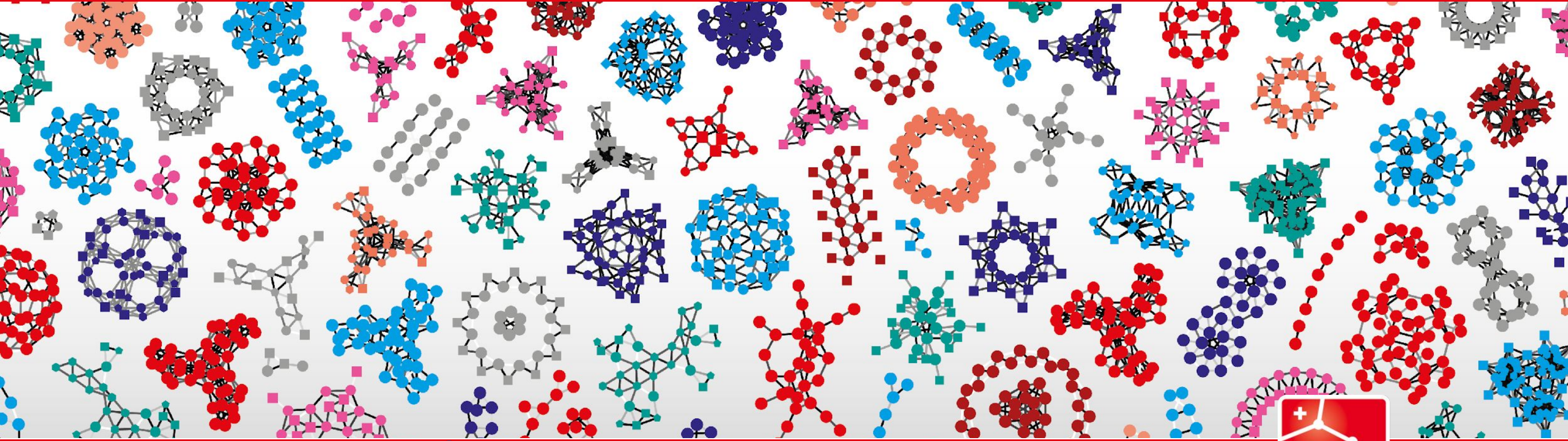


SIB Swiss Institute of Bioinformatics

UniProt 100 billion+ triple KG

PIR

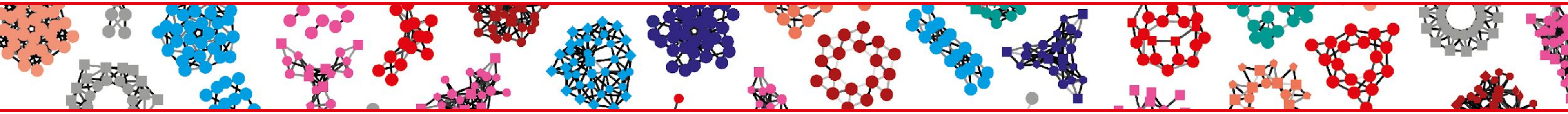
EMBL-EBI

Jerven Bolleman

Principal Software Engineer – Swiss-Prot group

www.sib.swiss

Overview



01 Background

02 Why RDF ?

03 How to be FAIR

04 Challenges

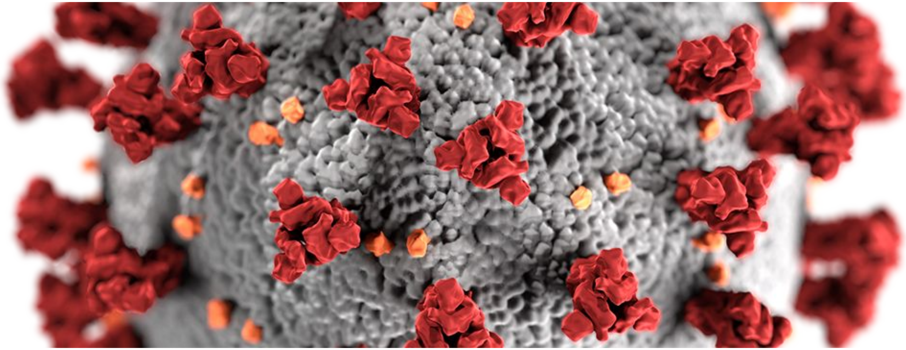
05 Questions (and hopefully answers)

Background

Proteins and a lot of what we know about them

Most of the parts of what makes us what we are.

