Enrique Solera Navarro

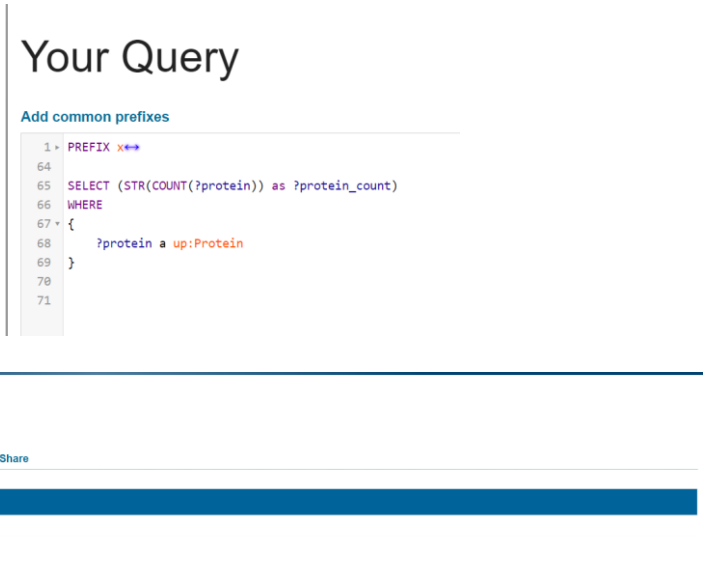
Assignment 5 SPARQL queries

I would like you to create the SPARQL query that will answer each of these questions. **Please submit the answers as screenshots of the query plus the first 2-3 lines of response. Submit to Moodle.** NO programming is required! Use whatever SPARQL client interface you want (Jupyter, Yasgui, etc.) Thanks!

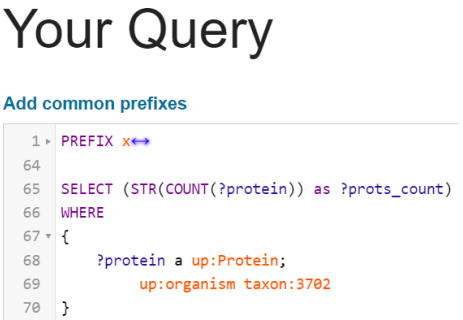
For many of these you will need to look-up how to use the SPARQL functions 'COUNT' and 'DISTINCT' (we used 'distinct' in class), and probably a few others...

UniProt SPARQL Endpoint: http://sparql.uniprot.org/sparql (note that you need to configure the endpoint to GET if you're using YASGUI)

Q1: 1 POINT How many protein records are in UniProt?



Q2: 1 POINT How many Arabidopsis thaliana protein records are in UniProt?



Enrique Solera Navarro

Results

Sparql XML Sparql JSON CSV Show query Share

prots_count
"272700"^^xsd:string

Q3: 1 POINT retrieve pictures of Arabidopsis thaliana from UniProt?

Your Query

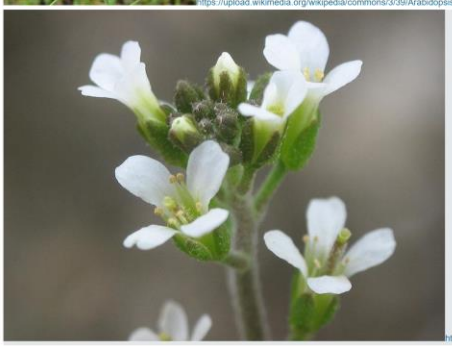
Add common prefixes

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT ?image
WHERE {
  taxon:3702 foaf:depiction ?image
}
```

Results

Sparql XML Sparql JSON CSV Show query Share

image
 https://upload.wikimedia.org/wikipedia/commons/3/39/Arabidopsis.jpg
 https://upload.wikimedia.org/wikipedia/commons/thumb/6/60/Arabidopsis_thaliana_inflorescencias.jpg/800px-Arabidopsis_thaliana_inflorescencias.jpg

Q4: 1 POINT: What is the description of the enzyme activity of UniProt Protein Q9SZZ8

Your Query

Add common prefixes

```
PREFIX uniprot:<http://purl.uniprot.org/uniprot/>
PREFIX up:<http://purl.uniprot.org/core/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>

