UniProt in RDF: Tackling Data Integration and Distributed Annotation with the Semantic Web

Nicole Redaschi[1] and the UniProt Consortium

[1]Swiss-Prot Group, Swiss Institute of Bioinformatics, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

The Data Integration Problem

The UniProt knowledgebase (UniProtKB) is a comprehensive repository of protein sequence and annotation data. We collect information from the scientific literature and other databases and provide links to over one hundred biological resources. Such links between different databases are an important basis for data integration, but the lack of a common standard to represent and link information makes data integration an expensive business. Data providers use diverse solutions to represent their data. Historically, many have used a proprietary syntax without formal description, making parsing an adventure. More and more are now offering their data also in standard data exchange formats, mainly XML. While XML schema languages allow to defined data structures, rendering parsing predictable, XML neither lowers the complexity of integrating many different database schemas, nor does it offer a standard solution to link resources.

The Semantic Web - A Solution?

The World Wide Web is a web of hyperlinked documents, whereas the Semantic Web (http://www.w3.org/2001/sw/) is a web of linked data. Its aim is to allow both humans and machines to navigate between databases that store information about the same thing. Creating such a web of data requires common formats for the combination of data from diverse sources, as well as language for describing the data. The two core technologies that address these requirements are the Resource Description Framework (http://www.w3.org/RDF/) and the Web Ontology Language (http://www.w3.org/2004/OWL/).

RDF in a Nutshell: Resource Description Framework

- ***Resource:*** Generalization of "Web resource": Things that can be identified on the Web, even when they cannot be directly retrieved on the Web.
- ***Description:*** A resource can be described by making ***statements*** that specify the properties and property values of the resource.
- ***Statement*** (aka ***Triple***): An RDF statement consists of:
  - ***Subject:*** Identifies the ***resource*** that the statement describes.
  - ***Predicate:*** Identifies a ***property*** of that resource.
  - ***Object:*** Identifies the ***value*** of that property.

The RDF data model is a directed graph, which can be represented as a set of simple statements of the form subject-predicate-object. To enable the linking of data on the Web, RDF requires that each resource must have a globally unique identifier. These identifiers allow everybody to make statements about a given resource and, together with the simple structure of the RDF data model, make it easy to combine statements made by different people (or databases) to allow queries across different datasets. RDF is thus an industry standard that can make a major contribution to solve two important problems of bioinformatics: distributed annotation and data integration.

Example: Combining Datasets in a Triple Store for SPARQL Queries

(http://www.w3.org/TR/rdf-sparql-query/)

```
Namespaces:
pharma  = http://www.pharma.com/owl/
target  = http://www.pharma.com/target/
up      = http://purl.uniprot.org/core/
uniprot = http://purl.uniprot.org/uniprot/

Subject:        Predicate:    Object:
target:ABC      rdf:type      pharma:Target
target:ABC      pharma:study  "phase 2"
target:ABC      owl:sameAs    uniprot:Q07820
...
uniprot:Q07820  rdf:type      up:Protein
uniprot:Q07820  up:encodedBy  "MCL1"

Query:
SELECT ?target, ?gene WHERE
{
  ?target owl:sameAs   ?x .
  ?x      rdf:type     up:Protein .
  ?x      up:encodedBy ?gene
}

Result:  target:ABC  "MCL1"
```

UniProtKB – A Case Study

- Contains many cross references to external resources and is referenced by many external resources.
- The data may form arbitrary graph structures and the data model is extended or modified several times a year.