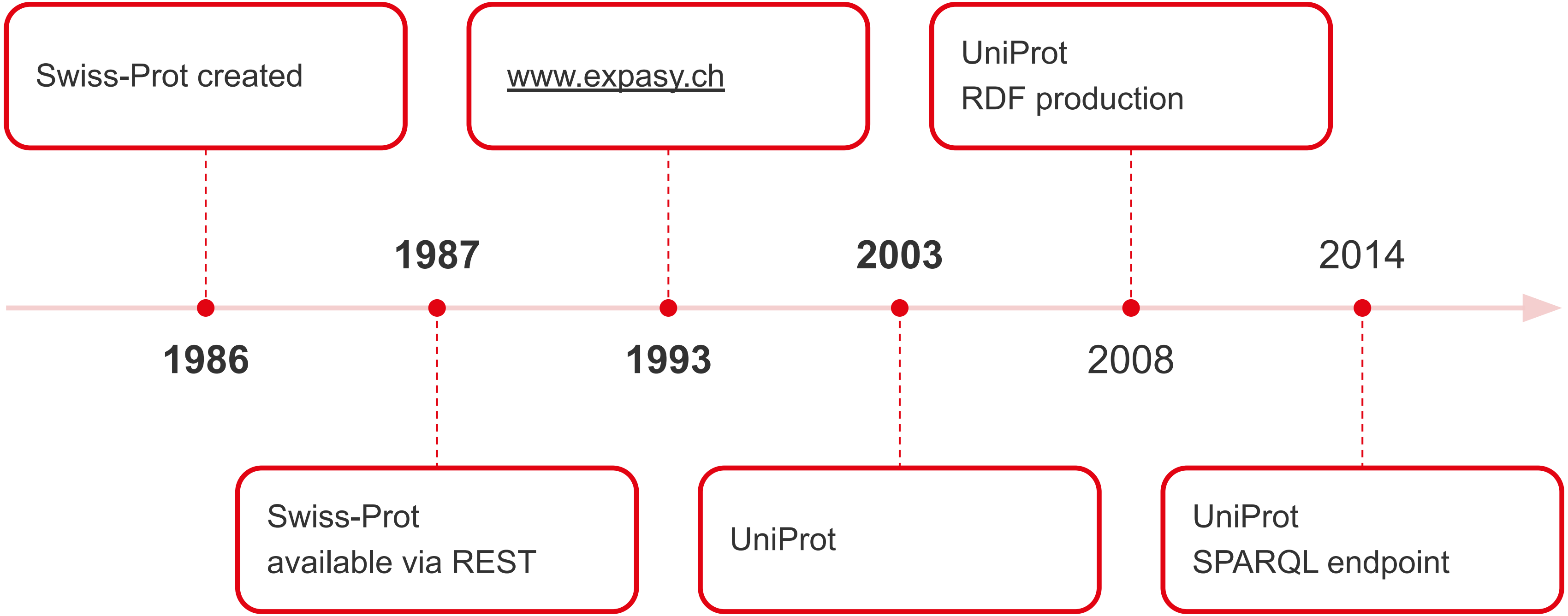
Kind of important ...



UniProt Knowledge?

