Implementation of flexible search for proteomics metadata

http://dev.jpost.org/px-rdf

Database Center for Life Science (DBCLS), Japan
Shin Kawano

Database Center for Life Science (DBCLS), Japan
Yuki Moriya

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), UK
Juan Antonio Vizcaino

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), UK
Tobias Ternent

Institute for Systems Biology (ISB), USA
Eric Deutsch

Introduction

The ProteomeXchange (PX) Consortium (http://www.proteomexchange.org) provides a globally coordinated data submission and dissemination platform for mass spectrometry proteomics data in the public domain, involving the main existing proteomics repositories.The members of the Consortium are PRIDE (https://www.ebi.ac.uk/pride), PeptideAtlas/PASSEL (http://www.peptideatlas.org/passel), MassIVE (https://massive.ucsd.edu), and jPOST, which has just joined the Consortium (http://jpost.org). Public datasets from the different members can be accessed into a common interface called ProteomeCentral (http://proteomecentral.proteomexchange.org). A set of technical and biological common metadata about the datasets has been agreed by the PX members. Although the ProteomeCentral web interface (Fig. 1) provides a state-of-the-art search functionality, it is not well-suited to construct more complex searches. In the context of 'Linked Open Data', a concept about connecting data independently of the involved biological data types, we chose the Resource Description Framework (RDF) data model to achieve this intended more advanced search functionality, to improve dataset discoverability.

Methods

We designed a PX-RDF schema (Fig. 3) based on the PX-XML schema (Fig. 2). In addition to well-known ontologies (such as Dublin Core and Friend of a Friend) and proteomics domain specific controlled vocabularies (such as PSI-MS and UNIMOD) (Fig. 4), we defined a new ontology, PX ontology, which makes up for deficiencies in existing vocabularies (Fig. 5). The PX ontology and conversion program from PX-XML to PX-RDF are available at https://github.com/PX-RDF/ontology and https://github.com/PX-RDF/RDF, respectively. The converted PX-RDF files were loaded into Virtuoso, which is a database management system for RDF. The Virtuoso server can be searched by SPARQL (SPARQL Protocol and RDF Query Language), which is a query language for RDF (Fig. 6).

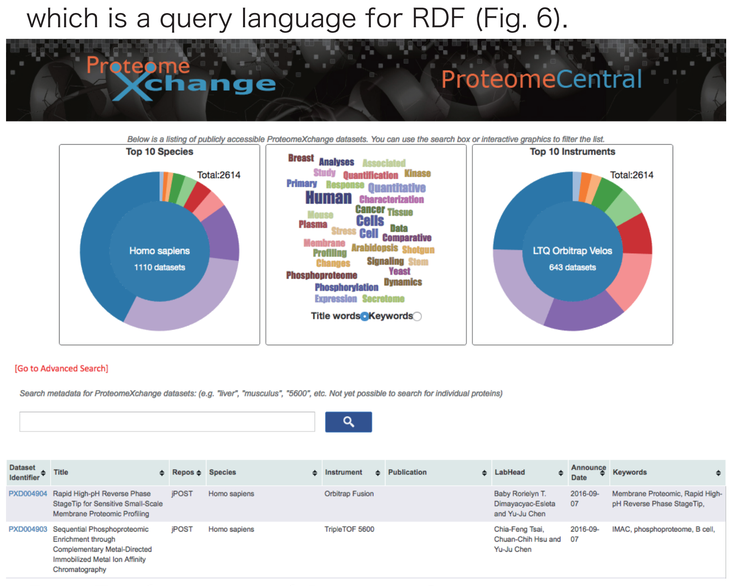

Fig. 1 The front page of ProteomeCentral
http://proteomecentral.proteomexchange.org/

