Assignment 5 SPARQL queries

I would like you to create the SPARQL query that will answer each of these questions. Please submit the queries simply as a text document (NO programming is required!) - submit to GitHub as usual.

*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/>**

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

PREFIX uniprotkb:<http://purl.uniprot.org/uniprot/>

PREFIX taxon:<http://purl.uniprot.org/taxonomy/>

PREFIX ec:<http://purl.uniprot.org/enzyme/>

PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>

**1 POINT How many protein records are in UniProt?**

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

select (STR(COUNT(?s)) as ?count)

where {

?s rdf:type up:Protein

}

Count = **281303435** proteins

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

SELECT ?taxon

WHERE

{ ?taxon a up:Taxon ;

up:scientificName "Arabidopsis thaliana" .

}

taxon = <http://purl.uniprot.org/taxonomy/3702>

SELECT distinct ?p

WHERE

{ ?s ?p taxon:3702

}

limit 10

p = 3. <http://purl.uniprot.org/core/organism>

select (STR(count(?s)) as ?Count)

where {

?s rdf:type up:Protein ;

up:organism taxon:3702 .

}

Count = **89182** Proteins

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

select distinct ?p

where {

uniprotkb:Q9SZZ8 ?p ?o

}

limit 100

p = 8. [http://purl.uniprot.org/core/enzyme](http://purl.uniprot.org/core/enzyme" \t "_blank)

select distinct ?pre

where {

uniprotkb:Q9SZZ8 up:enzyme ?o .

?o ?pre ?l

}

limit 100

pre = 3. <http://purl.uniprot.org/core/activity>

select ?desc

where {

uniprotkb:Q9SZZ8 up:enzyme ?o .

?o up:activity ?l .

?l rdfs:label ?desc .

}

desc = **Beta-carotene + 4 reduced ferredoxin [iron-sulfur] cluster + 2 H(+) + 2 O(2) = zeaxanthin + 4 oxidized ferredoxin [iron-sulfur] cluster + 2 H(2)O.**

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

select distinct ?p

where {

?prot a up:Protein ;

?p ?o .

}

p = 6. <http://purl.uniprot.org/core/created>

select ?p ?date

where {

?p a up:Protein ;

up:created ?date .

FILTER (?date > "2019-01-01"^^xsd:dateTime)

}

**1 POINT How many species are in the UniProt taxonomy?**

select distinct ?p

where {

?tax a up:Taxon ;

?p ?o

}

limit 30

p = 9. <http://purl.uniprot.org/core/rank>

select distinct ?o

where {

?tax a up:Taxon ;

up:rank ?o

}

limit 30

o = 1. http://purl.uniprot.org/core/Species

select (str(count(distinct ?tax)) as ?Count)

where {

?tax a up:Taxon ;

up:rank up:Species

}

Count = **1766921** Species

**1 POINT How many species have at least one protein record?**

select (str(count(distinct ?t)) as ?Count)

where {

?s rdf:type up:Protein ;

up:organism ?t .

?t up:rank up:Species .

}

Count = **984622**

**From the Atlas gene expression database SPARQL Endpoint: http://www.ebi.ac.uk/rdf/services/atlas/sparql**

**1 POINT What is the Affymetrix probe ID for the Arabiodopsis Apetala3 gene? (HINT - you cannot answer this directly from Atlas - you will first have to look at what kinds of database cross-references are in Atlas, and then construct the appropriate URI for the Apetala3 gene based on its ID number in \*that\* database)**

**3 POINTS - get the experimental description for all experiments where the Arabidopsis Apetala3 gene is DOWN regulated**

**From the REACTOME database SPARQL endpoint: http://www.ebi.ac.uk/rdf/services/reactome/sparql**

**2 POINTS: How many REACTOME pathways are assigned to Arabidopsis (taxon 3702)? (note that REACTOME uses different URLs to define their taxonomy compared to UniProt, so you will first have to learn how to structure those URLs….)**

PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX biopax3: <http://www.biopax.org/release/biopax-level3.owl#>

select (count (distinct ?pathway) AS ?Count)

where

{ ?pathway rdf:type biopax3:Pathway ;

biopax3:organism <http://identifiers.org/taxonomy/3702>

}

Count = 809 Pathways

**3 POINTS: get all PubMed references for the pathway with the name “Degradation of the extracellular matrix”**

PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX biopax3: <http://www.biopax.org/release/biopax-level3.owl#>

SELECT DISTINCT ?pathway ?pathwayname ?s

WHERE

{

?pathway rdf:type biopax3:Pathway ;

biopax3:displayName ?pathwayname .

FILTER(str(?pathwayname) = 'Degradation of the extracellular matrix') .

?s biopax3:db ?o .

FILTER(str(?o) ='Pubmed') .

}

# BONUS QUERIES

**UniProt BONUS 2 points: 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 keywords:<http://purl.uniprot.org/keywords/>

PREFIX uniprotkb:<http://purl.uniprot.org/uniprot/>

PREFIX taxon:<http://purl.uniprot.org/taxonomy/>

PREFIX ec:<http://purl.uniprot.org/enzyme/>

PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>

PREFIX skos:<http://www.w3.org/2004/02/skos/core#>

PREFIX owl:<http://www.w3.org/2002/07/owl#>

PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>

PREFIX GO:<http://purl.obolibrary.org/obo/GO\_>

select ?f ?nombre

where{

?protein a up:Protein ;

up:organism taxon:3702 ;

up:annotation ?a ;

up:encodedBy ?locus .

?locus up:locusName ?f ;

skos:prefLabel ?nombre .

?a a up:Function\_Annotation ;

rdfs:comment ?s

FILTER contains(?s, 'pattern formation') .

}

**REACTOME BONUS 2 points: write a query that proves that all Arabidopsis pathway annotations in Reactome are “inferred from electronic annotation” (evidence code) (...and therefore are probably garbage!!!)**

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX biopax3: <http://www.biopax.org/release/biopax-level3.owl#>

PREFIX taxon: <http://identifiers.org/taxonomy/>

SELECT (count(?pathway1) as ?Count1) (count(?term) as ?count2)

WHERE

{

?pathway1 rdf:type biopax3:Pathway ;

biopax3:organism taxon:3702.

?pathway2 rdf:type biopax3:Pathway ;

biopax3:organism taxon:3702;

biopax3:evidence ?evidence .

?evidence biopax3:evidenceCode ?evidenceCode .

?evidenceCode biopax3:term ?term .

}

Count1= 654481

Count2= 654481