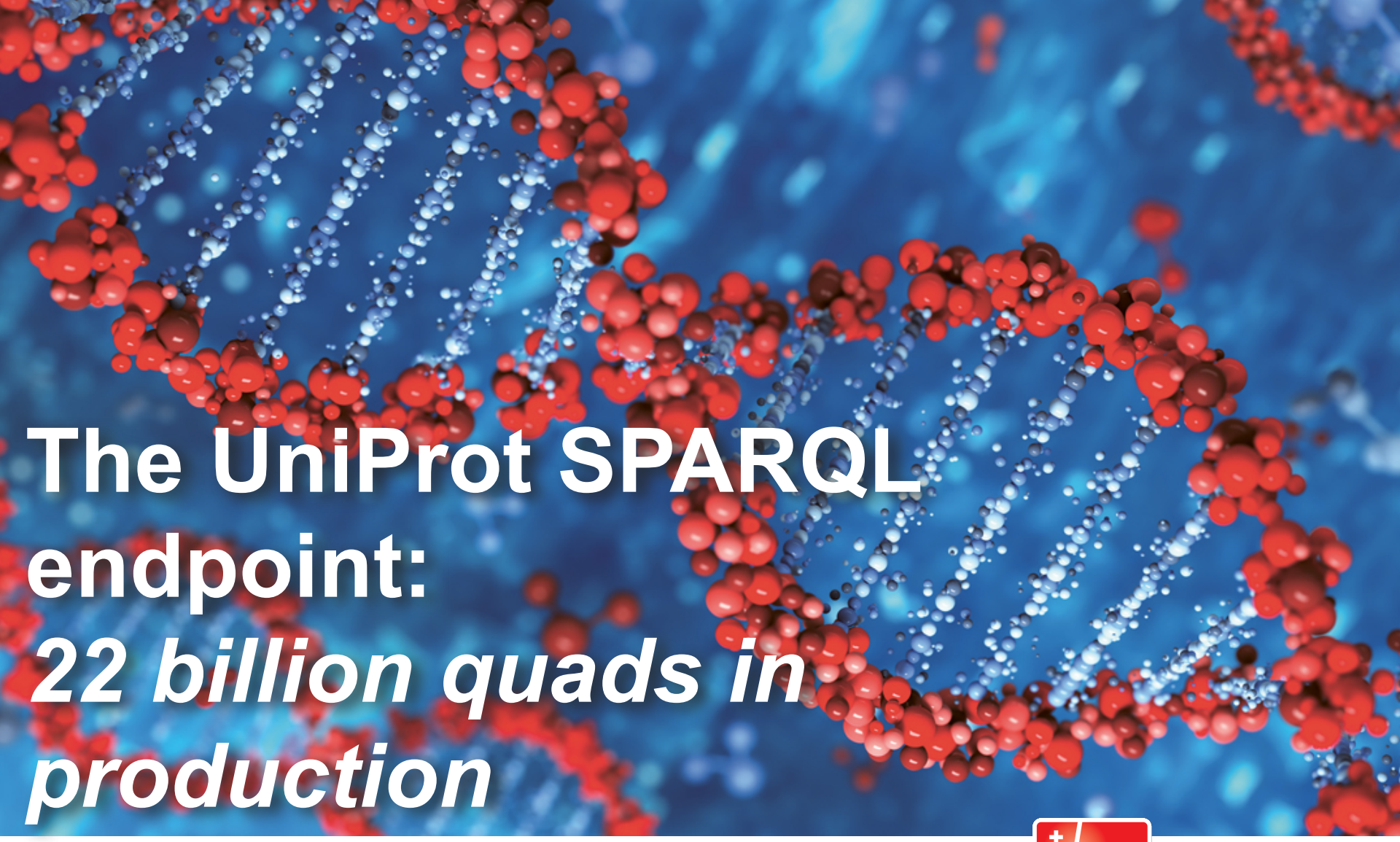


The UniProt SPARQL endpoint: *22 billion quads in production*

Jerven Bolleman
Lead Software Developer
Swiss-Prot Group

SIB
Swiss Institute of Bioinformatics

elixir SWITZERLAND

Recap RDF and SPARQL

Ambiguity in data

Recap RDF and SPARQL

Ambiguity in data

Novel questions

Recap RDF and SPARQL

Ambiguity in data

Novel questions

Variety of information

Recap RDF and SPARQL

Ambiguity in data

Novel questions

Variety of information

Public access

Recap RDF and SPARQL

Ambiguity in data

Novel questions

Variety of information

Public access

Community knows more

Recap RDF and SPARQL

Social solutions, not technical ones

Your SPARQL query

Add common prefixes

```
1
```

Submit Query

About

This SPARQL endpoint contains all UniProt data. It is free to access and supports the SPARQL 1.1 Standard.

There are 21,915,270,889 triples is this release (2016_05). The query timeout is 45 minutes. All triples are available in the default graph. There are 17 named graphs.

Documentation

The documentation about UniProt RDF is spread into 2 parts

1. Classes and predicates defined by the UniProt consortium
2. Statistics and diagrams

News

Forthcoming changes
Planned changes for UniProt

Slow/White and the 6 DWORFs | Cross-references to SIGNOR | Changes to the controlled vocabulary of human diseases
UniProt release 2016_05

Small changes, big effects | Changes to the controlled vocabulary of human diseases | New UniProt JAPI
UniProt release 2016_04

From the Zika forest to the Amazon, news from a viral wanderer | Cross-references to EPD | Cross-references to TopDownProteomics
UniProt release 2016_03

News archive

Examples

1. Select all taxa from the UniProt taxonomy: (show)
