## 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...

NOTES: I did this assignment in collaboration with the rest of my classmates. Some of us have previously worked with SPARQL queries for another course of the master's program (Semantic Technologies), so if you need me to explain the queries in details, please tell me.

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

### 1. 1 POINT How many protein records are in UniProt?

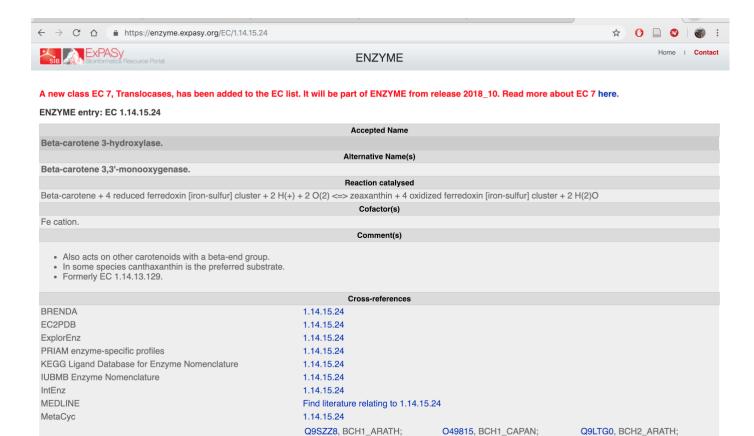
PREFIX up:<a href="http://purl.uniprot.org/core/">PREFIX up:<a href="http://purl.uniprot.org/">PREFIX up:<

There are 223102577 protein records.

#### 2. 1 POINT How many Arabidopsis thaliana protein records are in UniProt?

#### There are 89247 protein records

### 3. 1 POINT: What is the description of the enzyme activity of UniProt Protein Q9SZZ8



4. 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:<a href="http://purl.uniprot.org/core/">PREFIX up:<a href="http://purl.uniprot.org/">PREFIX up:<a href="http://

PREFIX xsd:<a href="http://www.w3.org/2001/XMLSchema#">http://www.w3.org/2001/XMLSchema#</a>

SELECT ?protein ?date

WHERE {

}

?protein a up:Protein .

?protein up:created ?date .

FILTER (?date >= "2018-01-01"^^xsd:date)

There are so many results, I am not going to put them here, but the code works!

#### 5. 1 POINT How many species are in the UniProt taxonomy?

PREFIX up:<a href="http://purl.uniprot.org/core/">PREFIX up:<a href="http://purl.uniprot.org/">PREFIX up:<a href="http://

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

FROM <a href="http://sparql.uniprot.org/taxonomy">http://sparql.uniprot.org/taxonomy</a>

WHERE {

?taxon a up:Taxon .

```
?taxon up:rank up:Species }
```

There are 1601900 species in the Uniprot taxonomy.

### 6. 1 POINT How many species have at least one protein record?

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

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

WHERE {

?protein a up:Protein .

?protein up:organism ?taxon .

?taxon up:rank up:Species}

833130 species have at least 1 protein record.

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

This link does not work, it leads you to http://www.ebi.ac.uk/rdf/services/sparql

None of these Atlas/SPARQL queries worked.

7. 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)

```
PREFIX dcterms: <a href="http://purl.org/dc/terms/">http://purl.org/dc/terms/</a>
PREFIX atlasterms: <a href="http://rdf.ebi.ac.uk/terms/expressionatlas">http://rdf.ebi.ac.uk/terms/expressionatlas</a>
PREFIX up:<a href="http://purl.uniprot.org/core/">http://purl.uniprot.org/core/</a>
PREFIX taxon:<a href="http://purl.uniprot.org/taxonomy/">http://purl.uniprot.org/taxonomy/</a>
SELECT ?id
WHERE {

SERVICE<a href="http://sparql.uniprot.org/sparql">http://sparql.uniprot.org/sparql</a>
{

?protein a up:Protein .

?protein up:organism taxon:3702 .

?protein up:recommendedName ?name .

?name up:fullName ?full .
```

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

```
}
 ?probe atlasterms:dbXref ?protein .
 ?probe dcterms:identifier ?id
}
```