Example: Identifiers and Cross References

1. UniProtKB Flat Text Format

While a biologist easily recognizes familiar database names in different contexts, it is impossible to write generic programs that recognize cross references that are represented in such heterogeneous ways.

```
AC   O22340; Q94FV9;
...
DE            EC=4.2.3.16;
...
OX   NCBI_TaxID=46611;
...
RX   MEDLINE=97413772; PubMed=9268308; DOI=10.1074/jbc.272.35.21784;
...
FT   VARIANT      42     42       R -> L (in dbSNP:rs41298293).
FT                                /FTId=VAR_034569.
...
DR   GO; GO:0000287; F:magnesium ion binding; IEA:InterPro.
...
DR   InterPro; IPR008949; Terpenoid_synth.
```

2. UniProtKB XML Format

An XML schema definition language can define all elements and properties that a program may encounter, but as there is no standard to represent cross references, each data provider is following own conventions.

```
<accession>O22340</accession>
<accession>Q94FV9</accession>
...
<organism key="2">
  <dbReference type="NCBI Taxonomy" id="46611" key="3" />
...
<reference key="4">
    <dbReference type="MEDLINE" id="97413772" key="5" />
    <dbReference type="PubMed" id="9268308" key="6" />
    <dbReference type="DOI" id="10.1074/jbc.272.35.21784" key="7" />
...
<feature type="sequence variant" description="In dbSNP:rs41298293."
                                 id="VAR_034569">
...
<dbReference type="EC" id="EC 4.2.3.16" key="1" />
...
<dbReference type="GO" id="GO:0000287" key="21">
  <property type="term" value="F:magnesium ion binding" />
  <property type="evidence" value="IEA:InterPro" />
</dbReference>
...
<dbReference type="InterPro" id="IPR008949" key="26">
  <property type="entry name" value="Terpenoid_synth" />
</dbReference>
```

3. UniProtKB RDF Format

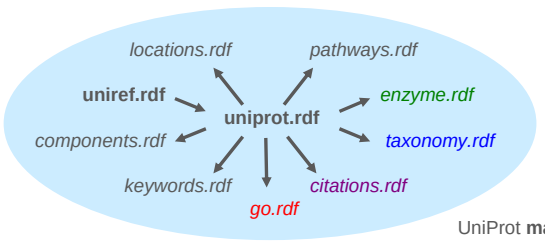
Every RDF resource is identified by at least one URIref (http://www.w3.org/Addressing/). These identifiers provide a standard way to link resources both within and between datasets.

```
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/

up       = http://purl.uniprot.org/core/
uniprot  = http://purl.uniprot.org/uniprot/
enzyme   = http://purl.uniprot.org/enzyme/
taxonomy = http://purl.uniprot.org/taxonomy/
...
```

Where possible we use standard properties (RDFS, OWL, DC). UniProt specific classes and properties are described by an OWL ontology (http://dev.isb-sib.ch/projects/uniprot-rdf/owl/).

Subject:	*Predicate:*	*Object:*
uniprot:O22340	rdf:type	up:Protein
uniprot:O22340	up:replaces	uniprot:Q94FV9
uniprot:O22340	up:enzyme	enzyme:4.2.3.16
uniprot:O22340	up:organism	taxonomy:46611
citations:9268308	owl:sameAs	medline:97413772
citations:9268308	owl:sameAs	pubmed:9268308
citations:9268308	dc:identifier	"doi:10.1074/jbc.272.35.21784"
annotation:VAR_034569	rdfs:seeAlso	dbsnp:rs41298293
uniprot:O22340	up:classifiedWith	go:0000287
uniprot:O22340	rdfs:seeAlso	interpro:IPR001906



UniProt **main** and *supporting* datasets

UniProt is mainly supported by the National Institutes of Health (NIH) grant 2 U01 HG02712-04. Additional support for the EBI's involvement in UniProt comes from the European Commission (EC)'s FELICS grant (021902RII3) and from the NIH grant 1R01HGO2273-01. Swiss-Prot activities at the SIB are supported by the Swiss Federal Government through the Federal Office of Education and Science and the European Commission contract FELICS (021902RII3). PIR activities are also supported by the NIH grants and contracts HHSN266200400061C, NCI-caBIG, and 1R01GM080646-01, and the National Science Foundation (NSF) grant IIS-0430743.

Contact
help@uniprot.org
www.uniprot.org

See also:
www.isb-sib.ch