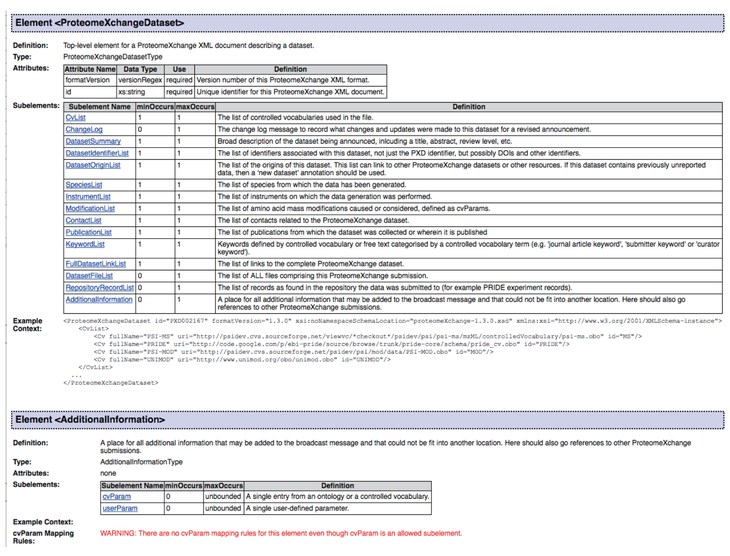

Fig. 2 A part of PX-XML schema
http://proteomecentral.proteomexchange.org/schemas/proteomeXchange-1.3.0.html

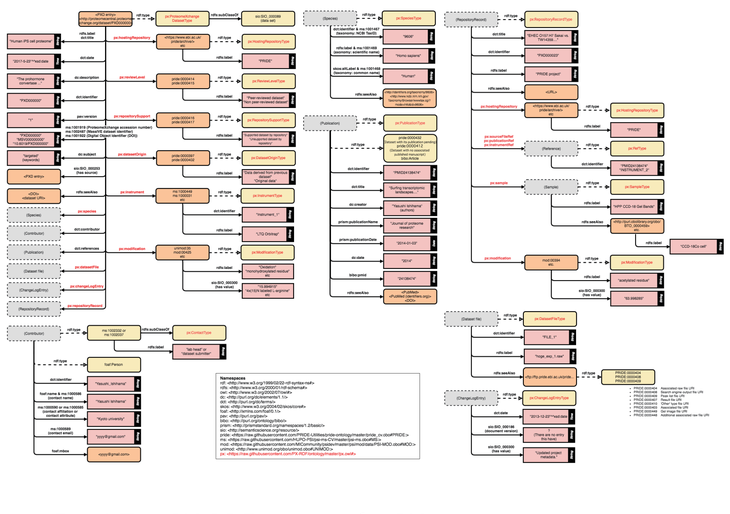

Fig. 3 The PX-RDF schema for PX metadata
https://github.com/PX-RDF/RDF/blob/master/PX-RDF.png

Namespaces
rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
owl: <http://www.w3.org/2002/07/owl#>
dc: <http://purl.org/dc/elements/1.1/>
dct: <http://purl.org/dc/terms/>
skos: <http://www.w3.org/2004/02/skos/core#>
foaf: <http://xmlns.com/foaf/0.1/>
pav: <http://purl.org/pav/>
bibo: <http://purl.org/ontology/bibo/>
prism: <http://prismstandard.org/namespaces/1.2/basic/>
sio: <http://semanticscience.org/resource/>
pride: <https://raw.githubusercontent.com/PRIDE-Utilities/pride-ontology/master/pride_cv.obo#PRIDE:>
ms: <https://raw.githubusercontent.com/HUPO-PSI/psi-ms-CV/master/psi-ms.obo#MS:>
mod: <https://raw.githubusercontent.com/MICommunity/psidev/master/psi/mod/data/PSI-MOD.obo#MOD:>
unimod: <http://www.unimod.org/obo/unimod.obo#UNIMOD:>
px: <https://raw.githubusercontent.com/PX-RDF/ontology/master/px.owl#>