- UniProtKB/Swiss-Prot (500,000)
  - Expert curated (AI accelerated)
  - Scientific literature
- UniProtKB/TrEMBL (50 billion)
  - Automated from DNA databases
  - AI annotated (learned on Swiss-Prot)
- UniRef (7 billion)
- UniParc (50 billion)
  - Archive

Background/History



1986

Swiss-Prot created

1987

Swiss-Prot
available via REST

1993

www.expasy.ch

2003

UniProt

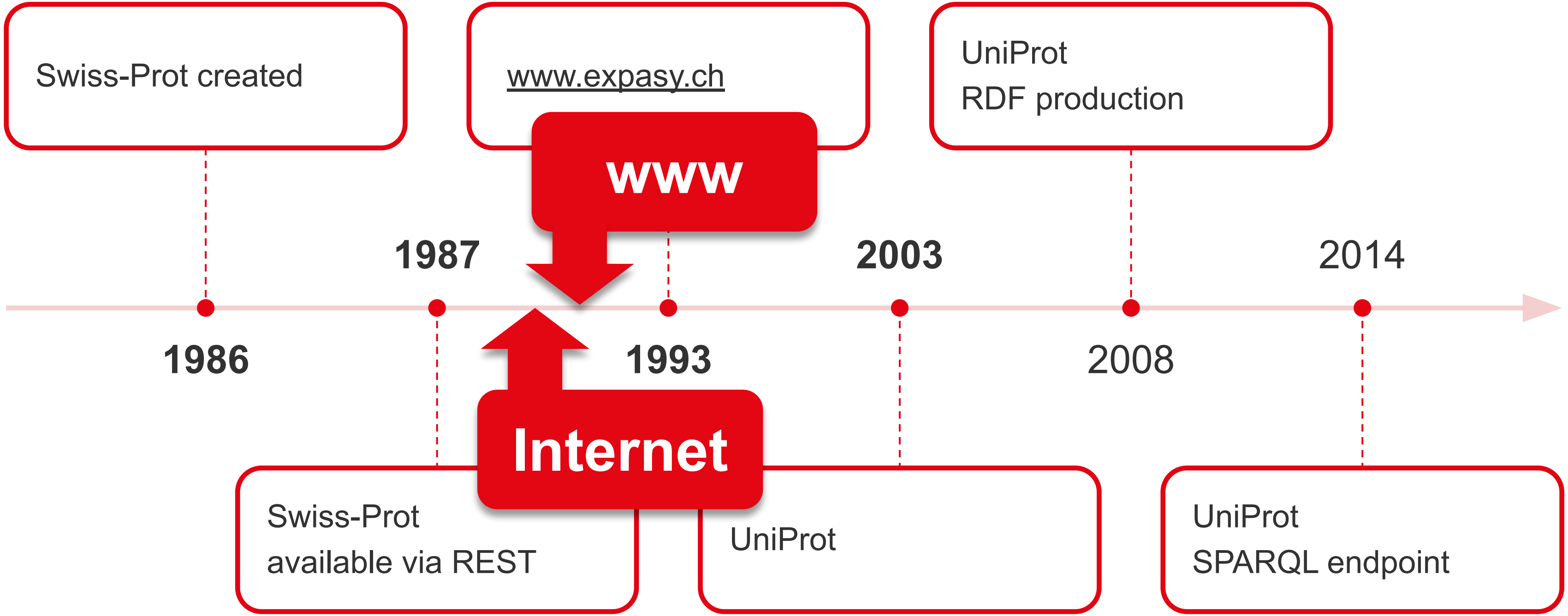
2008

UniProt
RDF production

2014

UniProt
SPARQL endpoint

Background/History



1986 Swiss-Prot created

1987 Swiss-Prot available via REST

Internet

WWW

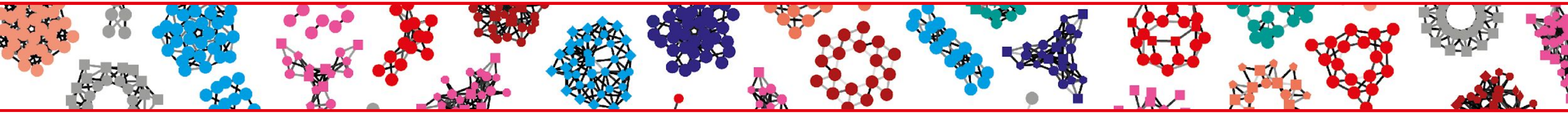
1993 www.expasy.ch

2003 UniProt

2008 UniProt RDF production

2014 UniProt SPARQL endpoint

Overview



01 Background

02 Why RDF ?

03 How to be FAIR

04 Challenges

05 Questions (and hopefully answers)

What is 9993 ?