2. Select all bacterial taxa, and their scientific name, from the UniProt taxonomy: (show)
3. Select all E-Coli K12 (including strains) UniProt entries and their amino acid sequence: (show)
4. Select the UniProt entry with the mnemonic 'A4_HUMAN': (show)
5. Select a mapping of UniProt to PDB entries using the UniProt cross-references to the PDB database: (show)
6. Select all cross-references to external databases of the category '3D structure databases' of UniProt entries that are classified with the keyword '3Fe-4S': (show)
7. Select all UniProt entries, and their recommended protein name, that have a preferred gene name that contains the text 'DNA': (show)
8. Select the preferred gene name and disease annotation of all human UniProt entries that are known to be involved in a disease: (show)
9. Select all human UniProt entries with a sequence variant that leads to a 'loss of function': (show)
10. Select all human UniProt entries with a sequence variant that leads to a tyrosine to phenylalanine substitution: (show)
11. Select all UniProt entries with annotated transmembrane regions and the regions' begin and end coordinates on the canonical sequence: (show)
12. Select all UniProt entries that were integrated on the 30th of November 2010: (show)
13. Was any UniProt entry integrated on the 9th of January 2013? (show)
14. Construct new triples of the type 'HumanProtein' from all human UniProt entries: (show)
15. Select all triples that relate to the EMBL CDS entry AA089367.1: (show)
16. Select all triples that relate to the taxon that describes *Homo sapiens*: (show)
17. Select the average number of cross-references to the PDB database of UniProt entries that have at least one cross-reference to the PDB database: (show)

