Assignment 5 SPARQL queries

Adrián López Beltrán

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/>

SELECT (STR(COUNT(?prot)) as ?prot_number)
WHERE
{       ?prot a up:Protein
}
```

prot_number: 281303435

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

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


SELECT (STR(COUNT(?protein)) as ?prot_number)
WHERE
{
        ?protein a up:Protein .
        ?protein ?p ?taxon .
        ?taxon <http://purl.uniprot.org/core/scientificName> "Arabidopsis thaliana"
}
```

prot_number: 89182

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_description
WHERE
{
        uniprot:Q9SZZ8 a up:Protein ;
                        up:enzyme ?enzyme.
        ?enzyme up:activity ?activity.
        ?activity <http://www.w3.org/2000/01/rdf-schema#label> ?activity_description
}
```

activity_description:

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 taxon: <http://purl.org/biodiversity/taxon/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX up:<http://purl.uniprot.org/core/>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>

SELECT ?protein (STR(?date) AS ?date_str)
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?

```
PREFIX taxon: <http://purl.org/biodiversity/taxon/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX up:<http://purl.uniprot.org/core/>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
PREFIX up:<http://purl.uniprot.org/core/>


SELECT DISTINCT (STR(COUNT(?specie)) as ?species_number)
WHERE
{
        ?taxon <http://purl.uniprot.org/core/scientificName> ?specie .
        ?taxon up:rank up:Species
}
```

species_number: 1766921

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

```
SELECT (STR(COUNT(DISTINCT ?taxon)) AS ?species_count)
WHERE
{
        ?protein a up:Protein ;
                up:organism ?taxon .
        ?taxon up:rank up:Species
}
```

species_count: 984622

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)

CANNOT BE DONE

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….)

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

SELECT (COUNT (DISTINCT ?pathway) AS ?pathways)
WHERE
{
        ?pathway a biopax3:Pathway ;
                        ?p tax:3702
}
```

?pathways: 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#>
PREFIX tax:<http://identifiers.org/taxonomy/>

SELECT DISTINCT (str(?pubmedId) AS ?pubmed_ID)
WHERE
{
        ?pathway a biopax3:Pathway ;
                        biopax3:displayName ?name ;
                        biopax3:xref ?ref .
        ?red biopax3:db ?db ;
                biopax3:id ?pubmedId .
        FILTER(str(?name) = 'Degradation of the extracellular matrix') .
        FILTER(str(?db) ='Pubmed')
}
```

?pubmed_ID: 25,959 entries

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 ?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')
}
```

number of entries: 15

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

SELECT (COUNT (?pathway1) AS ?all) (COUNT (?pathway2) AS ?electronic)
WHERE
{

   ?pathway1 a biopax3:Pathway ;
         biopax3:organism taxon:3702 ;
                  biopax3:evidence ?evidence1 .
   ?evidence1 biopax3:evidenceCode ?evidenceCode1 .
   ?evidenceCode1 biopax3:term ?term1 .

   ?pathway2 a biopax3:Pathway ;
         biopax3:organism taxon:3702 ;
          biopax3:evidence ?evidence2 .
   ?evidence2 biopax3:evidenceCode ?evidenceCode2 .
   ?evidenceCode2 biopax3:term ?term2 .

   FILTER REGEX(?term2, 'inferred from electronic annotation')

}
```

number of entries: 654481