# 8. 3 POINTS - get the experimental description for all experiments where the Arabidopsis Apetala3 gene is

```
DOWN regulated
PREFIX dcterms: <a href="http://purl.org/dc/terms/">http://purl.org/dc/terms/</a>
PREFIX sio: <a href="http://semanticscience.org/resource/">http://semanticscience.org/resource/</a>
PREFIX atlas: <a href="http://rdf.ebi.ac.uk/resource/expressionatlas/">http://rdf.ebi.ac.uk/resource/expressionatlas/</a>
PREFIX atlasterms: <a href="http://rdf.ebi.ac.uk/terms/expressionatlas/">http://rdf.ebi.ac.uk/terms/expressionatlas/</a>
PREFIX up:<a href="http://purl.uniprot.org/core/">PREFIX up:<a href="http://purl.uniprot.org/">PREFIX up:<a href="http://
PREFIX taxon:<a href="http://purl.uniprot.org/taxonomy/">http://purl.uniprot.org/taxonomy/>
SELECT ?desc
WHERE {
                             SERVICE<a href="http://sparql.uniprot.org/sparql">http://sparql.uniprot.org/sparql</a>
      {
                             ?protein a up:Protein .
                             ?protein up:organism taxon:3702.
                             ?protein up:recommendedName ?name .
                                                          ?name up:fullName ?full .
                                                          FILTER CONTAINS(?full, 'APETALA 3').
      }
        ?probe atlasterms:dbXref ?protein .
                             ?differential atlasterms:isMeasurementOf ?probe .
                             ?dea atlasterms:hasExpressionValue ?differential .
                             ?experiment atlasterms:hasAnalysis ?dea .
                             ?experiment dcterms:description ?desc .
                             ?differential sio:SIO 000300 ?value .
                             FILTER CONTAINS(?value, 'DOWN') }
```

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

#### Again, this link leads to <a href="http://www.ebi.ac.uk/rdf/services/sparql">http://www.ebi.ac.uk/rdf/services/sparql</a>

9. 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: <a href="http://www.biopax.org/release/biopax-level3.owl#>">

PREFIX tax:<a href="http://identifiers.org/taxonomy/">http://identifiers.org/taxonomy/>

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

WHERE {

?pathway a biopax3:Pathway .

?pathway biopax3:organism tax:3702}

#### There are 809 REATOME pathways for Arabidopsis.

# 10. 3 POINTS: get all PubMed references for the pathway with the name "Degradation of the extracellular matrix"

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

SELECT DISTINCT ?pubmedId

WHERE {

?pathway a biopax3:Pathway .

?pathway biopax3:displayName ?name .

?pathway biopax3:xref ?ref .

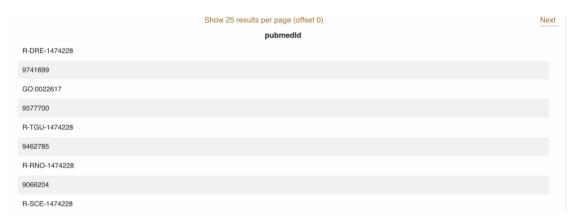
?red biopax3:db ?db.

?ref biopax3:id ?pubmedId .

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

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

#### Result example:



#### **BONUS QUERIES**

<u>UniProt BONUS 2 points</u>: find the AGI codes and gene names for all Arabidopsis thaliana proteins that have a protein function annotation description that mentions "pattern formation"

### Results

}

<b>★</b> Sparql XML	<b>★</b> SparqI JSON	<b>★</b> CSV	Share	
agi				name
At3g09090				DEX1
At3g02130				RPK2
At4g21750				ATML1
At2g46710				ROPGAP3
At5g55250				IAMT1
At1g13980				GN
At5g40260				SWEET8
At5g02010				ROPGEF7
At1g69670				CUL3B
At1g63700				YDA
At3g54220				SCR
At4g37650				SHR
At1g26830				CUL3A
At1g69270				RPK1
At2g42580				TTL3

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: <a href="http://www.w3.org/1999/02/22-rdf-syntax-ns#">http://www.w3.org/1999/02/22-rdf-syntax-ns#</a>

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

PREFIX taxon: <a href="http://identifiers.org/taxonomy/">http://identifiers.org/taxonomy/>

SELECT(COUNT (DISTINCT ?evidence1) AS ?all) (COUNT (DISTINCT ?evidence2) AS ?electronic)

WHERE {

?pathway a biopax3:Pathway .

?pathway biopax3:organism taxon:3702.

?pathway biopax3:evidence ?evidence1.

?pathway biopax3:evidence ?evidence2.

?evidence1 biopax3:evidenceCode ?evidenceCode1.

?evidence2 biopax3:evidenceCode ?evidenceCode2 .

?evidenceCode1 biopax3:term ?term1 .

?evidenceCode2 biopax3:term ?term2 .

FILTER (str(?term2) = 'inferred from electronic annotation')

}

Results	Query history	Named Graphs				
		all	Show 25 results pe	r page (offset 0)	electronic	
809			809			