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 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 bibo:<http://purl.org/ontology/bibo/>

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

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

PREFIX faldo:<http://biohackathon.org/resource/faldo#>

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

PREFIX allie:<http://allie.dbcls.jp/>

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

PREFIX chebihash:<http://purl.obolibrary.org/obo/chebi#>

PREFIX cco:<http://rdf.ebi.ac.uk/terms/chembl#>

PREFIX codoa:<http://purl.glycoinfo.org/ontology/codao#>

PREFIX ensembl:<http://rdf.ebi.ac.uk/resource/ensembl/>

PREFIX ensemblexon:<http://rdf.ebi.ac.uk/resource/ensembl.exon/>

PREFIX ensemblprotein:<http://rdf.ebi.ac.uk/resource/ensembl.protein/>

PREFIX ensemblterms:<http://rdf.ebi.ac.uk/terms/ensembl/>

PREFIX ensembltranscript:<http://rdf.ebi.ac.uk/resource/ensembl.transcript/>

PREFIX glycan:<http://purl.jp/bio/12/glyco/glycan#>

PREFIX glyconnect:<https://purl.org/glyconnect/>

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

PREFIX mesh:<http://id.nlm.nih.gov/mesh/>

PREFIX mnet:<https://rdf.metanetx.org/mnet/>

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

PREFIX orthodb:<http://purl.orthodb.org/>

PREFIX orthodbGroup:<http://purl.orthodb.org/odbgroup/>

PREFIX patent:<http://data.epo.org/linked-data/def/patent/>

PREFIX pubmed:<http://rdf.ncbi.nlm.nih.gov/pubmed/>

PREFIX rh:<http://rdf.rhea-db.org/>

PREFIX schema:<http://schema.org/>

PREFIX sh:<http://www.w3.org/ns/shacl#>

PREFIX sio:<http://semanticscience.org/resource/>

PREFIX slm:<https://swisslipids.org/rdf/>

PREFIX sp:<http://spinrdf.org/sp#>

PREFIX uberon:<http://purl.obolibrary.org/obo/uo#>

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

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

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

Answer: 281303435

SELECT (STR(COUNT(?protein)) as ?count)

WHERE{?protein rdf:type up:Protein}

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

Answer: 89182

SELECT (STR(COUNT(?protein)) as ?count)

WHERE{

?protein a up:Protein;

up:organism taxon:3702

}

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

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

SELECT ?activity\_label

WHERE

{

uniprot:Q9SZZ8 a up:Protein ;

up:enzyme ?enzyme.

?enzyme up:activity ?activity.

?activity rdfs:label ?activity\_label

}

**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”)

Answer: 58202805

SELECT ?id (COUNT(?date) AS ?date\_count)

WHERE {

?protein a up:Protein ;

up:created ?date

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

}

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

Answer: 1766921

SELECT (STR(COUNT(DISTINCT ?taxon)) AS ?count)

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

WHERE{

?taxon a up:Taxon ;

up:rank up:Species

}

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

Answer: 984622

SELECT (STR(COUNT(DISTINCT ?taxon)) AS ?count)

WHERE

{

?protein a up:Protein ;

up:organism ?taxon .

?taxon up:rank up:Species

}

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

IM JUST ADDING THIS PREFIX TO THE ONES I HAD, BECAUSE I HAD TO USE SOME OF THE UNIPROT PREFIXES TOO

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

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

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

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

PREFIX dc: <http://purl.org/dc/elements/1.1/>

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

PREFIX dbpedia2: <http://dbpedia.org/property/>

PREFIX dbpedia: <http://dbpedia.org/>

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

PREFIX skos: [http://www.w3.org/2004/02/skos/core#](http://www.w3.org/2004/02/skos/core)

**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)

Answer: NO ANSWER (email)

SELECT distinct ?pred

WHERE

{

SERVICE<http://sparql.uniprot.org/sparql>

{

?protein a up:Protein ;

up:organism taxon:3702 ;

up:recommendedName ?name .

?name up:fullName ?full .