Coupling of the Penicillium Duponti Acid Protease to Ethylene-Maleic Acid (1 : 1) Linear Copolymer. Preparation and Properties of the Water-Soluble Derivative

PubMed

NCBI
Taxonomy

DGCR2 DiGeorge syndrome critical region gene 2 [Homo sapiens (human)]

NCBI Gene

Is why we use IRIs

Coupling of the Penicillium Duponti Acid Protease to Ethylene-Maleic Acid (1 : 1) Linear Copolymer. Preparation and Properties of the Water-Soluble Derivative

https://pubmed.ncbi.nlm.nih.gov/9993/

PubMed

I am 9993,
would you
like a flower

http://identifiers.org/taxonomy/9993

NCBI
Taxonomy

DGCR2 DiGeorge syndrome critical region gene 2 [Homo sapiens (human)]

http://identifiers.org/ncbigene/9993

NCBI Gene

RDF not just another standard

UniProt is not exhaustive

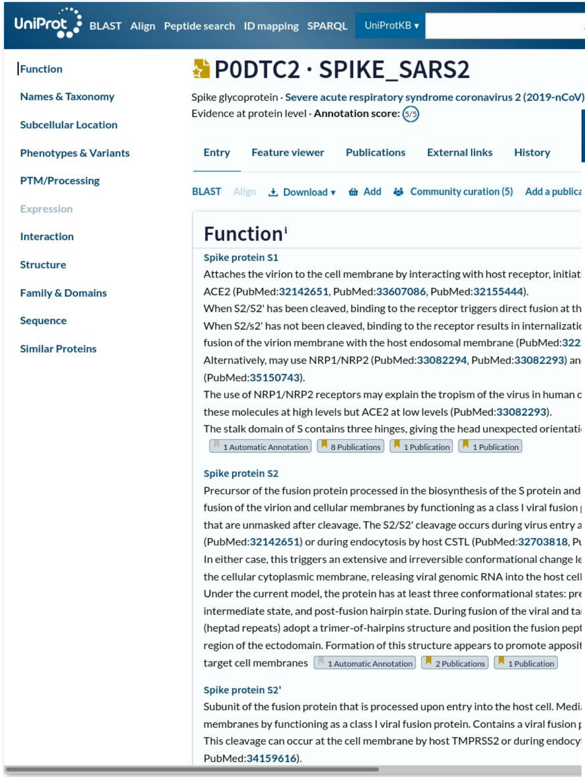
- Central point in the life sciences
- Linking to other resources

Trees have just one root (XML, JSON, Protocol Buffer etc)

Ecosystems are bigger than a single organism

Can't dictate vendors on our users

Links



P0DTC2 · SPIKE_SARS2

Spike glycoprotein · Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) (SARS-CoV-2) · Gene: S · 1273 amino acids · Evidence at protein level · Annotation score: 5/5

Function

Spike protein S1

Attaches the virion to the cell membrane by interacting with host receptor, initiat ACE2 (PubMed:32142651, PubMed:33607086, PubMed:32155444).

When S2/S2' has been cleaved, binding to the receptor triggers direct fusion at th When S2/s2' has not been cleaved, binding to the receptor results in internalizatic fusion of the virion membrane with the host endosomal membrane (PubMed:322 Alternatively, may use NRP1/NRP2 (PubMed:33082294, PubMed:33082293) an (PubMed:35150743).

The use of NRP1/NRP2 receptors may explain the tropism of the virus in human c these molecules at high levels but ACE2 at low levels (PubMed:33082293).

