Assignment 5 SPARQL queries

I would like you to create the SPARQL query that will answer each of these questions. Please submit the queries as a Jupyter notebook with the SPARQL kernel activated. NO programming is required! Submit to GitHub as usual, WITH THE ANSWERS STILL VISIBLE IN THE NOTEBOOK. Thanks!

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

Q1: 1 POINT How many protein records are in UniProt?
For this question, I just called protein records from Uniprot.
There are 360 157 660 protein records.

Q2: 1 POINT How many Arabidopsis thaliana protein records are in UniProt?
For this question I called protein records like in the previous question and I called the organism which have the taxon id 3702, corresponding to Arabidopsis thaliana. I found it on Uniprot.
There are 136 782 Arabidopsis thaliana protein records.

Q3: 1 POINT retrieve pictures of Arabidopsis thaliana from UniProt?
For this question, I helped me with an example from the lesson 8.

https://upload.wikimedia.org/wikipedia/commons/3/39/Arabidopsis.jpg
https://upload.wikimedia.org/wikipedia/commons/thumb/6/60/Arabidopsis_thaliana_inflorescencias.jpg/800px-Arabidopsis_thaliana_inflorescencias.jpg

Q4: 1 POINT: What is the description of the enzyme activity of UniProt Protein Q9SZZ8

For this question, I called the protein Q9SZZ8 and I identified it as an enzyme to find its activity. "Beta-carotene + 4 reduced ferredoxin [iron-sulfur] cluster + 2 H(+) + 2 O(2) = zeaxanthin + 4 oxidized ferredoxin [iron-sulfur] cluster + 2 H(2)O."

Q5: 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")
For this assignment, it didn't work on jupyter notebook or it was too long. However, it worked on https://sparql.uniprot.org/sparql.

id	date
"A0A1H7ADE3_PAEPO"xsd:string	"2021-06-02"xsd:date
"A0A1V1AIL4_ACIBA"xsd:string	"2021-06-02"xsd:date
"A0A2Z0L603_ACIBA"xsd:string	"2021-06-02"xsd:date
"A0A4J5GG53_STREE"xsd:string	"2021-04-07"xsd:date
"A0A6G8SU52_AERHY"xsd:string	"2021-02-10"xsd:date
"A0A6G8SU69_AERHY"xsd:string	"2021-02-10"xsd:date
"A0A7C9JLR7_9BACT"xsd:string	"2021-02-10"xsd:date
"A0A7C9JMZ7_9BACT"xsd:string	"2021-02-10"xsd:date
"A0A7C9KUQ4_9RHIZ"xsd:string	"2021-02-10"xsd:date
"A0A7D4HP61_NEIMU"xsd:string	"2021-02-10"xsd:date
"A0A7D6A5N9_SERMA"xsd:string	"2021-06-02"xsd:date
"A0A7D6FMY9_9ENTR"xsd:string	"2021-02-10"xsd:date
"A0A7D6VKU9_CITFR"xsd:string	"2021-02-10"xsd:date
"A0A7D6VKZ9_CITFR"xsd:string	"2021-02-10"xsd:date
"A0A7D7EJU1_CITFR"xsd:string	"2021-02-10"xsd:date
"A0A7D7HYH9_ECOLX"xsd:string	"2021-02-10"xsd:date
"A0A7G5HK20_9BSED"xsd:string	"2021-02-10"xsd:date

Q6: 1 POINT How many species are in the UniProt taxonomy?
2 029 846 species

Q7: 2 POINT How many species have at least one protein record? (this might take a long time to execute, so do this one last!)
This command was too long so I didn't succeed to run it. But it's present on the jupyter notebook. As in sparkql, the succession of orders is like successive additions (&) that need to be all true. So, for this question, I did like the previous question and I just add the command calling protein records.

Q8: 3 points: find the AGI codes and gene names for all Arabidopsis thaliana proteins that have a protein function annotation description that mentions "pattern formation"
I found 19 results:

AGI	At3g54220	At1g13980	At4g21750	At5g40260	At1g69670	At1g63700
Gene name	SCR	GN	ATML1	SWEET8	CUL3B	YDA

At2g46710	At1g26830	At1g55325	At3g09090	At4g37650	At5g55250	At3g02130
ROPGAP3	CUL3A	MED13	DEX1	SHR	IAMT1	RPK2

At2g42580	At1g69270	At5g02010	At1g66470	At1g49770	At5g37800
TTL3	RPK1	ROPGEF7	RHD6	BHLH95	RSL1

From the MetaNetX metabolic networks for metagenomics database SPARQL Endpoint: https://rdf.metanetx.org/sparql

(this slide deck will make it much easier for you! https://www.metanetx.org/cgi-bin/mnxget/mnxref/MetaNetX_RDF_schema.pdf)

Q9: 4 POINTS: what is the MetaNetX Reaction identifier (starts with "mnxr") for the UniProt Protein uniprotkb:Q18A79

Here the MetaNetX Reaction identifier for the UniProt Protein uniprotkb:Q18A79 :

mnxr165934
mnxr145046c3

FEDERATED QUERY - UniProt and MetaNetX

Q10: 5 POINTS: What is the official Gene ID (UniProt calls this a "mnemonic") and the MetaNetX Reaction identifier (mnxr…..) for the protein that has "Starch synthase" catalytic activity in Clostridium difficile (taxon 272563).

I don't succeed but I try some manipulations on the jupyter notebook.