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?

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

**PREFIX taxon:<https://www.uniprot.org/taxonomy/>**

**SELECT (STR(COUNT(?protein)) as ?prots\_count)**

**WHERE**

**{**

**?protein a up:Protein .**

**}**

***Result: 281.303.435***

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

**PREFIX taxon: <**[**http://purl.uniprot.org/taxonomy/**](http://purl.uniprot.org/taxonomy/)**>**

**PREFIX up:<**[**http://purl.uniprot.org/core/**](http://purl.uniprot.org/core/)**>**

**SELECT (COUNT(DISTINCT?protein) as ?proteins)**

**WHERE**

**{**

**?protein up:organism taxon:3702 .**

**?protein a up:Protein .**

**}**

***Result:89.182***

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

**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**

**}**

***Result: 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”)

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

**SELECT ?protein ?id ?date**

**WHERE**

**{**

**?protein a up:Protein .**

**?protein up:mnemonic ?id .**

**?protein up:created ?date .**

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

**}**

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

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

**SELECT (STR(COUNT(DISTINCT ?taxon)) AS ?number\_species)**

**FROM <http://sparql.uniprot.org/taxonomy>**

**WHERE**

**{**

**?taxon a up:Taxon .**

**?taxon up:rank up:Species .**

**}**

***Result: 1.766.921***

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

**SELECT (STR(COUNT(DISTINCT ?taxon)) AS ?number\_species)**

**FROM <http://sparql.uniprot.org/taxonomy>**

**WHERE**

**{**

**?protein a up:Protein .**

**?protein up:organism ?taxon .**

**?taxon a up:Taxon .**

**?taxon up:rank up:Species .**

**}**

***Result: 984.622***

**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

**PREFIX dcterms: <http://purl.org/dc/terms/>**

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

**PREFIX atlasterms: <http://rdf.ebi.ac.uk/terms/expressionatlas/>**

**PREFIX atlas: <http://rdf.ebi.ac.uk/resource/expressionatlas/>**

**SELECT distinct ?diffValue ?expUri ?description**

**FROM <http://rdf.ebi.ac.uk/dataset/expressionatlas>**

**WHERE {**

**?expUri atlasterms:hasPart ?analysis .**

**?expUri dcterms:description ?description .**

**?analysis atlasterms:hasOutput ?value .**

**?value rdfs:label ?diffValue .**

**FILTER REGEX(?diffValue, 'DOWN')**

**?value atlasterms:refersTo <http://rdf.ebi.ac.uk/resource/ensembl/AT3G54340> .**

**}**

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

**SELECT (COUNT(DISTINCT?pathway) as ?number\_path)**

**WHERE**

**{**

**?pathway a biopax3:Pathway .**

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

**}**

***Result: 809***

**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 (str(?pmId) AS ?pubmed\_ID)**

**WHERE**

**{**

**?pathway a biopax3:Pathway ;**

**biopax3:displayName ?name ;**

**biopax3:xref ?reference .**

**?reference biopax3:db ?db .**

**?reference biopax3:id ?pmId .**

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

**FILTER(str(?db) ='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 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 ?name ?id**

**WHERE**

**{**

**?protein a up:Protein .**

**?protein up:organism taxon:3702 .**

**?protein up:annotation ?annotation .**

**?annotation a up:Function\_Annotation .**

**?protein up:encodedBy ?gene .**

**?gene up:locusName ?id .**

**?gene skos:prefLabel ?name .**

**?annotation rdfs:comment ?text .**

**FILTER (CONTAINS(?text, 'pattern formation'))**

**}**

***Result: 15 entries***

**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 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 (?pathway) AS ?inferred)**

**WHERE**

**{**

**?pathway a biopax3:Pathway ;**

**biopax3:organism taxon:3702 ;**

**biopax3:evidence ?evidence .**

**?evidence biopax3:evidenceCode ?evidenceCode.**

**?evidenceCode biopax3:term ?term .**

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

**}**

***Result: 809***