The stalk domain of S contains three hinges, giving the head unexpected orientati

1 Automatic Annotation | 8 Publications | 1 Publication | 1 Publication

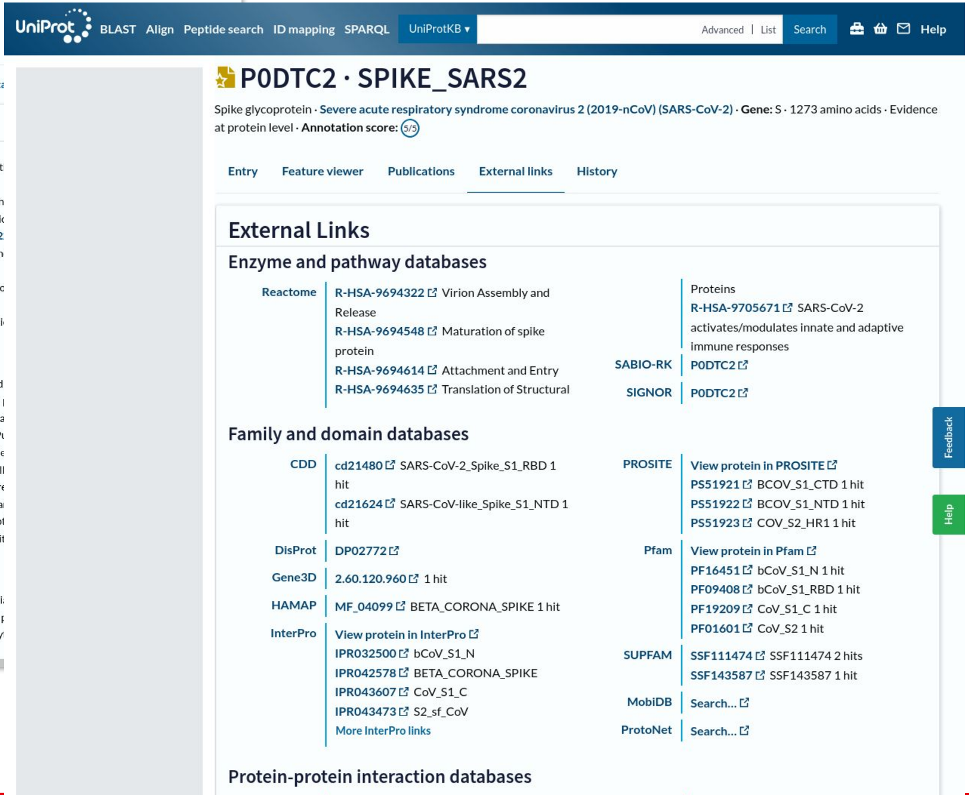
Spike protein S2

Precursor of the fusion protein processed in the biosynthesis of the S protein and fusion of the virion and cellular membranes by functioning as a class I viral fusion that are unmasked after cleavage. The S2/S2' cleavage occurs during virus entry a (PubMed:32142651) or during endocytosis by host CSTL (PubMed:32703818, Pu In either case, this triggers an extensive and irreversible conformational change le the cellular cytoplasmic membrane, releasing viral genomic RNA into the host cell Under the current model, the protein has at least three conformational states: pre intermediate state, and post-fusion hairpin state. During fusion of the viral and ta (heptad repeats) adopt a trimer-of-hairpins structure and position the fusion pept region of the ectodomain. Formation of this structure appears to promote apposit target cell membranes

1 Automatic Annotation | 2 Publications | 1 Publication

Spike protein S2'

Subunit of the fusion protein that is processed upon entry into the host cell. Medi membranes by functioning as a class I viral fusion protein. Contains a viral fusion This cleavage can occur at the cell membrane by host TMPRSS2 or during endocy PubMed:34159616).



