Assignment 5 SPARQL queries

I would like you to create the SPARQL query that will answer each of these questions. Please submit the queries as a Jupyter notebook with the SPARQL kernel activated. NO programming is required! Submit to GitHub as usual, WITH THE ANSWERS STILL VISIBLE IN THE NOTEBOOK. 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**

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

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

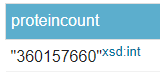
SELECT (COUNT (DISTINCT ?protein) AS ?proteincount)

WHERE

{

?protein a up:Protein .

}



**RESULT:**

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

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

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

SELECT (COUNT(DISTINCT ?protein) AS ?proteincount)

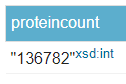
WHERE

{

?protein a up:Protein .

?protein up:organism taxon:3702 .

}



**RESULT:**

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

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

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

SELECT ?pictures

WHERE

{

?taxon a up:Taxon.

?taxon up:scientificName "Arabidopsis thaliana".

?taxon foaf:depiction ?pictures .

}

**RESULT:**

****

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

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

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

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

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

SELECT ?name ?activity

WHERE

{

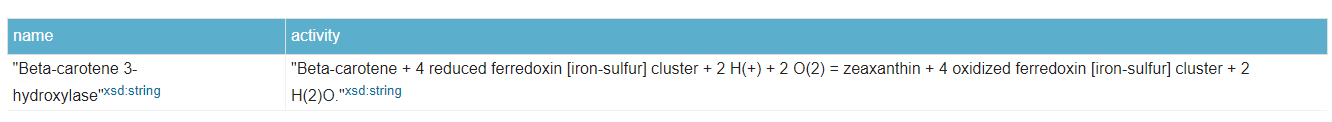
uniprotkb:Q9SZZ8 up:enzyme ?enzyme .

?enzyme skos:prefLabel ?name .

?enzyme up:activity ?act .

?act rdfs:label ?activity.

}

**RESULTS:** 

**Q5: 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”)**

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

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

SELECT ?id ?date

WHERE

{

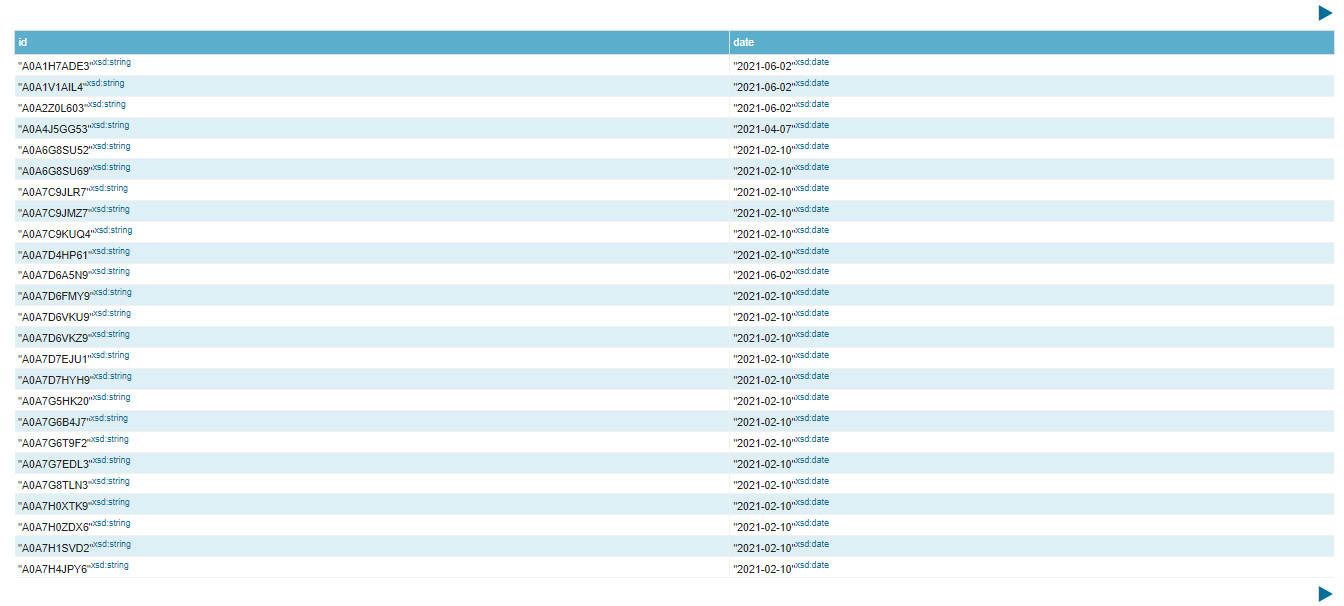
?protein a up:Protein .

