))
8. Select the preferred gene name and disease annotation of all human UniProt entries that are known to be involved in a disease: ([show](#))
9. Select all human UniProt entries with a sequence variant that leads to a 'loss of function': ([show](#))
10. Select all human UniProt entries with a sequence variant that leads to a tyrosine to phenylalanine substitution: ([show](#))
11. Select all UniProt entries with annotated transmembrane regions and the regions' begin and end coordinates on the canonical sequence: ([show](#))
12. Select all UniProt entries that were integrated on the 30th of November 2010: ([show](#))
13. Was any UniProt entry integrated on the 9th of January 2013? ([show](#))
14. Construct new triples of the type 'HumanProtein' from all human UniProt entries: ([show](#))
15. Select all triples that relate to the EMBL CDS entry AA089367.1: ([show](#))
16. Select all triples that relate to the taxon that describes *Homo sapiens*: ([show](#))
17. Select the average number of cross-references to the PDB database of UniProt entries that have at least one cross-reference to the PDB database: ([show](#))
18. Select the number of UniProt entries for each of the EC (Enzyme Commission) second level categories: ([show](#))

<http://www.uniprot.org/core/>

Core

Navigation Panel

Classes (167)	Absorption_Annotation
Properties (138)	
Object properties (57)	alternativeName
Datatype properties (81)	abstract

Ontology description

<http://purl.uniprot.org/core/> (rdf:type owl:Ontology)

rdfs:comment Properties and classes used for protein annotation. xsd:string

Classes

Absorption_Annotation (rdf:type owl:Class)

rdfs:label	Absorption xsd:string
rdfs:subClassOf	Biophysicochemical_Annotation
rdfs:comment	Indicates the wavelength in nm at which photoreactive proteins such as opsins and DNA photolyases show maximal absorption. xsd:string

Active_Site_Annotation (rdf:type owl:Class)

rdfs:label	Active Site xsd:string
rdfs:subClassOf	Site_Annotation
rdfs:comment	Amino acid(s) involved in the activity of an enzyme. xsd:string

<http://www.uniprot.org/format/>

!! work in progress !!

- Description of the different formats
 - Examples in Turtle and JSON-LD
- Bug reports very welcome
- Focuses on UniProtKB

RDF can be downloaded from uniprot.org

Filter by

- Reviewed (548,586)
Swiss-Prot
- Unreviewed (48,744,721)
EMBL
- Popular organisms
 - Human (145,892)
 - Mouse (99,892)
 - Mouse (76,082)
 - Zebrafish (56,536)
 - C. thaliana* (52,326)
 - Other organisms

View by

- taxonomy
- Keywords
- Gene Ontology
- Enzyme class
- Pathway
- UniRef

Your results in sequence clusters with identity of:

Entry	Entry	Gene names	Organism	Length
<input type="checkbox"/> Q6GZX3	002L	002L	Frog virus 3 (isolate Goorha) (FV-3)	320
<input type="checkbox"/> Q6GZX4	001R	002R	Frog virus 3 (isolate Goorha) (FV-3)	256
<input type="checkbox"/> Q197F7	003L	003L	Invertebrate iridescent virus 3 (IIV-3) (Mosquito iridescent virus)	156
<input type="checkbox"/> Q197F8	002R_IIV3	Uncharacterized protein 002R	Invertebrate iridescent virus 3 (IIV-3) (Mosquito iridescent virus)	458
<input type="checkbox"/> Q6GZX2	003R_FRG3G	Uncharacterized protein 3R	Frog virus 3 (isolate Goorha) (FV-3)	438
<input type="checkbox"/> Q6GZX1	004R_FRG3G	Uncharacterized protein 004R	Frog virus 3 (isolate Goorha) (FV-3)	60
<input type="checkbox"/> Q197F5	005L_IIV3	Uncharacterized protein 005L	Invertebrate iridescent virus 3 (IIV-3) (Mosquito iridescent virus)	217
<input type="checkbox"/> Q6GZX0	005R_FRG3G	Uncharacterized protein 005R	Frog virus 3 (isolate Goorha) (FV-3)	204
<input type="checkbox"/> Q91G88	006L_IIV6	Putative Kila-N domain-containing p...	IIV6-006L	352
<input type="checkbox"/> Q6GZW9	006R_FRG3G	Uncharacterized protein 006R	Frog virus 3 (isolate Goorha) (FV-3)	75
<input type="checkbox"/> Q6GZW8	007R_FRG3G	Uncharacterized protein 007R	Frog virus 3 (isolate Goorha) (FV-3)	128
<input type="checkbox"/> Q197F3	007R_IIV3	Uncharacterized protein 007R	Invertebrate iridescent virus 3 (IIV-3) (Mosquito iridescent virus)	447
<input type="checkbox"/> Q197F2	008L_IIV3	Uncharacterized protein 008L	Invertebrate iridescent virus 3 (IIV-3) (Mosquito iridescent virus)	347
<input type="checkbox"/> Q6GZW6	009L_FRG3G	Putative helicase 009L	Frog virus 3 (isolate Goorha) (FV-3)	948
<input type="checkbox"/> Q91G85	009R_IIV6	Uncharacterized protein 009R	Invertebrate iridescent virus 6 (IIV-6) (Chilo iridescent virus)	85

1 to 25 of 49,293,307 ► Show 25

Download Add to basket Columns

Download selected (0)
 Download all (49293307)

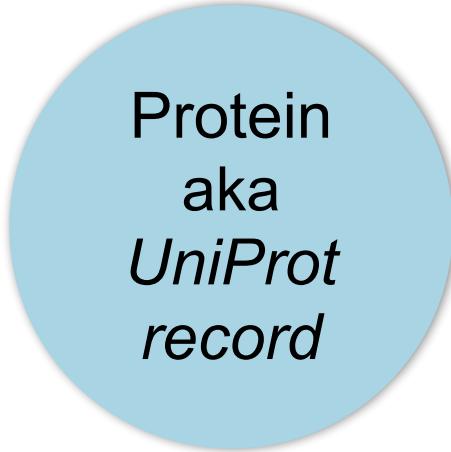
Format: RDF/XML
 Compressed Uncompressed

Preview first 10

The screenshot shows a search results page from Uniprot.org. A red box highlights a download modal window. The modal contains options to 'Download selected (0)' or 'Download all (49293307)', a 'Format' dropdown set to 'RDF/XML', and a radio button for 'Compressed' (which is checked). Below these are buttons for 'Preview first 10' and 'Go'. The main table lists protein entries with columns for Entry, Gene names, Organism, and Length. Several entries are highlighted with yellow star icons and bolded gene names, such as 'Uncharacterized protein 002R', 'Uncharacterized protein 3R', and 'Uncharacterized protein 004R'.