P0DTC2 · SPIKE_SARS2

Spike glycoprotein · Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) (SARS-CoV-2) · Gene: S · 1273 amino acids · Evidence at protein level · Annotation score: 5/5

External Links

Enzyme and pathway databases

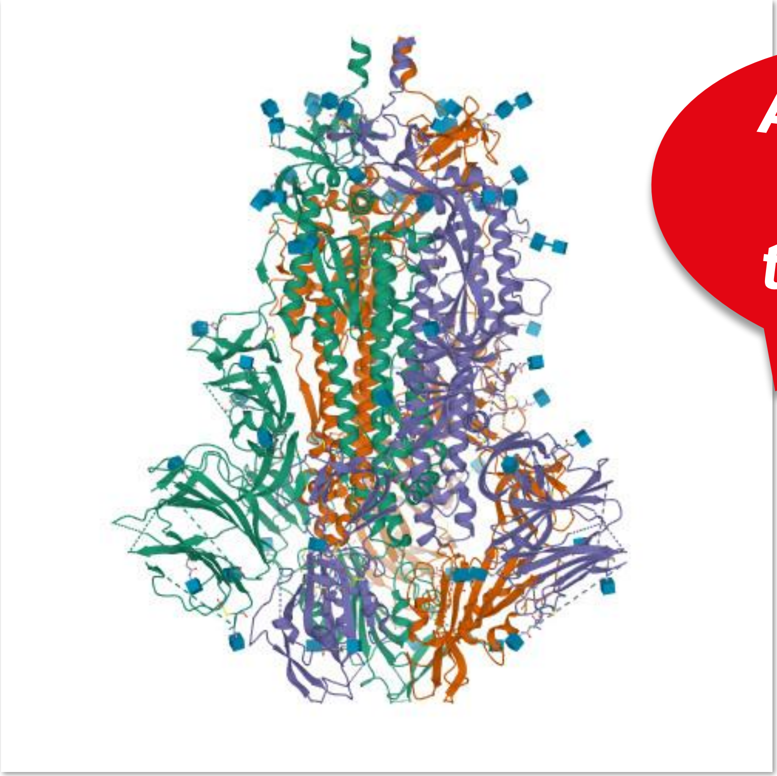
Reactome	R-HSA-9694322 Virion Assembly and Release R-HSA-9694548 Maturation of spike protein R-HSA-9694614 Attachment and Entry R-HSA-9694635 Translation of Structural Proteins R-HSA-9705671 SARS-CoV-2 activates/modulates innate and adaptive immune responses
SABIO-RK	P0DTC2
SIGNOR	P0DTC2

Family and domain databases

CDD	cd21480 SARS-CoV-2_Spike_S1_RBD 1 hit cd21624 SARS-CoV-like_Spike_S1_NTD 1 hit
DisProt	DP02772
Gene3D	2.60.120.960 1 hit
HAMAP	MF_04099 BETA_CORONA_SPIKE 1 hit
InterPro	View protein in InterPro IPR032500 bCoV_S1_N IPR042578 BETA_CORONA_SPIKE IPR043607 CoV_S1_C IPR043473 S2_sf_CoV More InterPro links
PROSITE	View protein in PROSITE PS51921 BCOV_S1_CTD 1 hit PS51922 BCOV_S1_NTD 1 hit PS51923 COV_S2_HR1 1 hit
Pfam	View protein in Pfam PF16451 bCoV_S1_N 1 hit PF09408 bCoV_S1_RBD 1 hit PF19209 CoV_S1_C 1 hit PF01601 CoV_S2 1 hit
SUPFAM	SSF111474 SSF111474 2 hits SSF143587 SSF143587 1 hit
MobiDB	Search...
ProtoNet	Search...