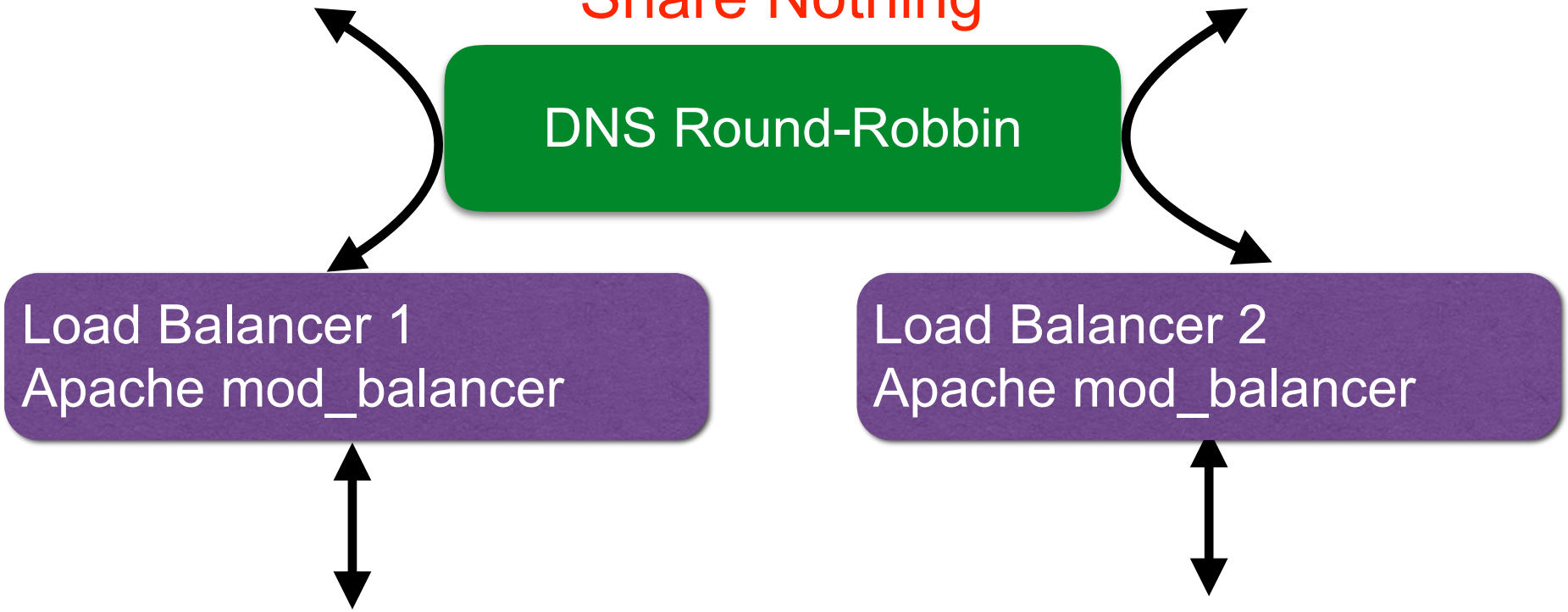
Dedicated machine for loading and testing

- Loading RDF data “solved” problem
  - 1,500,000 triples per second
  - no full text index
- 400+ RDF files for you on FTP
  - Check:
    - void.rdf
    - RELEASE.meta files