[Download](#)[Add to basket](#)[Columns](#) Download selected (0) Download all (49293307)**Format:**  Compressed Uncompressed[Preview first 10](#)[Go](#)

UniProtKB



UniRef

Cluster
aka
UniRef record

50%

90%

100%

UniParc

Sequence
aka
UniParc

Taxonomy

UniProtKB

Protein
aka
UniProt
record

Taxonomy

UniRef

Cluster
aka
UniRef record

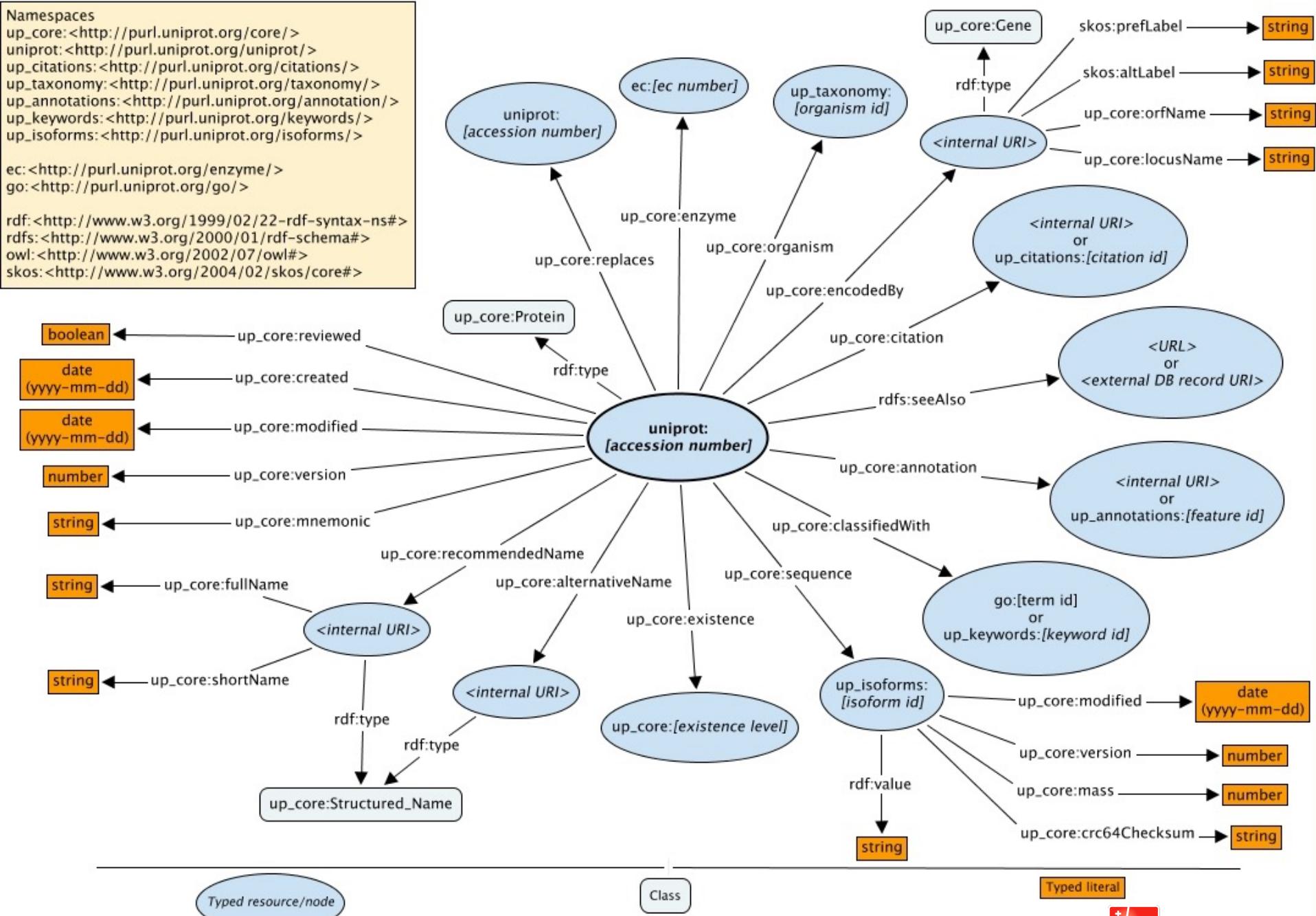
50%

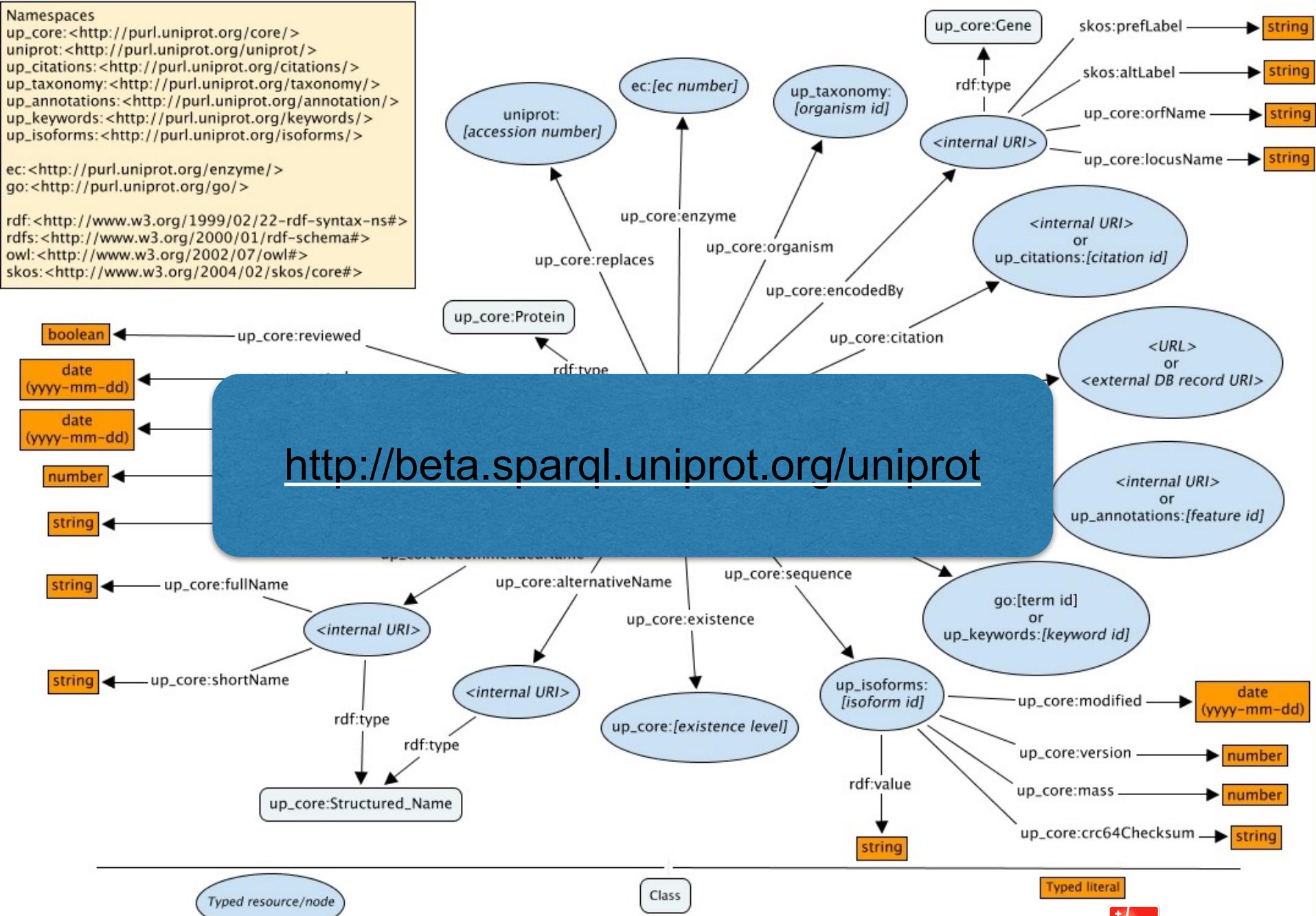
90%

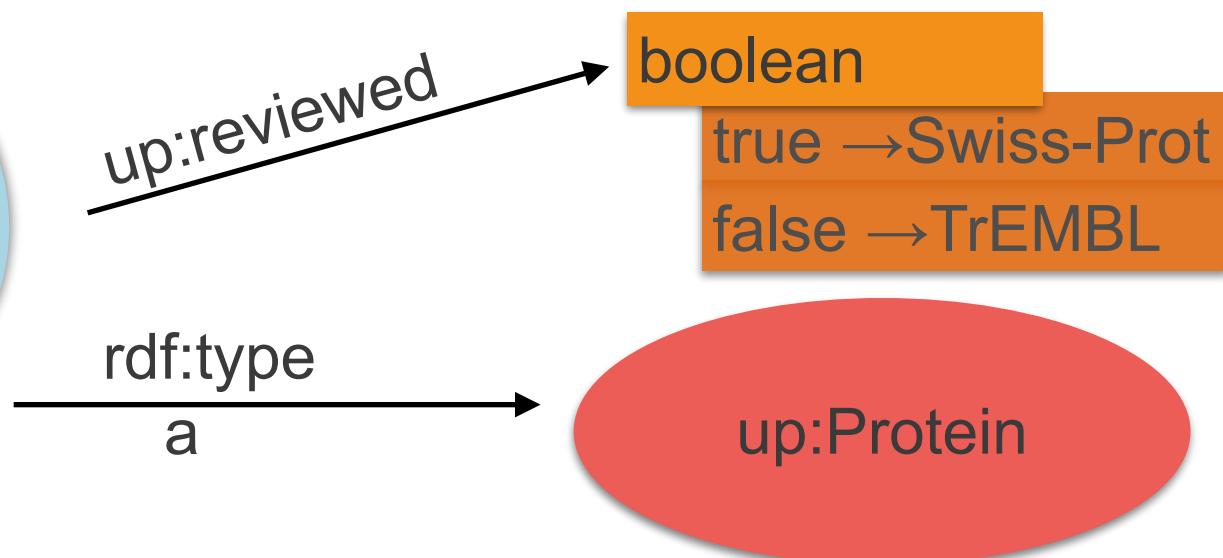
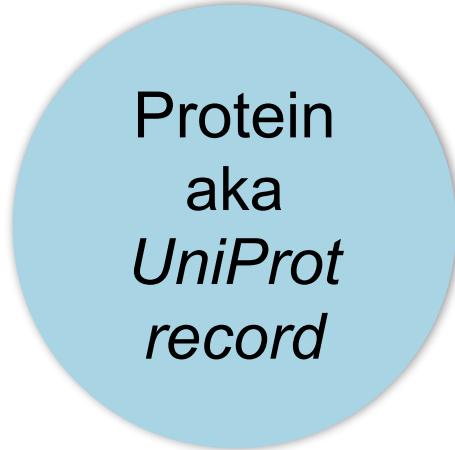
100%