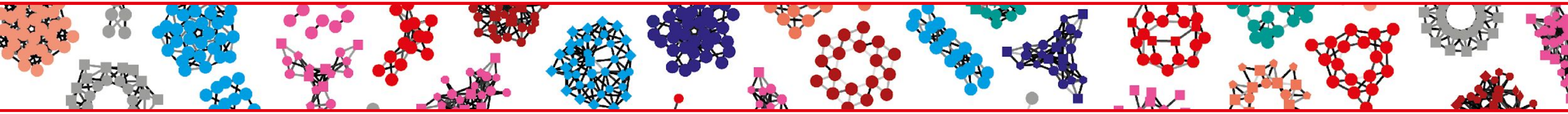
Protein-protein interaction databases

wwPDB "another" Protein database



AlphaFold is about this stuff

Overview



01 Background

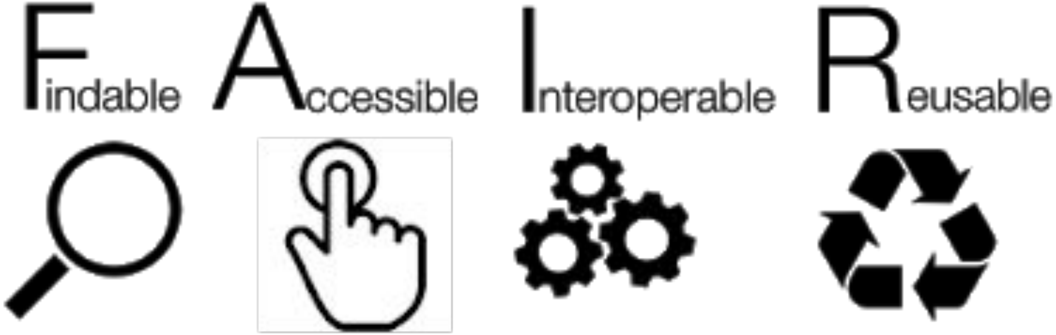
02 Why RDF ?