Fig. 4 Used controlled vocabularies and ontologies for PX-RDF

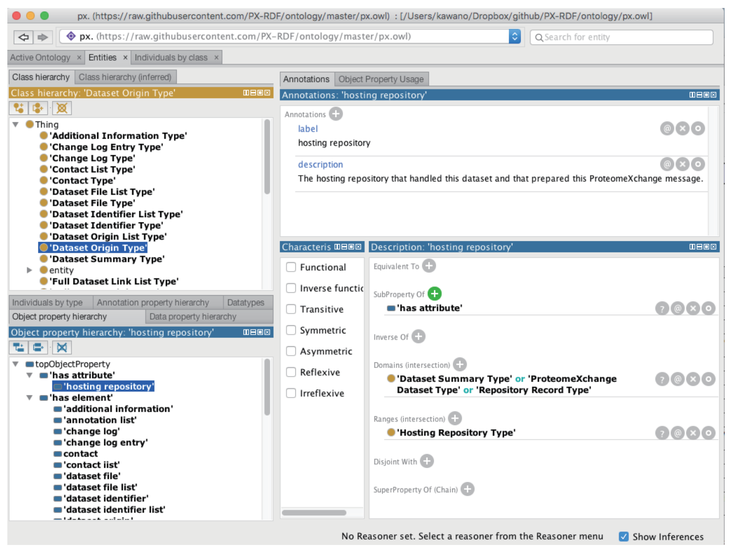

Fig. 5 The PX ontology on the Protege ontology editor

Fig. 6 SPARQL endpoint for PX-RDF
http://dev.jpost.org/px-rdf

Results and Discussion

Statistics at the end of August were:
- Entries: 2,611
- Triples: 2,064,628
- File size (tar.gz): 9.5 GB

Figs. 7 and 8 show an example SPARQL query and its output, respectively.

Since we employed the RDF data model, which is globally used e.g. in federated queries, we will not only be able to search for proteomics datasets, but also integrate these searches with datasets from other fields such as genomics, transcriptomics and metabolomics.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>      Prefixes for simplifing query
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX px: <https://raw.githubusercontent.com/PX-RDF/ontology/master/px.owl#>

SELECT ?id
WHERE {
  ?px dct:identifier ?id .                                  Getting ID
  ?px a px:ProteomeXchangeDatasetType .

  ?px px:instrument ?instrument .                           Filtering by an instrument
  ?instrument rdfs:label "LTQ Orbitrap Velos" .

  ?px px:species ?species .                                 Filtering by a species
  ?species rdfs:label "Homo sapiens (Human)"  .

  ?px dct:date ?date .                                      Filtering by date
  FILTER (?date > "2016-01-01"^^xsd:date)
}
```

Fig. 7 An example of SPARQL query. List IDs of datasets which were measured by "LTQ Orbitrap Velos" instruments, for "Homo sapiens" samples, and published after 2016.

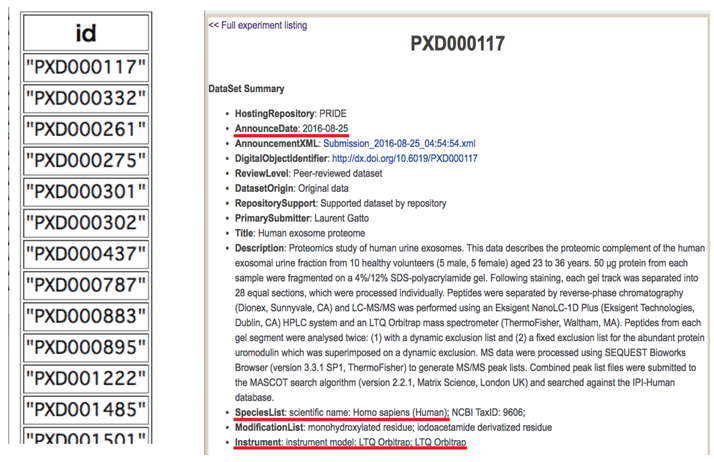

Fig. 8 Output of the query and one of the listed entry.

Future work

- Implementation of automatic update of PX-RDF using the PX RSS notification feed.
- Implementation of SPARQL endpoint on ProteomeCentral.

Conclusion

- We designed RDF schema of PX metadata, and converted from PX-XML to PX-RDF.
- PX-RDF enables users to perform much more complex and flexible query searches using SPARQL.

Acknowledgment

This work has been supported by the Database Integration Coordination Program, operated by the National Bioscience Database Center, Japan Science and Technology Agency. We would like to thank all data submitters and curators for thier contribution.