

# 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:<http://purl.uniprot.org/core/>
```

```
SELECT (COUNT(DISTINCT ?protein) AS ?count)
```

```
WHERE {?protein a up:Protein}
```

**There are 223102577 protein records.**

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

```
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}
```

**There are 89247 protein records**

## 3. 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/>
```



```
SELECT ?activity
```

```
WHERE {
```

```
    uniprot:Q9SZZ8 a up:Protein ;
```

```
    up:enzyme ?activity}
```

← → ↻ 🏠 🔒 https://enzyme.expasy.org/EC/1.14.15.24 ☆ ⓘ 📄 📺 🌐 ⋮

  **ExPASy**  
Bionformatics Resource Portal

**ENZYME**

Home | **Contact**

**A new class EC 7, Translocases, has been added to the EC list. It will be part of ENZYME from release 2018\_10. Read more about EC 7 [here](#).**

**ENZYME entry: EC 1.14.15.24**

Accepted Name	
Beta-carotene 3-hydroxylase.	
Alternative Name(s)	
Beta-carotene 3,3'-monooxygenase.	
Reaction catalysed	
Beta-carotene + 4 reduced ferredoxin [iron-sulfur] cluster + 2 H(+) + 2 O(2) <=> zeaxanthin + 4 oxidized ferredoxin [iron-sulfur] cluster + 2 H(2)O	
Cofactor(s)	
Fe cation.	
Comment(s)	
<ul style="list-style-type: none"> <li>Also acts on other carotenoids with a beta-end group.</li> <li>In some species canthaxanthin is the preferred substrate.</li> <li>Formerly EC 1.14.13.129.</li> </ul>	
Cross-references	
BRENDA	1.14.15.24
EC2PDB	1.14.15.24
ExplorEnz	1.14.15.24
PRIAM enzyme-specific profiles	1.14.15.24
KEGG Ligand Database for Enzyme Nomenclature	1.14.15.24
IUBMB Enzyme Nomenclature	1.14.15.24
IntEnz	1.14.15.24
MEDLINE	<a href="#">Find literature relating to 1.14.15.24</a>
MetaCyc	1.14.15.24
Q9SZZ8, BCH1_ARATH;      O49815, BCH1_CAPAN;      Q9LTG0, BCH2_ARATH;	

#### 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:<http://purl.uniprot.org/core/>

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

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

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

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

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

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

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

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

SELECT ?id

WHERE {

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

    {

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

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

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

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

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

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

SELECT ?desc

WHERE {

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

    {

        ?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 <http://www.ebi.ac.uk/rdf/services/sparql>

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

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

SELECT DISTINCT ?pubmedId

WHERE {

?pathway a biopax3:Pathway .

?pathway biopax3:displayName ?name .

?pathway biopax3:xref ?ref .

?ref biopax3:db ?db .

?ref biopax3:id ?pubmedId .

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

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

Result example:

Show 25 results per page (offset 0)		Next
	pubmedId	
R-DRE-1474228		
9741699		
GO:0022617		
9577700		
R-TGU-1474228		
9462785		
R-RNO-1474228		
9066204		
R-SCE-1474228		

## 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 .

    ?protein up:organism taxon:3702 .

    ?protein up:encodedBy ?gene .

    ?gene up:locusName ?agi .

    ?gene skos:prefLabel ?name .

    ?protein up:annotation ?annot .

    ?annot a up:Function\_Annotation ;

        rdfs:comment ?annotComment .

    FILTER CONTAINS(str(?annotComment), 'pattern formation')

}

## Results

[Sparql XML](#) [Sparql JSON](#) [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: <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 (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
Show 25 results per page (offset 0)		
809	all	electronic
	809	