FILTER CONTAINS( ?full, 'APETALA 3') .

}

?probe ?pred ?o

}

LIMIT 30

This is the last query I got results. I asked for all the predicates and got a list of more than 200, but none of them had a result when I asked for the predicates.

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

SELECT distinct ?s

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

WHERE {

?s a atlasterms:EnsemblDatabaseReference;

rdfs:label "AP3"

}

http://rdf.ebi.ac.uk/resource/ensembl/AT3G54340

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

I ALSO ADDED THIS TWO TO THE ONES I HAD

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

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

**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….)

Answer: 809

SELECT (COUNT (DISTINCT ?pathway) AS ?count)

WHERE

{

?pathway a biopax3:Pathway ;

biopax3:organism tax:3702

}

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

Answer: all references in a list

SELECT DISTINCT (str(?Id) AS ?pubmed\_ID)

WHERE

{

?pathway a biopax3:Pathway ;

biopax3:displayName ?name ;

biopax3:xref ?ref .

?red biopax3:db ?db ;

biopax3:id ?Id .

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

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

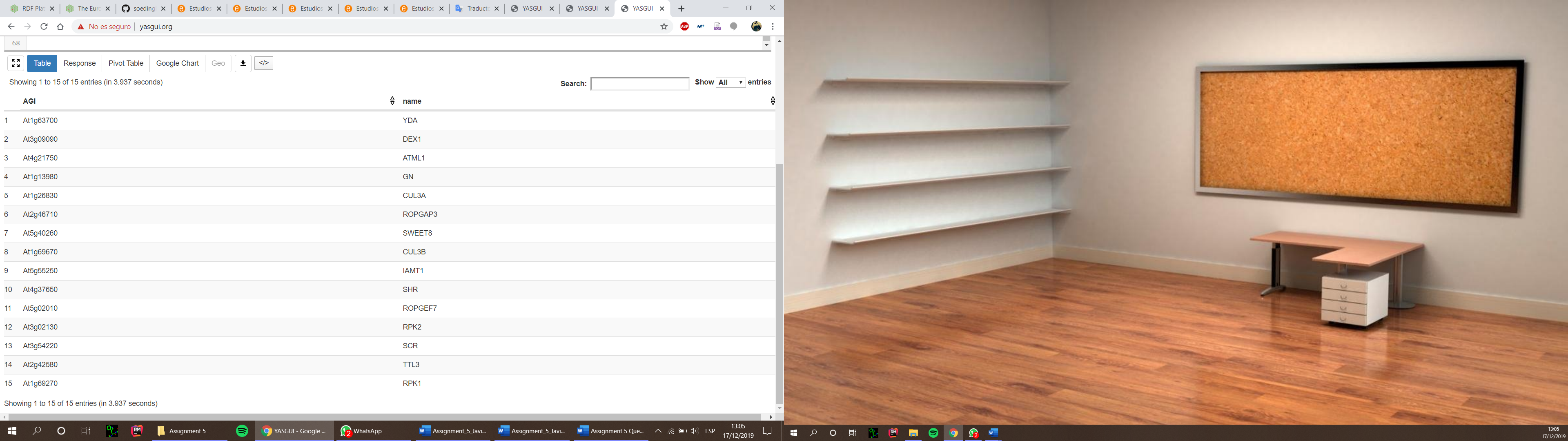
}

LIMIT 100 # I put this limit to ensure that I’m receiving the output correctly, in order to get all, don’t put the limit

# **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”

Answer: table with AGIs and the name of the gene



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')

}

**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!!!)

Answer: count1: 654481 // count2: 654481

THIS MEANS THAT THE 100% of the pathways annotations are “inferred from electronic annotation”

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

WHERE

{

?pathway1 a biopax3:Pathway ;

biopax3:organism taxon:3702.

?pathway2 a biopax3:Pathway ;

biopax3:organism taxon:3702;

biopax3:evidence ?evidence .

?evidence biopax3:evidenceCode ?evidenceCode .

?evidenceCode biopax3:term ?term .

}