Uptime/SLA

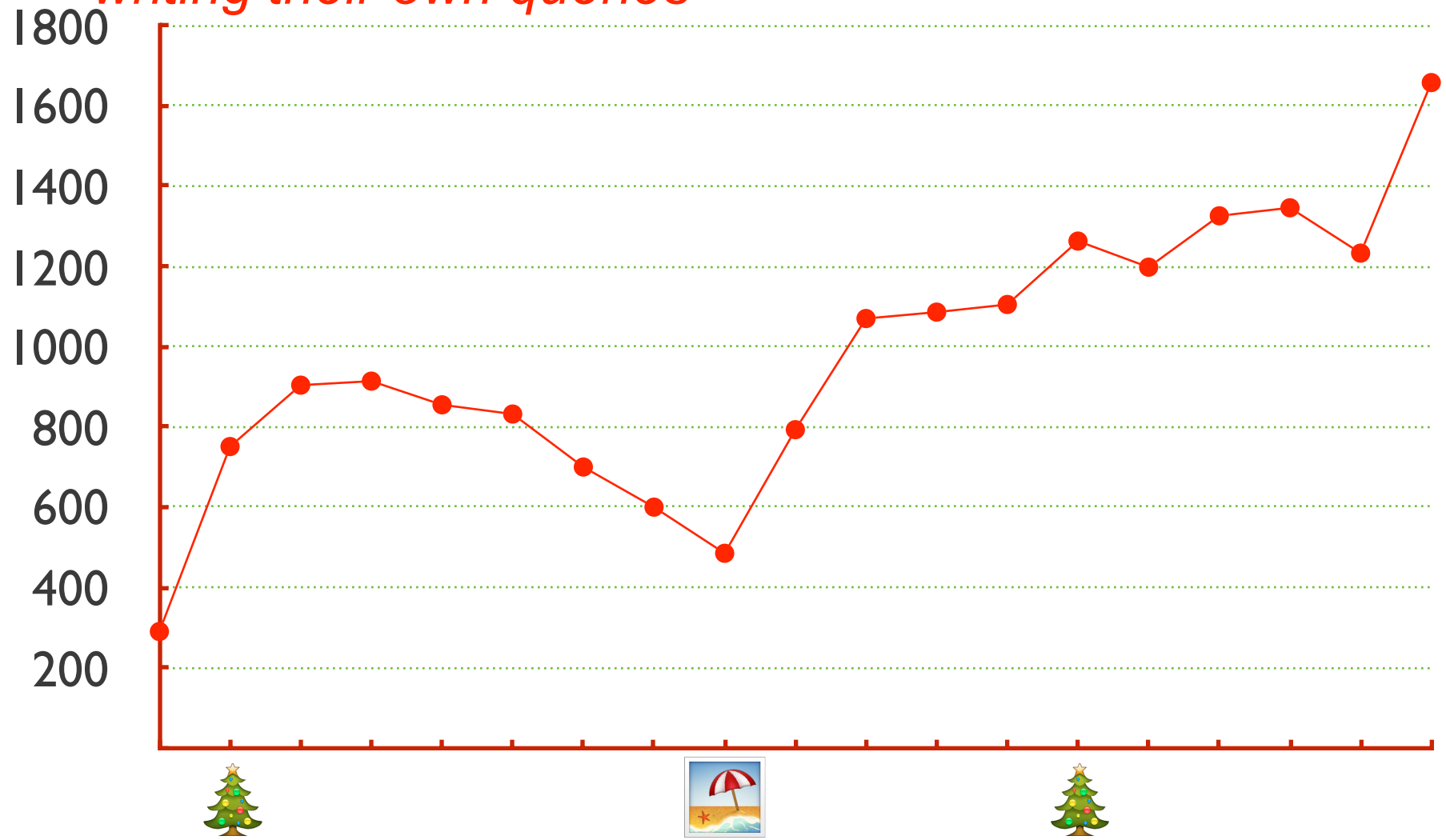
- Best effort
  - hey it's free
  - 99.5% goal
- Challenges
  - Pile-up of long running queries
  - HTTP connection instability
  - Semantic web researchers
  - Bugs in the implementation