SELECT ?activity_label
WHERE
{
    uniprot:Q9SZZ8 a up:Protein ;
             up:enzyme ?enzyme.
    ?enzyme up:activity ?activity.
        ?activity rdfs:label ?activity_label
}
```

Enrique Solera Navarro

Results

Sparql XML Sparql JSON CSV Show query Share

activity_label
"all-trans-beta-carotene + 4 H(+) + 2 O2 + 4 reduced [2Fe-2S]-[ferredoxin] = all-trans-zeaxanthin + 2 H2O + 4 oxidized [2Fe-2S]-[ferredoxin]."xsd:string

Q5: 1 POINT: Retrieve the proteins ids, and date of submission, for 5 proteins that have been added to UniProt this year (HINT Google for “SPARQL FILTER by date”)

Your Query

Add common prefixes

```
PREFIX up:<http://purl.uniprot.org/core/>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>

SELECT ?id (STR(?date) AS ?date_str)
WHERE
{
    ?protein a up:Protein ;
            up:created ?date .
    FILTER ( ?date >= "2023-01-01"^^xsd:date)

    BIND (SUBSTR(STR(?protein),33) AS?id)

}
```

Results

Sparql XML Sparql JSON CSV Show query Share

id	date_str
"C4M4P4"xsd:string	"2023-02-22"xsd:string
"C4M4P4"xsd:string	"2023-02-22"xsd:string
"P83003"xsd:string	"2023-02-22"xsd:string
"Q04770"xsd:string	"2023-02-22"xsd:string
"P49009"xsd:string	"2023-02-22"xsd:string

Q6: 1 POINT How many species are in the UniProt taxonomy?

Your Query

Add common prefixes

```
PREFIX up:<http://purl.uniprot.org/core/>

SELECT (STR(COUNT(DISTINCT ?taxon)) AS ?count)
FROM <http://sparql.uniprot.org/taxonomy>
WHERE
{
    ?taxon a up:Taxon ;
           up:rank up:Species
}
```

Results

Sparql XML Sparql JSON CSV Show query Share

count
"2058509"xsd:string

Enrique Solera Navarro

Q7: 2 POINT How many species have at least one protein record? (this might take a long time to execute, so do this one last!)

Your Query

Add common prefixes

```
PREFIX up:<http://purl.uniprot.org/core/>

SELECT (STR(COUNT(DISTINCT ?taxon)) AS ?count)
WHERE
{
    ?protein a up:Protein ;
           up:organism ?taxon .
    ?taxon up:rank up:Species
}
```

Results

Sparql XML Sparql JSON CSV Show query Share

count
"1090630"xsd:string

<u>Q8: 3 points</u>: find the AGI codes and gene names for all Arabidopsis thaliana proteins that have a protein function annotation description that mentions “pattern formation”

```
PREFIX up:<http://purl.uniprot.org/core/>
PREFIX taxon:<http://purl.uniprot.org/taxonomy/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos:<http://www.w3.org/2004/02/skos/core#>

SELECT ?AGI ?name
WHERE
{
    ?protein a up:Protein ;
             up:organism taxon:3702 ;
             up:encodedBy ?gene ;
             up:annotation  ?annot .
    ?gene up:locusName ?AGI ;
          skos:prefLabel ?name .
    ?annot a up:Function_Annotation ;
           rdfs:comment ?annotComment .