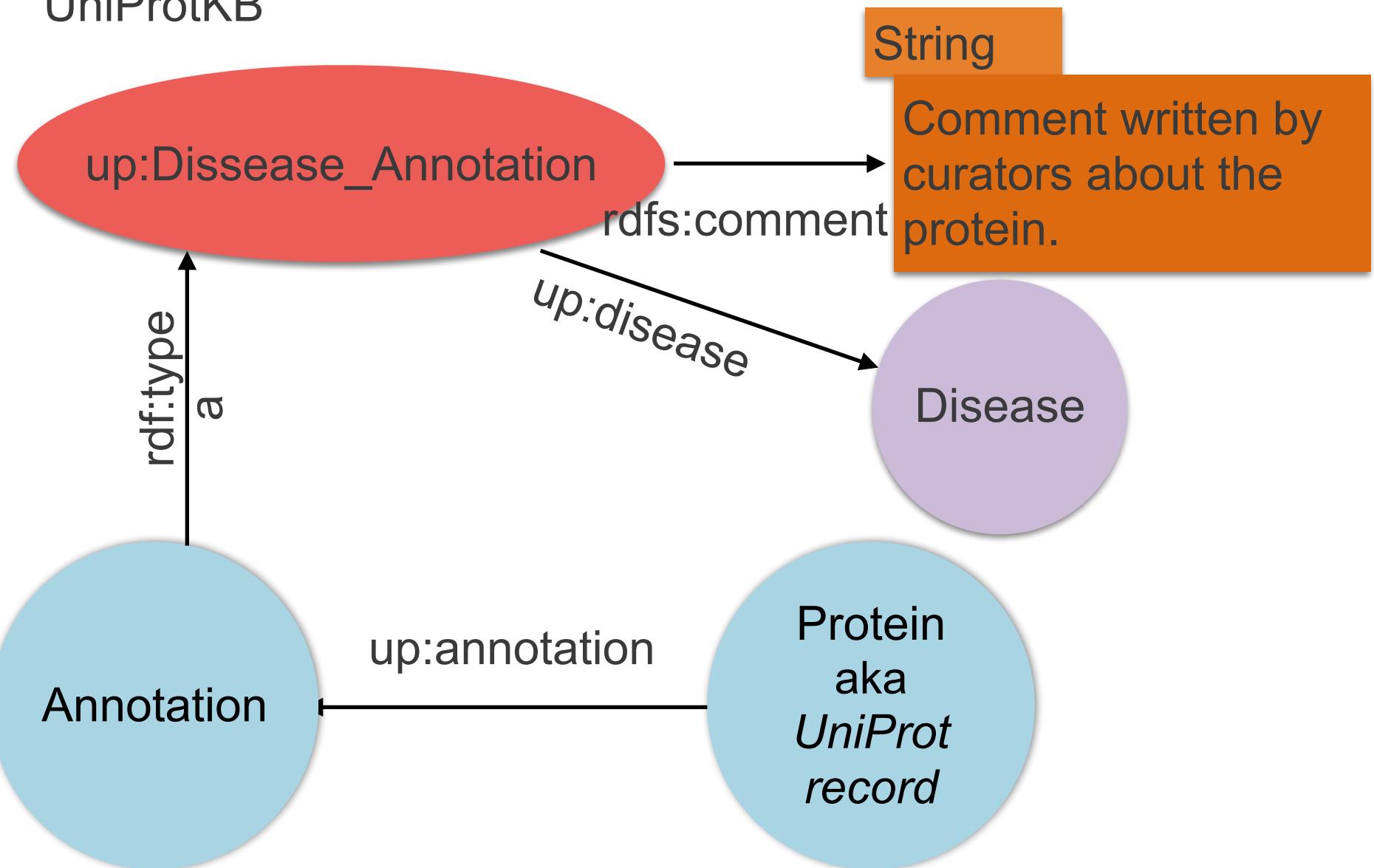
UniParc

Sequence
aka
UniParc









UniProtKB

Protein
aka
UniProt
record

Taxonomy

Locations

Keywords

Controlled vocabularies
to help you find information

ChEBI

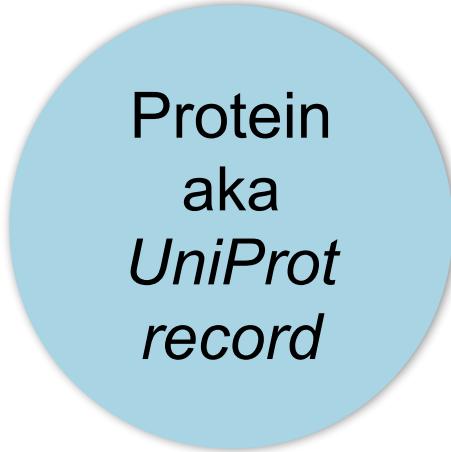
GO

FALDO

BIBO

Reusing ontologies
for ease of integration

UniProtKB



UniRef

Cluster
aka
UniRef record

50%

90%

100%

UniParc

Sequence
aka