Share Nothing



Users

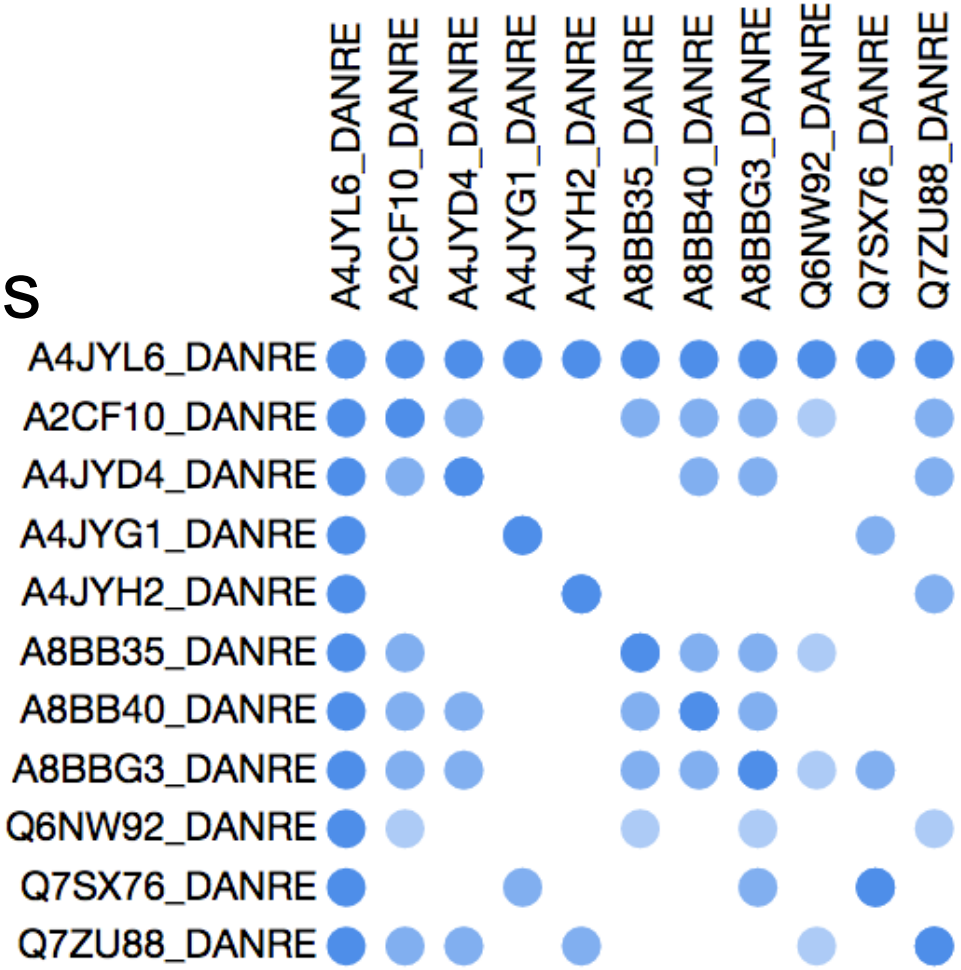
writing their own queries



Enabling new visualisations

2 Level Protein-Protein interaction

- www.uniprot.org
  - entry focused
- demand for new vis
  - major work on server side
- SPARQL
  - ✓ fast
  - ✓ maintained
  - ✓ no new dev



Differences from Benchmarks

- Query load unpredictable
  - Peak waves
  - Long running ones
  - It never stops
- Breaking point key knowledge

Queries are bigger

- 4000 chars is not enough
- 32000 chars is not enough
- Computation is part of the query
  - custom functions key part of the solution

Biology is complicated

```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX uniprot:<http://purl.uniprot.org/uniprot/>
PREFIX sequence:<http://purl.uniprot.org/sequences/>
PREFIX unirule:<http://purl.uniprot.org/unirules/>
PREFIX taxon:<http://purl.uniprot.org/taxonomy/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX hamap-sparql:<http://example.org/hamap_sparql/>
PREFIX up:<http://purl.uniprot.org/core/>
PREFIX faldo:<http://biohackathon.org/resource/faldo#>
PREFIX method:<http://example.org/method/>
PREFIX keyword:<http://purl.uniprot.org/keywords/>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
PREFIX proteome:<http://purl.uniprot.org/proteomes/>
PREFIX hamap:<http://purl.uniprot.org/hamap/>
PREFIX annotation:<http://purl.uniprot.org/annotation/>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
CONSTRUCT {
    ?this up:annotation ?annotation0,
        ?annotation1,
        ?annotation2,
        ?annotation3,
        ?annotation5;
      up:classifiedWith <http://purl.obolibrary.org/obo/19805>,
        <http://purl.obolibrary.org/obo/334>,
        <http://purl.obolibrary.org/obo/34354>,
        <http://purl.obolibrary.org/obo/43420>,
        <http://purl.obolibrary.org/obo/6569>,
        <http://purl.obolibrary.org/obo/8198>,
        keyword:223,
        keyword:560,
        keyword:662 .
   ?annotation0 a up:Function_Annotation;
      rdfs:comment "Catalyzes the oxidative ring opening of 3-hydroxyanthranilate to 2-amino-3-carboxymuconate semialdehyde, which spontaneously cyclizes to quinolinate." .
   ?pfs1t0 up:annotation ?annotation17 .
   ?pfs2t0 up:annotation ?annotation27 .
```