    FILTER CONTAINS(?annotComment, 'pattern formation')
}
```

Results

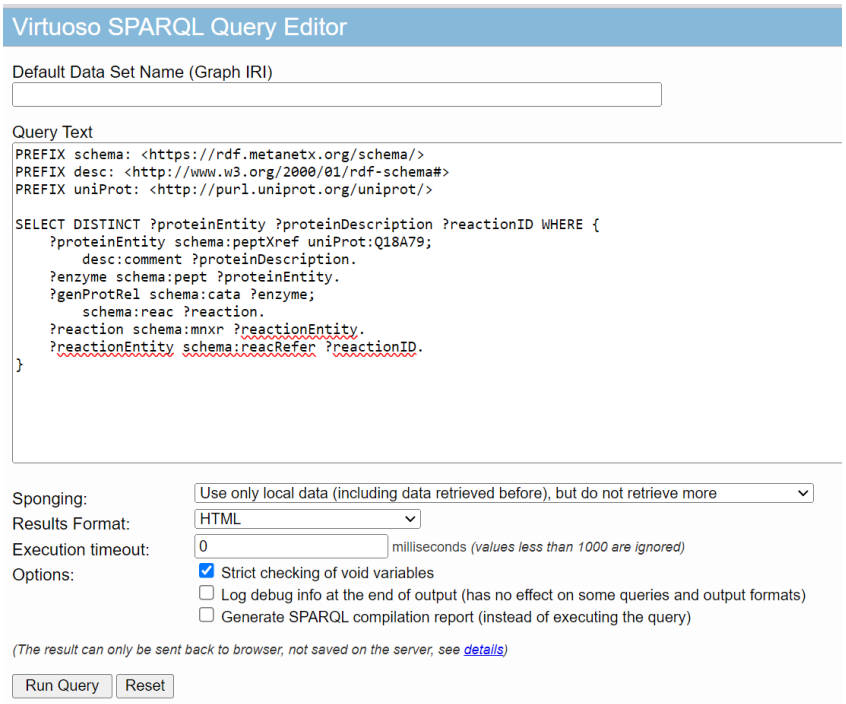
Sparql XML Sparql JSON CSV Show query Share

AGI	name
"At1g66470"xsd:string	"RHD6"xsd:string
"At1g66470"xsd:string	"RHD6"xsd:string
"At1g13980"xsd:string	"GN"xsd:string

Enrique Solera Navarro

From the MetaNetX metabolic networks for metagenomics database
SPARQL Endpoint: https://rdf.metanetx.org/sparql
(this slide deck will make it much easier for you!
https://www.metanetx.org/cgi-bin/mnxget/mnxref/MetaNetX_RDF_schema.pdf)

Q9: 4 POINTS: what is the MetaNetX Reaction identifier (starts with “mnxr”) for the UniProt Protein uniprotkb:Q18A79



proteinEntity	proteinDescription	reactionID
https://rdf.metanetx.org/pept/GLGA_CLOD6	"Glycogen synthase (EC 2.4.1.21) (Starch [bacterial glycogen] synthase)"	http://bigg.ucsd.edu/universal/reactions/GLCS1
https://rdf.metanetx.org/pept/GLGA_CLOD6	"Glycogen synthase (EC 2.4.1.21) (Starch [bacterial glycogen] synthase)"	http://rdf.rhea-db.org/18189

FEDERATED QUERY - UniProt and MetaNetX

Q10: 5 POINTS: What is the official locus name, and the MetaNetX Reaction identifier (mnxr…..) for the protein that has “glycine reductase” catalytic activity in Clostridium difficile (taxon 272563). (this must be executed on the https://rdf.metanetx.org/sparql endpoint)

Enrique Solera Navarro

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

Query Text

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX mnx: <https://rdf.metanetx.org/schema/>
PREFIX taxon: <http://purl.uniprot.org/taxonomy/>

SELECT DISTINCT ?locusName ?reactionID WHERE {
    SERVICE <https://sparql.uniprot.org/sparql>{
        SELECT ?UPquery ?locusName WHERE {
            ?UPquery a up:Protein;
                up:organism taxon:272563;
                up:encodedBy ?gene.
            ?gene up:locusName ?locusName.
            ?UPquery up:enzyme ?act.
            ?act skos:prefLabel 'glycine reductase'.
        }
    }
    ?pept mnx:peptXref ?UPquery.
    ?cata mnx:pept ?pept.
    ?gpr mnx:cata ?cata;
        mnx:reac ?reac.
    ?reac mnx:mnxr ?rxn.
    ?rxn mnx:reacRefer ?reactionID.
}
```

Sponging: Use only local data (including data retrieved before), but do not retrieve more

Results Format: HTML

Execution timeout: 0 milliseconds *(values less than 1000 are ignored)*

Options:
- [x] Strict checking of void variables
- [ ] Log debug info at the end of output (has no effect on some queries and output formats)
- [ ] Generate SPARQL compilation report (instead of executing the query)

(The result can only be sent back to browser, not saved on the server, see details)

Run Query | Reset

locusName	reactionID
"CD630_23490"	http://bigg.ucsd.edu/universal/reactions/SR5
"CD630_23520"	http://bigg.ucsd.edu/universal/reactions/SR5
"CD630_23510"	http://bigg.ucsd.edu/universal/reactions/SR5
"CD630_23480"	http://bigg.ucsd.edu/universal/reactions/SR5
"CD630_23490"	http://bigg.ucsd.edu/universal/reactions/GLYR
"CD630_23520"	http://bigg.ucsd.edu/universal/reactions/GLYR
"CD630_23510"	http://bigg.ucsd.edu/universal/reactions/GLYR
"CD630_23540"	http://bigg.ucsd.edu/universal/reactions/GLYR
"CD630_23480"	http://bigg.ucsd.edu/universal/reactions/GLYR