?protein up:created ?date .

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

FILTER (?date >= '2021-01-01'^^xsd:date)

}

**RESULT:** 

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

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

SELECT (COUNT (DISTINCT ?species) AS ?speciesnum)

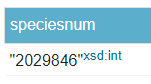
WHERE

{

?species a up:Taxon .

?species up:rank up:Species .

}



**RESULT:**

**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!)**

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

SELECT (COUNT(DISTINCT ?num) AS ?species\_atl\_oneprot)

WHERE

{

?protein a up:Protein .

?protein up:organism ?num .

?num a up:Taxon .

?num up:rank up:Species .

}



**RESULT:**

**Q8: 3 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 rdfs:<http://www.w3.org/2000/01/rdf-schema#>

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

SELECT DISTINCT ?AGI\_name ?name

WHERE

{

?protein a up:Protein ;

up:organism ?taxon\_id ;

up:encodedBy ?g ;

up:annotation ?function\_annot .

?taxon\_id a up:Taxon ;

up:scientificName "Arabidopsis thaliana" .

?g skos:prefLabel ?name .

?g up:locusName ?AGI\_name .

?protein up:annotation ?annotation .

?annotation rdfs:comment ?f\_annot .

FILTER REGEX (?f\_annot, "pattern formation", "i") .

}

**RESULT: **

**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**](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**

PREFIX mtnx: <https://rdf.metanetx.org/schema/>

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

SELECT DISTINCT ?id

WHERE{

?pept mtnx:peptXref up:Q18A79 .

?cata a mtnx:CATA ;

mtnx:pept ?pept .

?gpr mtnx:cata ?cata ;

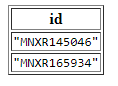
mtnx:reac ?reac .

?reac a mtnx:REAC ;

mtnx:mnxr ?mnxr .

?mnxr rdfs:label ?id .

}

****

**RESULT:**

**FEDERATED QUERY - UniProt and MetaNetX**

**Q10: 5 POINTS: What is the official Gene ID (UniProt calls this a “mnemonic”) and the MetaNetX Reaction identifier (mnxr…..) for the protein that has “Starch synthase” catalytic activity in Clostridium difficile (taxon 272563).**

PREFIX mnx: <https://rdf.metanetx.org/schema/>

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

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

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

SELECT DISTINCT ?ID ?MNXID ?activity

WHERE{

service <http://sparql.uniprot.org/sparql> {

?protein a up:Protein ;

up:organism taxon:272563 ;

up:mnemonic ?ID ;

up:classifiedWith ?GO .

?GO rdfs:label ?activity .

filter contains(?activity, "starch synthase")

bind (substr(str(?protein),33) as ?prot\_ac)

bind (IRI(CONCAT(uniprotkb:,?prot\_ac)) as ?uniprotRef)

}

service <https://rdf.metanetx.org/sparql> {

?pept mnx:peptXref ?uniprotRef .

?cata mnx:pept ?pept .

?gpr mnx:cata ?cata ;

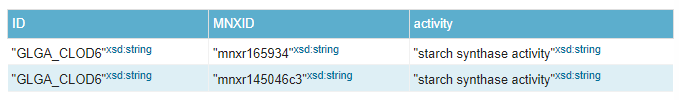
mnx:reac ?reac .

?reac rdfs:label ?MNXID .

}

}

**RESULT:**

****

**Antonio Porlán Miñarro Bioinformatics Programming Challeges**

**Assignment 5**