Differences from Benchmarks

- Data model larger
  - 170 classes in UniProt
    - 12 SNB
  - 139 predicates
    - 15 SNB

Differences from Benchmarks

- Data model changes
  - 2005 First benchmarking using UniProt
    - 80 million triples
  - 2016 …
    - 22 billion triples

The Team

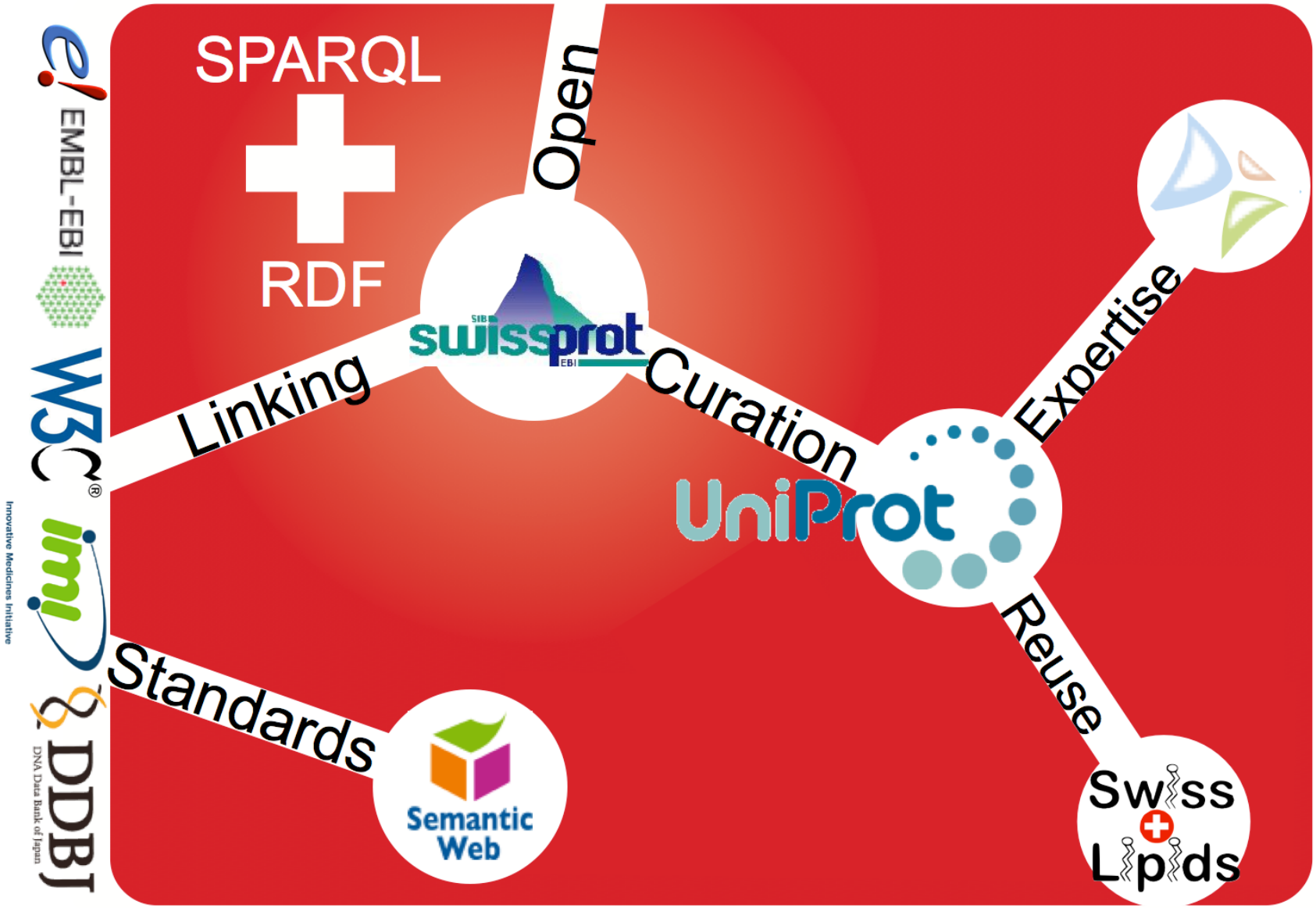
PIs: Alex Bateman, Cathy Wu, Ioannis Xenarios

