Implementing Schema.org markup on www.UniProt.org

Jerven Bolleman[1], **Parit Bansal**[1], **Sebastien Gehant**[1], **Alan Bridge**[1], **Nicole Redaschi**[1], **and the UniProt Consortium**[1,2,3]

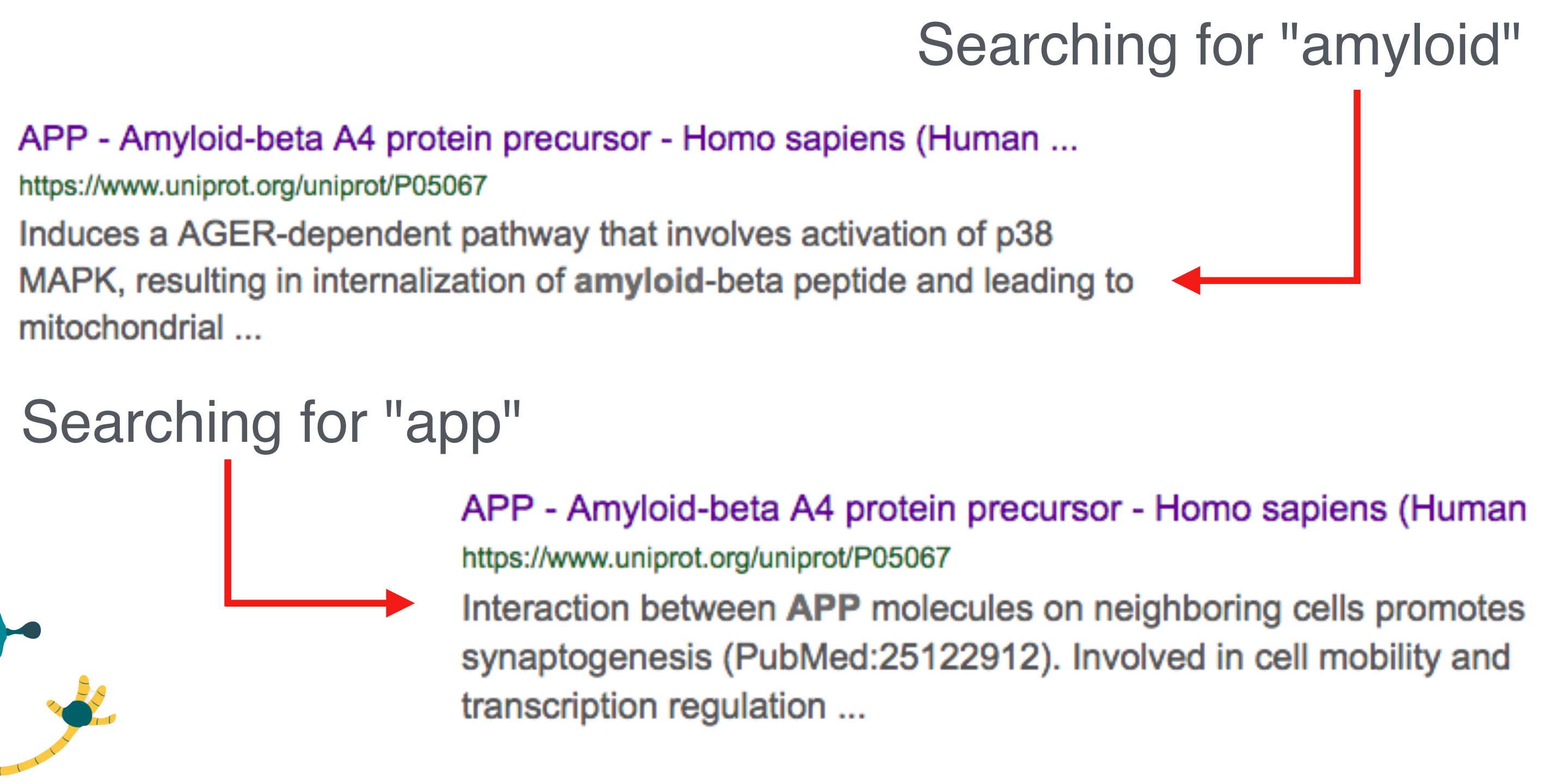
[1] Swiss-Prot Group, SIB Swiss Institute of Bioinformatics, Geneva, Switzerland
[2] EMBL-European Bioinformatics Institute, Cambridge, UK
[3] Protein Information Resource, Georgetown University, Washington DC & University of Delaware, USA

Schema.org

Is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond.

Why?

We hope to make our data more findable in search engines. It also seems to improve relevance of snippets in search results:



Searching for "amyloid"

APP - Amyloid-beta A4 protein precursor - Homo sapiens (Human ...
https://www.uniprot.org/uniprot/P05067
Induces a AGER-dependent pathway that involves activation of p38 MAPK, resulting in internalization of **amyloid**-beta peptide and leading to mitochondrial ...

