Assignment 5

UniProt SPARQL Endpoint: http://sparql.uniprot.org/sparql/

1 POINT How many protein records are in UniProt?

281303435 records

```
PREFIX up:<http://purl.uniprot.org/core/>
SELECT (COUNT(DISTINCT  ?protein) AS ?count)
WHERE{
      ?protein a up:Protein .
}
```

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

89182 records

```
PREFIX up:<http://purl.uniprot.org/core/>
PREFIX taxon:<http://purl.uniprot.org/taxonomy/>
SELECT (COUNT(DISTINCT ?protein) AS ?count)
WHERE{
      ?protein a up:Protein .
      ?protein up:organism taxon:3702 .  #select only A. thaliana proteins
}
```

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 up:<http://purl.uniprot.org/core/>
PREFIX uniprot:<http://purl.uniprot.org/uniprot/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX label:<http://www.w3.org/2004/02/skos/core#>

SELECT ?description
WHERE {
  uniprot:Q9SZZ8 a up:Protein ;
                 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

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

SELECT ?id ?date
WHERE{
  ?protein a up:Protein .
  ?protein up:mnemonic ?id .
  ?protein up:created ?date .
  FILTER (contains(STR(?date), "2019"))
}
```

1 POINT How many species are in the UniProt taxonomy?

1766921 species

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


SELECT (COUNT(DISTINCT ?taxon) AS ?count)
WHERE{
  ?taxon a up:Taxon .
  ?taxon up:rank up:Species
}
```

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

The query takes too long to load

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

SELECT (COUNT(DISTINCT ?taxon) AS ?count)
WHERE{
  ?taxon a up:Taxon .
  ?taxon up:rank up:Species .
  ?protein a up:Protein .
}
```

Atlas gene expression database

SPARQL Endpoint: http://www.ebi.ac.uk/rdf/services/atlas/sparql

NO 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 rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX atlasterms: <http://rdf.ebi.ac.uk/terms/expressionatlas/>

SELECT distinct ?description
FROM <http://rdf.ebi.ac.uk/dataset/expressionatlas>
WHERE {
  ?gene rdfs:label 'AP3' .
  ?s atlasterms:refersTo ?gene .
  ?s a atlasterms:DecreasedDifferentialExpressionRatio .
  ?s atlasterms:isOutputOf ?a .
  ?a rdfs:label ?description .
}
```

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 pathways

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


SELECT (COUNT(DISTINCT ?pathway) AS ?count)
WHERE {
  ?pathway a biopax3:Pathway .
  ?pathway biopax3:organism ?a .
  FILTER contains(STR(?a),'3702')
 }
```

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

7 references

```
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 ?ref) AS ?count)
WHERE {
  ?pathway rdf:type biopax3:Pathway .
  ?pathway biopax3:displayName ?pathwayname .
  FILTER CONTAINS(STR(?pathwayname),'Degradation of the extracellular matrix') .
  ?pathway biopax3:xref ?ref .
  ?ref a biopax3:PublicationXref .
  FILTER CONTAINS(STR(?ref),'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”

There are 15 results (SELECT (count(distinct ?AGI) as ?count))

```
PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
PREFIX up:<http://purl.uniprot.org/core/>
PREFIX taxon:<http://purl.uniprot.org/taxonomy/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>

SELECT ?AGI ?name
WHERE{
  ?protein a up:Protein .
  ?protein up:organism taxon:3702 .
  ?protein up:annotation ?annotation .
  ?annotation a up:Function_Annotation ;
          rdfs:comment ?c .
  FILTER (contains(STR(?c), "pattern formation")).

 ?protein up:encodedBy ?gene .
  ?gene up:locusName ?AGI ;
               skos:prefLabel ?name .

}
```

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!!!)

809 have IEA

```
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 ?evidence) as ?e2)
WHERE {
  ?pathway a biopax3:Pathway .
  ?pathway biopax3:organism ?a .
  FILTER contains(STR(?a),'3702') .
  ?pathway biopax3:evidence ?evidence . #delete code after this to get all annotations
  ?evidence biopax3:evidenceCode ?code .
  FILTER CONTAINS(STR(?code),'EvidenceCodeVocabulary1').
}
```