Key staff: Cecilia Arighi (Curation), Lydie Bougueleret (Co-Direction), Alan Bridge (Content), Hongzhan Huang (Development), Michele Magrane (Curation), Maria Martin (Development), Peter McGarvey (Content), Darren Natale (Content), Claire O'Donovan (Content), Sylvain Poux (Curation), Manuela Pruess (Coordination), Nicole Redaschi (Development)

Content/Curation: Lucila Aimo, Ghislaine Argoud-Puy, Andrea Auchincloss, Kristian Axelsen, Brigitte Boeckmann, Emmanuel Boutet, Lionel Breuza, Ramona Britto, Hema Bye-A-Jee, Cristina Casals Casas, Elisabeth Coudert, Melanie Courtot, Anne Estreicher, Livia Famiglietti, Marc Feuermann, John S. Garavelli, Penelope Garmiri, Daniel Gonzalez, Arnaud Gos, Nadine Gruaz, Emma Hatton-Ellis, Ursula Hinz, Alex Holmes, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Guillaume Keller, Kati Laiho, Philippe Lemercier, Damien Lieberherr, Alistair MacDougall, Patrick Masson, Anne Morgat, Barbara Palka, Ivo Pedruzzi, Klemens Pichler, Sandrine Pilbout, Catherine Rivoire, Bernd Roechert, Karen Ross, Michel Schneider, Aleksandra Shypitsyna, Christian Sigrist, Elena Speretta, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Nidhi Tyagi, C. R. Vinayaka, Qinghua Wang, Kate Warner, Lai-Su Yeh, Rosanna Zaru

Development: Emanuele Alpi, Ricardo Antunes, Leslie Arminski, Parit Bansal, Delphine Baratin, Teresa Batista Neto, Benoit Bely, Mark Bingley, Jerven Bolleman, Borisas Bursteinas, Chuming Chen, Yongxing Chen, Beatrice Cuche, Alan Da Silva, Edouard De Castro, Maurizio De Giorgi, Tunca Dogan, Leyla Garcia Castro, Elisabeth Gasteiger, Sebastien Gehant, Arnaud Kerhornou, Vicente Lara, Wudong Liu, Thierry Lombardot, Jie Luo, Xavier Martin, Andrew Nightingale, Joseph Onwubiko, Diego Poggioli, Monica Pozzato, Sangya Pundir, Guoying Qi, Alexandre Renaux, Steven Rosanoff, Rabie Saidi, Tony Sawford, Edward Turner, Vladimir Volynkin, Yuqi Wang, Tony Wardell, Xavier Watkins, Hermann Zellner, Jian Zhang

European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, UK
Protein Information Resource (PIR), Washington DC and Delaware, USA
SIB Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland


SPARQL
RDF
Open
Linking
Curation
Expertise
Reuse
Standards
swissprot
UniProt
Semantic Web
Swiss Lipids
EMBL-EBI
W3C
imi
Innovative Medicines Initiative
DDBJ
DNA Data Bank of Japan