Searching for "app"

APP - Amyloid-beta A4 protein precursor - Homo sapiens (Human
https://www.uniprot.org/uniprot/P05067
Interaction between **APP** molecules on neighboring cells promotes synaptogenesis (PubMed:25122912). Involved in cell mobility and transcription regulation ...

How: RDFa, JSON-LD or Microdata encoding

Search engine robots understand three encodings for Schema.org markup: RDFa and JSON-LD are RDF syntaxes, while Microdata was promoted by the WhatWG ?

JSON-LD

```
{
  "@context" :
     [ "schema.org",
       { "uniprot": "http://purl.uniprot.org/uniprot" }
     ],
  "id" : "uniprot:P05067",
  "name" : "Amyloid-beta A4 protein"
}
```

RDFa

```
<main
  vocab="https://schema.org"
  prefix="uniprot: http://purl.uniprot.org/uniprot/"
  property="mainEntity"
  resource="uniprot:P05067"
  <h1 property="name">
     Amyloid-beta A4 protein
  </h1>
</main>
```

Microdata

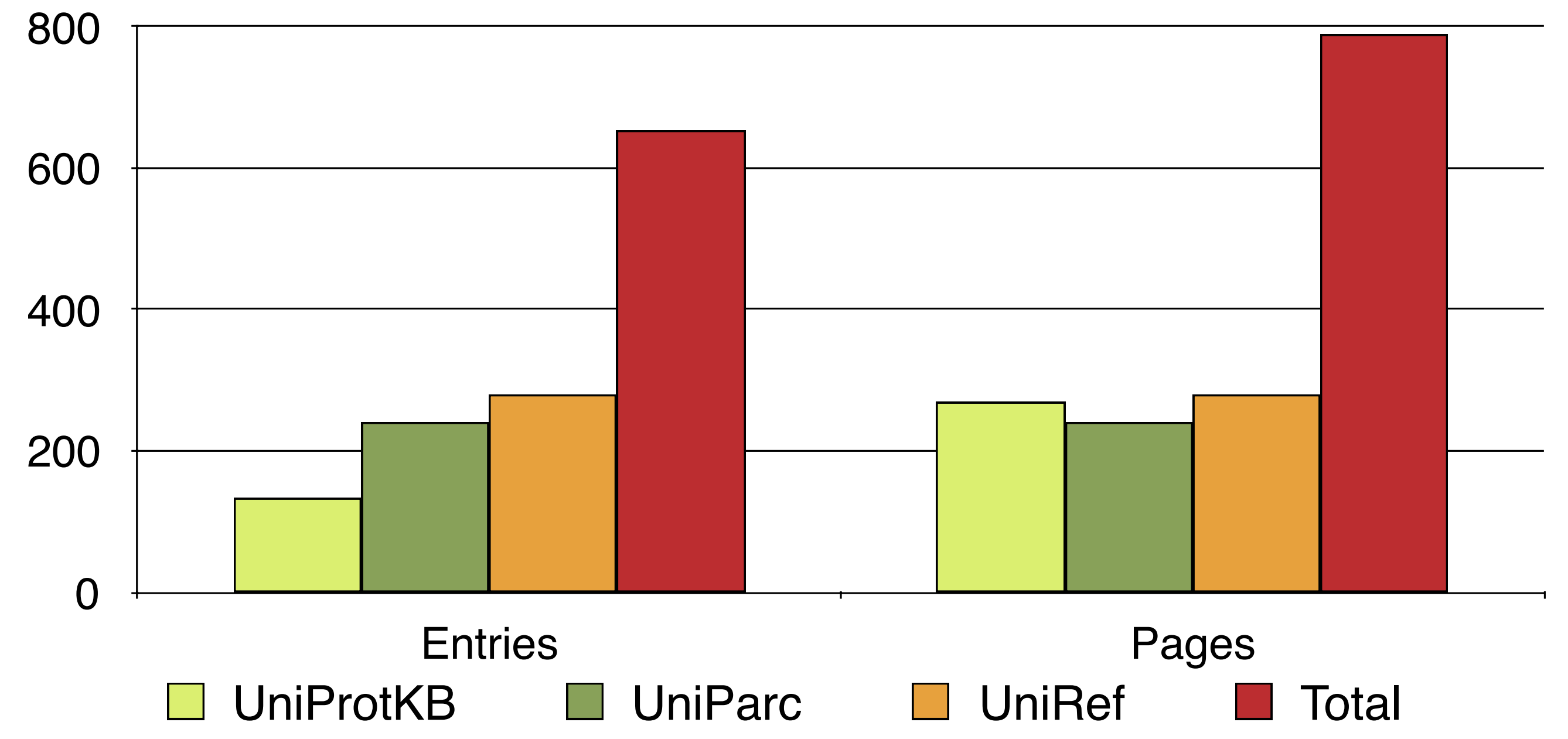
```
<main
  itemscope
  itemprop="mainEntity"
  itemid="http://purl.uniprot.org/uniprot/P05067">
  <h1 itemprop="name">
     Amyloid-beta A4 protein
  </h1>
</main>
```

UniProt on the web

UniProt is a comprehensive resource for protein sequence and annotation data that has been available on the web since its creation in 2002 (and its predecessors Swiss-Prot and TrEMBL much longer…).