UniParc

Taxonomy

UniProtKB

Protein
aka
UniProt record

Taxonomy

Taxon

UniRef

Cluster
aka
UniRef record

50%

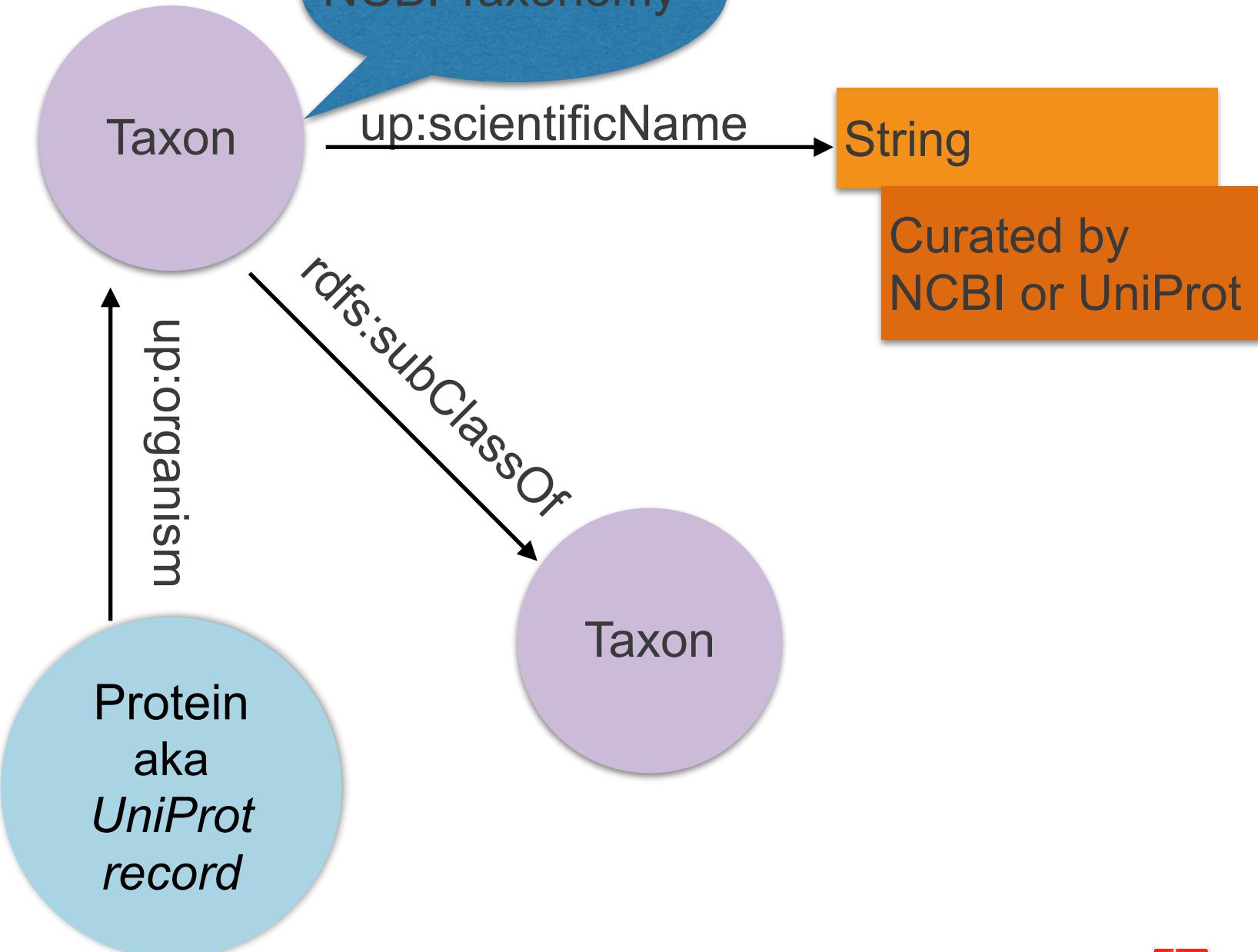
90%

100%

UniParc

Sequence
aka
UniParc

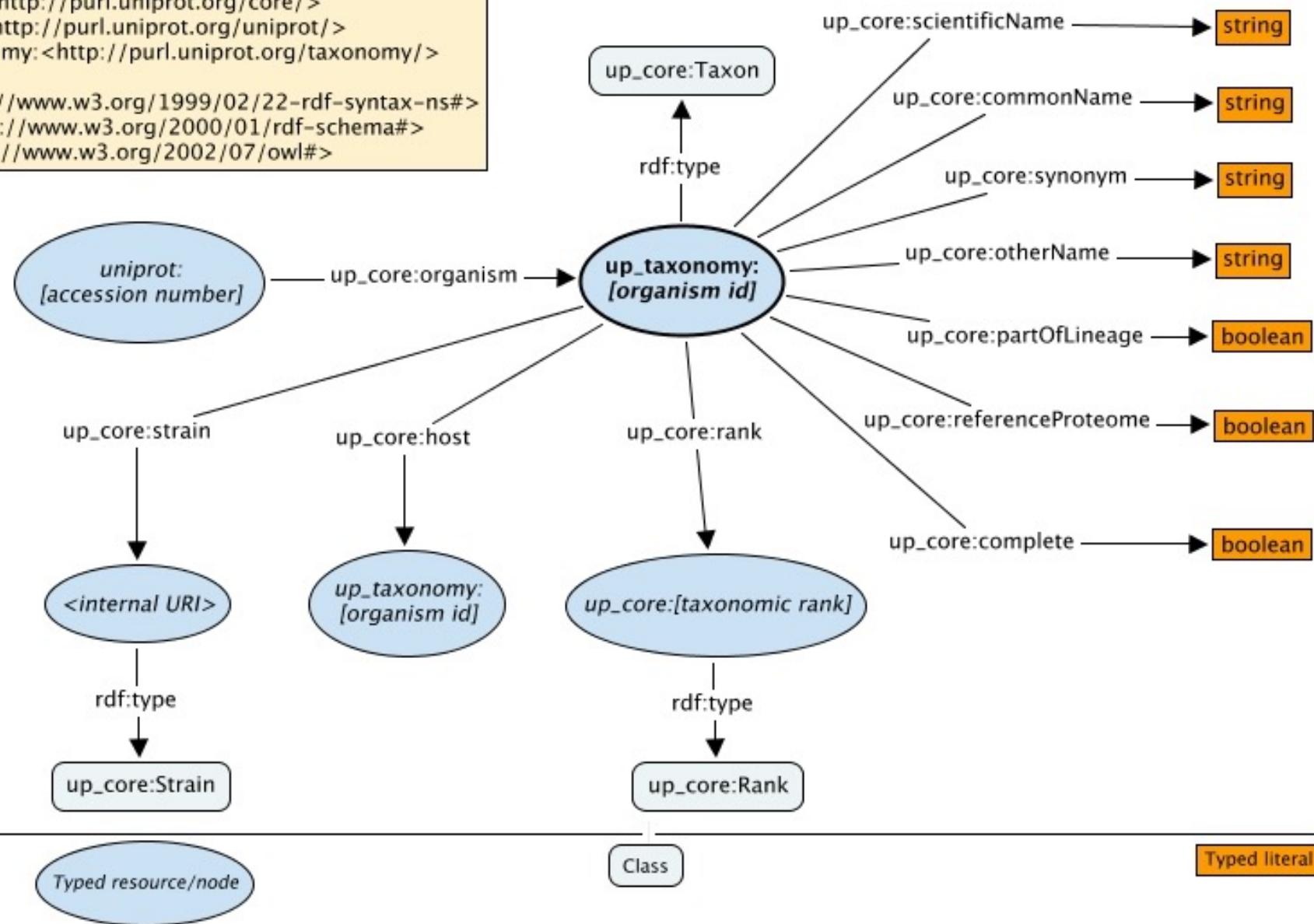
Taxonomy



Namespaces

up_core:<<http://purl.uniprot.org/core/>>
 uniprot:<<http://purl.uniprot.org/uniprot/>>
 up_taxonomy:<<http://purl.uniprot.org/taxonomy/>>

