Using SPARQL to assemble data on human proteins - All that glitters is not gold!

Lydie Lane
www.nextprot.org is the reference knowledge base for HPP

- MS-data
- Antibody-based data
- RNA-seq data
- PTMs
- Variants (genomic and somatic)
- PPIs
- Functional annotations

- Tools for the MS community
- Third-party tool for function prediction
- SPARQL-based search tools to query this data
www.nextprot.org homepage
neXtProt provides metrics for the C-HPP
neXtProt provides tools for the C-HPP
Advanced, SPARQL-based interface to explore the data

Retrieves lists of entries with associated information

>140 pre-made queries available
SNORQL interface [https://snorql.nextprot.org](https://snorql.nextprot.org)

Retrieves any data as a table

>170 pre-made queries available
www.nextprot.org does not contain everything....

- MS-data
- Antibody-based data
- RNA-seq data
- PTMs
- Variants (genomic and somatic)
- PPIs
- Functional annotations
  - Phylogenetic information
  - Data on model organisms
  - Pharmacology data
  - Clinical proteogenomic data
  - Structural data
  - Protein interactions with pathogens
  - .....
Other resources and tools could fill these gaps

- Phylogenetic information
- Pharmacology data
- Clinical proteogenomic data
- Structural data
- Data on model organisms
- Protein interactions with pathogens
- ......
Federated SPARQL queries in neXtProt
Example of federated query

SPARQL endpoint: https://api.nextprot.org/sparql

PREFIX...  
PREFIX wp: <http://vocabularies.wikipathways.org/wp#>  
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>  
PREFIX dcterms: <http://purl.org/dc/terms/>  

```sparql
select distinct ?pathname group_concat(distinct ?gen , ',') as ?genes where {
  service <http://sparql.wikipathways.org/sparql> {    
  SELECT DISTINCT ?pathname ?gen WHERE {
    { ?geneProduct a wp:GeneProduct . }
    union
    { ?geneProduct a wp:Protein . }
    filter(lregexp(?gen,"^[a-z]+") ). # ensures official gene names for subsequent nextprot matching
    ?pathway a wp:Pathway .
    FILTER(contains(str(?organism),"9606"))
  } } }
}
```

Human pathways in which at least one protein is mitochondrial GOLD
Vision: enhance neXtProt contents by federating existing resources

Pharmacology/Clinical data

LinkedOmics

Systems biology

Phylogeny

Host-pathogen interactions

Model organisms

Structural biology
Federated queries rely on (functional) SPARQL endpoints...
How can we engage more resources and users into the semantic data field?

**Issues**

- Not all resources have a (fully functional) SPARQL endpoint
- Users need to have some experience in database querying and understand the data model(s) of the resource(s) to be queried

**Potential solutions (all ongoing)**

- Help other resources develop their own SPARQL endpoint (« BYODs »)
- Develop user-friendly query interfaces (graphical design and/or natural language)
- Develop training
A one-day tutorial on 9 independent SPARQL endpoints hosted by the SIB, including neXtProt, UniProt, GlyConnect, Rhea, OrthoDB, OMA and Bgee.