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/**

**1 POINT** How many protein records are in UniProt? – **281303435**

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

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

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

SELECT (STR(COUNT(?protein)) AS ?number)

WHERE {

?protein a up:Protein

}

**1 POINT** How many Arabidopsis thaliana protein records are in UniProt? **– 89182**

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

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

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

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

SELECT (STR(COUNT(DISTINCT ?protein)) AS ?ArabidopsisProtein)

WHERE {

?protein up:organism taxon:3702 .

?protein a up:Protein

}

**1 POINT**: What is the description of the enzyme activity of UniProt Protein Q9SZZ8 - **Beta-carotene + 4 reduced ferredoxin [iron-sulfur] cluster + 2 H(+) + 2 O(2) = zeaxanthin + 4 oxidized ferredoxin [iron-sulfur] cluster + 2 H(2)O.**

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

PREFIX rdfs: [http://www.w3.org/2000/01/rdf-schema#](http://www.w3.org/2000/01/rdf-schema)

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

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

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

SELECT ?description

WHERE{

uniprotkb:Q9SZZ8 up:enzyme ?enzyme.

?enzyme up:activity ?activity.

?activity rdfs:label ?description

}

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

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

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

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

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

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

SELECT ?protein ?date

WHERE {

?protein up:created ?date .

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

?protein a up:Protein .

}limit 10

**1 POINT** How many species are in the UniProt taxonomy? **- 1766921**

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

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

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

SELECT (STR(COUNT(DISTINCT ?taxon)) as ?taxon\_count)

WHERE{

?taxon a up:Taxon.

?taxon up:rank up:Species

}

**1 POINT** How many species have at least one protein record? - **984622**

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

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

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

SELECT (STR(COUNT (DISTINCT ?species)) AS ?number)

WHERE {

SELECT ?species

WHERE {

?species up:rank up:Species .

?protein a up:Protein .

?protein up:organism ?species

}

GROUP BY ?species

HAVING (COUNT (?protein) >= 1)

}

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

**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….) - **809**

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

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

SELECT (STR(COUNT(DISTINCT ?pathway)) AS ?number)

WHERE

{

?pathway a biopax3:Pathway .

?pathway biopax3:organism taxonomy:3702

}

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

# 

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

SELECT DISTINCT ?refs

WHERE {

?pathway a biopax3:Pathway .

?pathway biopax3:displayName ?name .

?pathway biopax3:evidence ?evidence .

?evidence biopax3:xref ?refs .

FILTER (CONTAINS(?name, "Degradation of the extracellular matrix"))

}

# 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 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#>

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

SELECT ?AGIcode ?genename

WHERE

{

?protein a up:Protein .

?protein up:organism taxon:3702 .

?protein up:encodedBy ?gene .

?gene up:locusName ?AGIcode .

?gene skos:prefLabel ?genename .

?protein up:annotation ?annotation .

?annotation a up:Function\_Annotation.

?annotation rdfs:comment ?annotationComment.

FILTER CONTAINS(?annotationComment, '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!!!) – 654481 & 654481

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

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

SELECT (STR(COUNT(?path\_all)) AS ?all) (STR(COUNT(?path\_e)) AS ?electronic\_records)

WHERE

{

?path\_all a biopax3:Pathway .

?path\_all biopax3:organism taxon:3702 .

?path\_all biopax3:evidence ?evidence\_all .

?evidence\_all biopax3:evidenceCode ?evidenceCode\_all .

?evidenceCode\_all biopax3:term ?term\_all .

?path\_e a biopax3:Pathway .

?path\_e biopax3:organism taxon:3702 .

?path\_e biopax3:evidence ?evidence\_e .

?evidence\_e biopax3:evidenceCode ?evidenceCode\_e .

?evidenceCode\_e biopax3:term ?term\_e .

FILTER REGEX(?term\_e, 'inferred from electronic annotation')

}