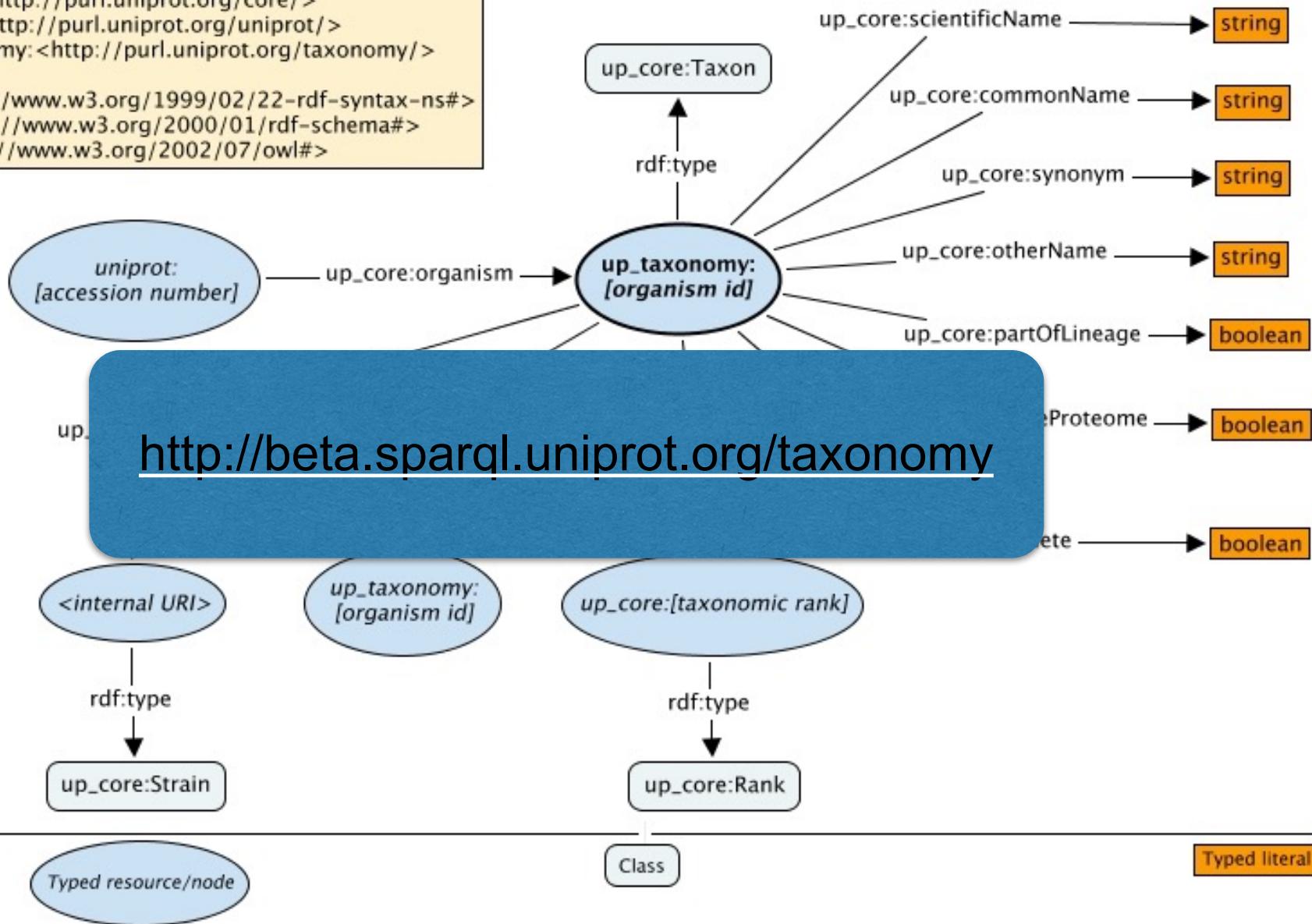
rdf:<<http://www.w3.org/1999/02/22-rdf-syntax-ns#>>
 rdfs:<<http://www.w3.org/2000/01/rdf-schema#>>
 owl:<<http://www.w3.org/2002/07/owl#>>



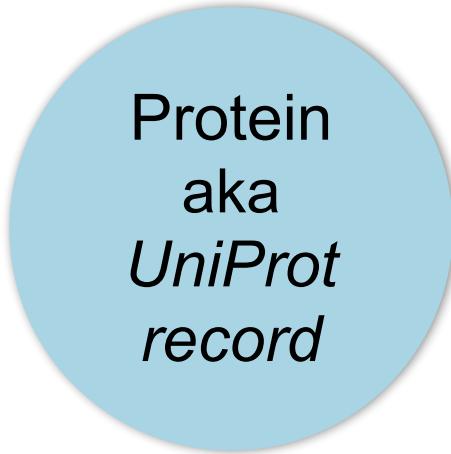
Namespaces

up_core:<<http://purl.uniprot.org/core/>>
uniprot:<<http://purl.uniprot.org/uniprot/>>
up_taxonomy:<<http://purl.uniprot.org/taxonomy/>>

rdf:<<http://www.w3.org/1999/02/22-rdf-syntax-ns#>>
rdfs:<<http://www.w3.org/2000/01/rdf-schema#>>
owl:<<http://www.w3.org/2002/07/owl#>>



