Implementing Schema.org markup on www.UniProt.org

Jerven Bolleman[1][0000−0002−7449−1266], Parit Bansal[0000−0002−0875−1680], Sebastien Gehant[0000−0003−1608−9954], Nicole Redaschi[0000−0001−8890−2268], Alan Bridge[0000−0003−2148−9135], and the UniProt Consortium

Swiss Institute of Bioinformatics
jerven.bolleman@sib.swiss, parit.bansal@sib.swiss, sebastien.gehant@sib.swiss, nicole.redaschi@sib.swiss, alan.bridge@sib.swiss,

Abstract. The UniProt knowledgebase (UniProtKB) is a resource of protein sequences and functional information whose centerpiece is the expert-curated UniProtKB/Swiss-Prot section. UniProt data is accessible at www.uniprot.org via a user-friendly interface and a REST API that serves the data in several formats, including our own RDF formalization since 2008. In September 2014 we added Schema.org markup in RDFa encoding on our webpages. Schema.org is a community vocabulary for marking up webpages to help search engines understand the published content.

1 RDF on the UniProt website

Since UniProt provides its information as a detailed RDF model[4] via a SPARQL endpoint https://sparql.uniprot.org/, the reason to add Schema.org markup to UniProt webpages is for Search Engine Optimization. RDF and its strong semantics, expressed with RDFS/OWL[2], allow us to introduce hierarchical relations between the UniProt RDF model and the Schema.org vocabulary via `rdfs:subPropertyOf` relations. But because the major search engines do not implement this functionality, we must materialize these relations (Example 1).

We decided to encode our Schema.org markup as RDFa[1] instead of the often prefered choice JSON-LD [3] for the following reasons:

1. RDFa allows us to encode all new triples with a smaller increase in document size than an embedded JSON snippet.
2. Google, and other search engines, interpret both JavaScript and JSON-LD, but we see that websites with significant JavaScript content are crawled slower than those with a more static profile. More importantly, sites with heavy use of JavaScript are usually crawled in two runs, static and dynamic. Considering that Google makes only 300,000 runs per domain and month, and that the UniProt website has over 500 million pages, halving our crawl rate significantly impacts our visibility in search results.

The 14 Terabyte of HTML encoding nearly 6 billion Schema.org triples for all of UniProtKB is retrievable with a simple script (Example 2). The full UniProt data in our own RDF model is available at https://www.uniprot.org/downloads/ and via our SPARQL endpoint https://sparql.uniprot.org/.

```
base <http://purl.uniprot.org/citations/>
prefix up:<http://purl.uniprot.org/core/>
prefix schema:<http://schema.org>
prefix rdfs:<http://www.w3.org/2000/01/rdf-schema#>

up:title rdfs:subPropertyOf schema:name .

<26055106> a up:Journal_Citation ;
  up:title "Uncoupling protein-1 is protective of ..." ;
  schema:name "Uncoupling protein-1 is protective of ..." .
```

Example 1: Declaring that the UniProt title concept is a specialization of the schema:name concept, with materialized `schema:name` predicate.

```
for entry in $(wget -q "https://www.uniprot.org/uniprot/?format=list")
do
    rapper -i rdfa https://www.uniprot.org/uniprot/${entry} -o turtle
done
```

Example 2: Retrieving the Schema.org markup for all UniProtKB entries and converting it to Turtle format.

References

1. Adida, B., Herman, I., Birbeck, M., McCarron, S.: RDFa core 1.1 - third edition. W3C recommendation, W3C (Mar 2015), http://www.w3.org/TR/2015/REC-rdfa-core-20150317/
2. Schneider, M.: OWL 2 web ontology language RDF-based semantics (second edition). W3C recommendation, W3C (Dec 2012), http://www.w3.org/TR/2012/REC-owl2-rdf-based-semantics-20121211/
3. Sporny, M., Kellogg, G., Lanthaler, M.: JSON-ld 1.0. W3C recommendation, W3C (Jan 2014), http://www.w3.org/TR/2014/REC-json-ld-20140116/
4. The UniProt Consortium: Uniprot: the universal protein knowledgebase