The UniProt website is ranked by Alexa in the world's top 13,000 websites and serves nearly a million users per month.

12,361 Alexa



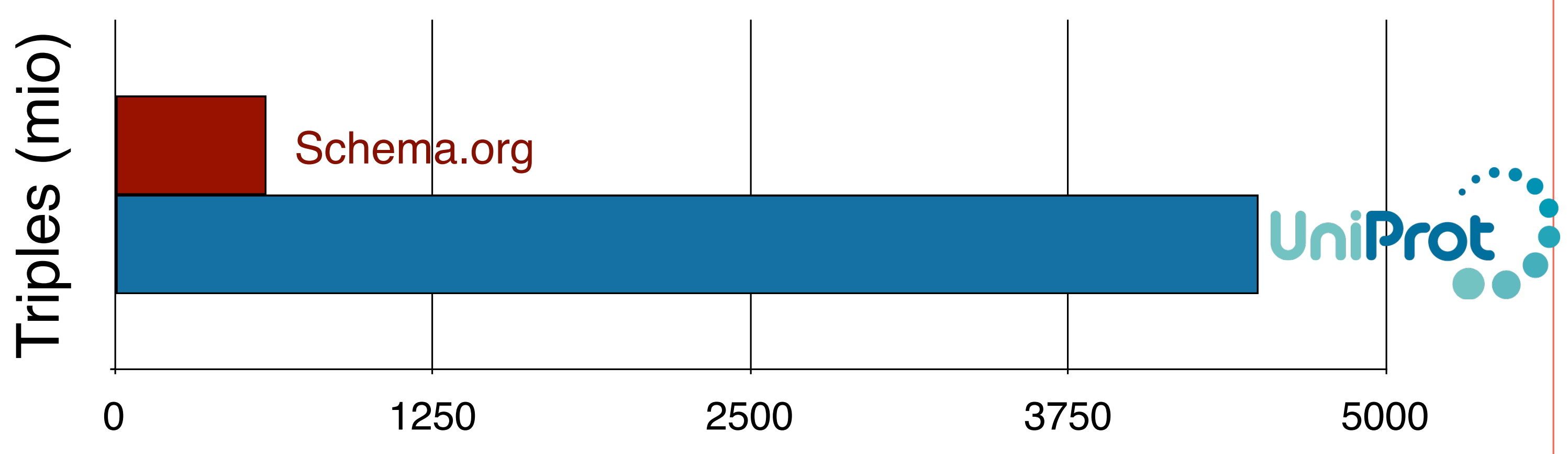
The UniProt website is trending towards a billion HTML pages for the entries of its three core databases.

RDFa or JSON-LD?

Search engine robots only crawl a limited number of pages (around half a million) per month and site. We thus aim for only one crawl per page and must use inline RDFa or JSON-LD. A JSON-LD encoding would nearly double our page sizes because many literals would be repeated, once for the human website visitors and once for the robots. Increasing the page size increases the rendering time and negatively impacts our users, especially on older and mobile devices. We therefore decided to use RDFa for UniProt.org.

Schema.org or UniProt RDF?

All the UniProt data is available in a custom RDF schema and can be queried on our public SPARQL endpoint at sparql.uniprot.org. The Schema.org vocabulary is very generic and allows us to describe only a small part of the UniProt data. Specific biological concepts are marked-up as the closest possible Schema.org thing, but this incurs a significant loss of precision.



Our community (**Bioschema**) is improving the modelling with a proposed extension to schema.org.

Tools that try to extract the markup would take 93 days to crawl all of UniProt.org (assuming 10ms per page) and download about 15 Terabytes of HTML. This would cost app. USD 1,500 in bandwidth cost if uniprot.org was running on Google Cloud.

UniProt is supported by several institutes of the National Institutes of Health (NIH) under grant U24HG007822. Additional support comes mainly from the Swiss State Secretariat for Education, Research and Innovation (SERI), the European Molecular Biology Laboratory (EMBL) and the National Science Foundation (NSF)

Contact
help@uniprot.org
www.uniprot.org

See also:
www.sib.swiss