UniProtKB



UniRef

Cluster
aka
UniRef record

50%

90%

100%

UniParc

Sequence
aka
UniParc

Taxonomy

UniProtKB

UniRef

50%

Cluster
aka
UniRef record

90%

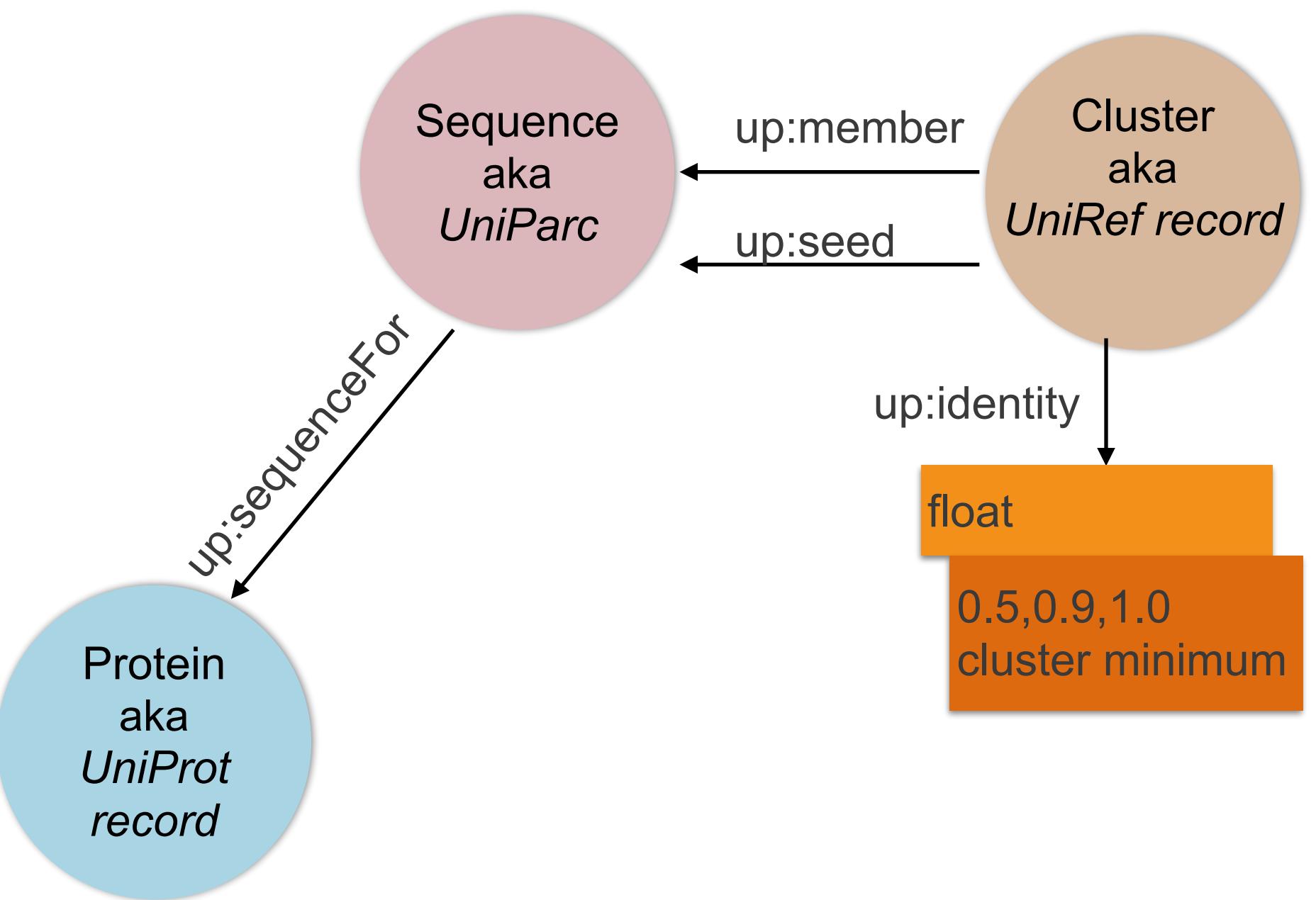
100%

Protein
aka
*UniProt
record*

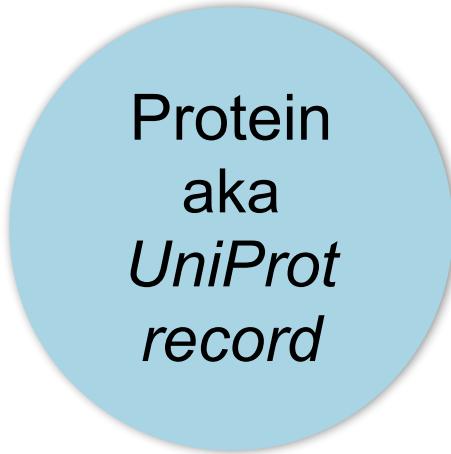
UniParc

Sequence
aka
UniParc

Taxonomy



UniProtKB



UniRef

Cluster
aka
UniRef record

50%

90%

100%

UniParc

Sequence
aka
UniParc

Taxonomy

UniProtKB

UniRef

Cluster
aka
UniRef record

50%

90%

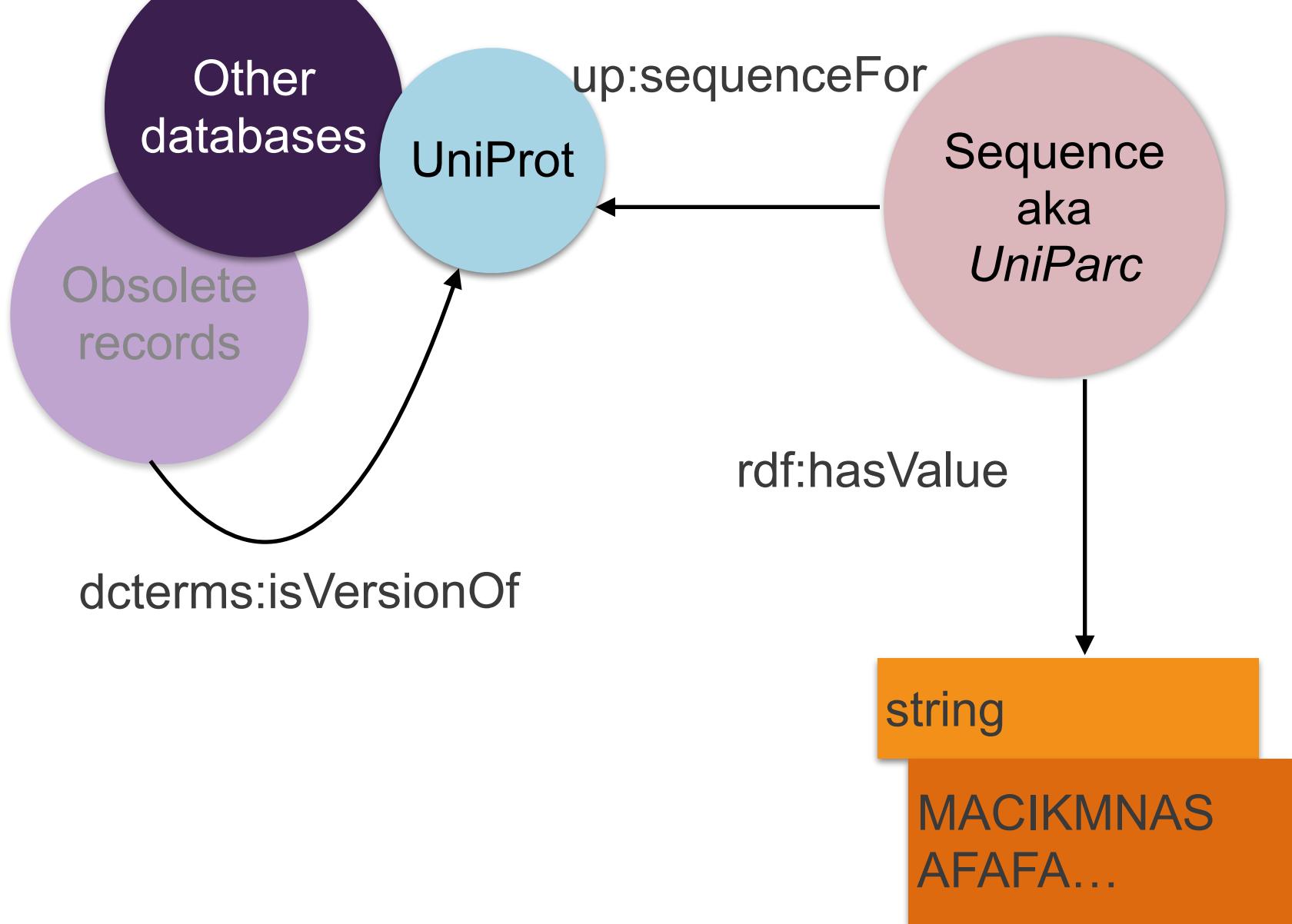
100%

Protein
aka
*UniProt
record*

UniParc

Sequence
aka
UniParc

Taxonomy



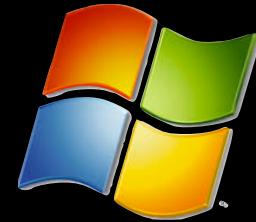
We are using EPO and EPO receptors as training data

- <http://www.uniprot.org/uniprot/P19235>
 - EPO_HUMAN
- <http://www.uniprot.org/uniprot/P01588>
 - EPOR_HUMAN
- <http://www.uniprot.org/uniparc/UPI000012A0AD>
- http://www.uniprot.org/uniref/UniRef100_P19235
- etc...

`./start-uniprot.sh`

or

`./start-uniprot.bat`



<http://localhost:9999>

<http://bc2-2015.sparql.uniprot.org/>

► SPARQL playground Documentation FAQs Data About

Resources ▾



6) Select Proteins similar to Human EPO isform 3

Show prefixes ...

endpoint: <http://localhost:9999/sparql>

```
SELECT ?epoLike
WHERE {?cluster up:identity "0.9"^^xsd:float;
          up:member ?member, ?member2 .
          ?member up:sequenceFor ?epoLike .
          ?member2 up:sequenceFor uniprot:P19235-3 .
          FILTER(?epoLike != uniprot:P19235-3)
}
```

html Go Reset

Query time is 0.188[s] for 9 rows

epoLike

```
<http://purl.uniprot.org/EPO/JA684580.1>
<http://purl.uniprot.org/USPTO/ADA04868.1>
```

Tags ▾ Filter sparql examples

1) Select all taxa from the UniProt taxonomy

2) Select all taxa, and their scientific name, from the UniProt taxonomy

3) All Swiss-Prot entries

4) Select UniProt records with a related disease

4) Select UniProt records with a related disease and a comment

4) Select UniProt records with a related disease plus more information about the disease

5) Proteins that are in the genus Hominidae

6) Select Proteins similar to Human EPO isform 3



File Edit View History Bookmarks Dots Help

http://www.google.com/search?hl=fr&q=0&sa=N&tbo=0&utb=0&usq=0&msa=0&osq=0

Google Error

We're sorry...

... but your query looks similar to automated requests from a computer virus or spyware application. To protect our users, we can't process your request right now.

We'll restore your access as quickly as possible, so try again soon. In the meantime, if you suspect that your computer or network has been infected, you might want to run a [virus checker](#) or [spyware remover](#) to make sure that your systems are free of viruses and other spurious software.

If you're continually receiving this error, you may be able to resolve the problem by deleting your Google cookies and revisiting Google. For browser-specific instructions, please consult your browser's online support center.

We apologize for the inconvenience, and hope we'll see you again on Google.

To continue searching, please type the characters you see below:



Date: [redacted] | Tor Disabled: [redacted] | IP: 69.239.181.129:43 | Proxy/Proxy-Disabled: [redacted] | Profile: [redacted] | NW: [redacted]