03 How to be FAIR

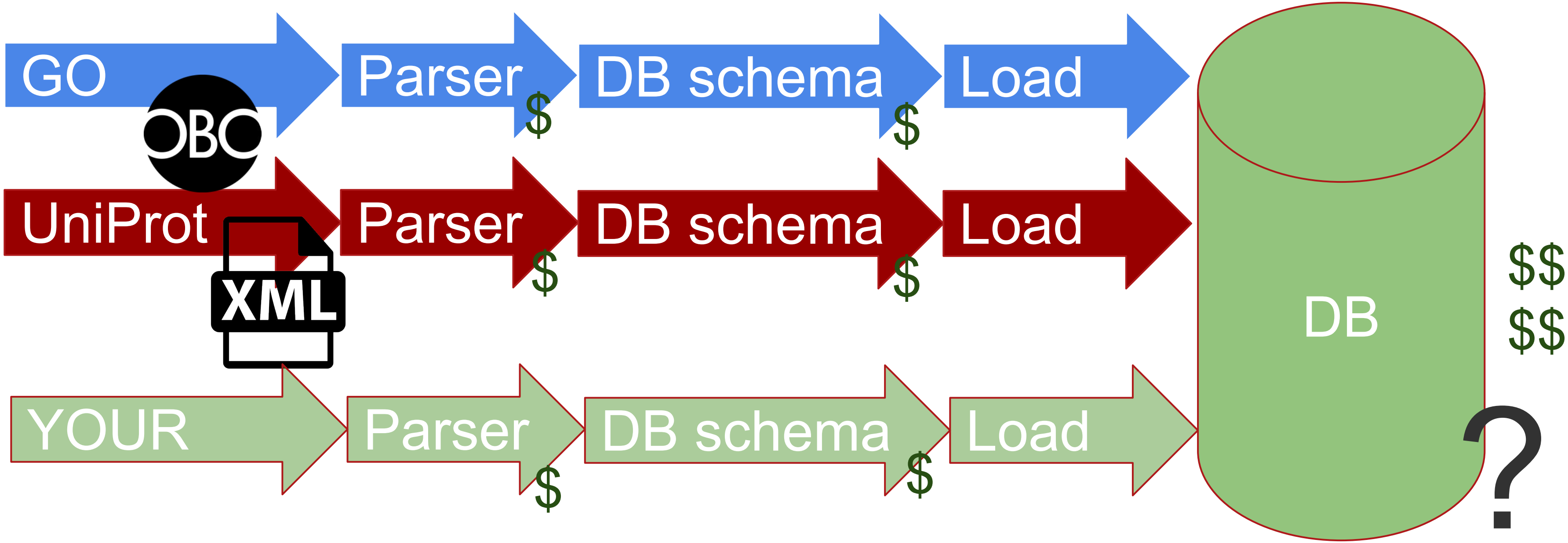
04 Challenges

05 Questions (and hopefully answers)

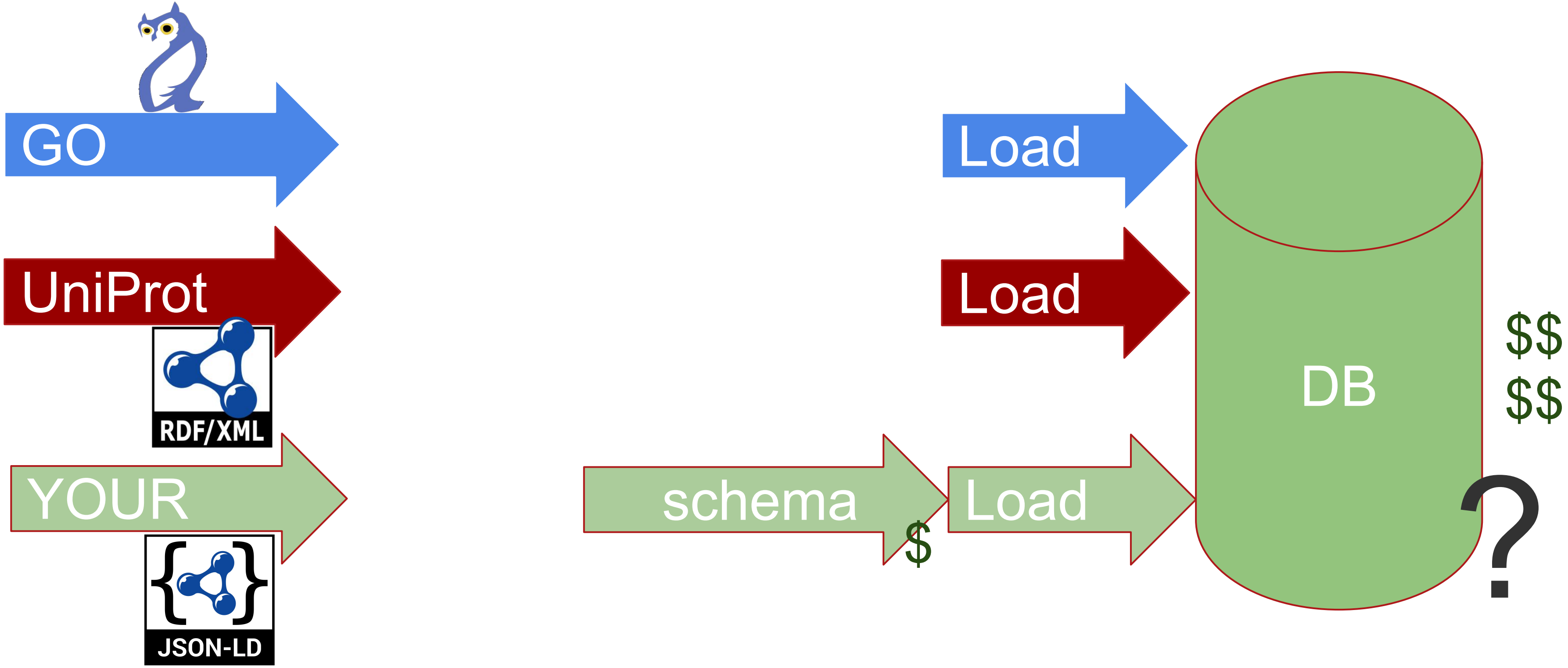
The FAIRest format of them all



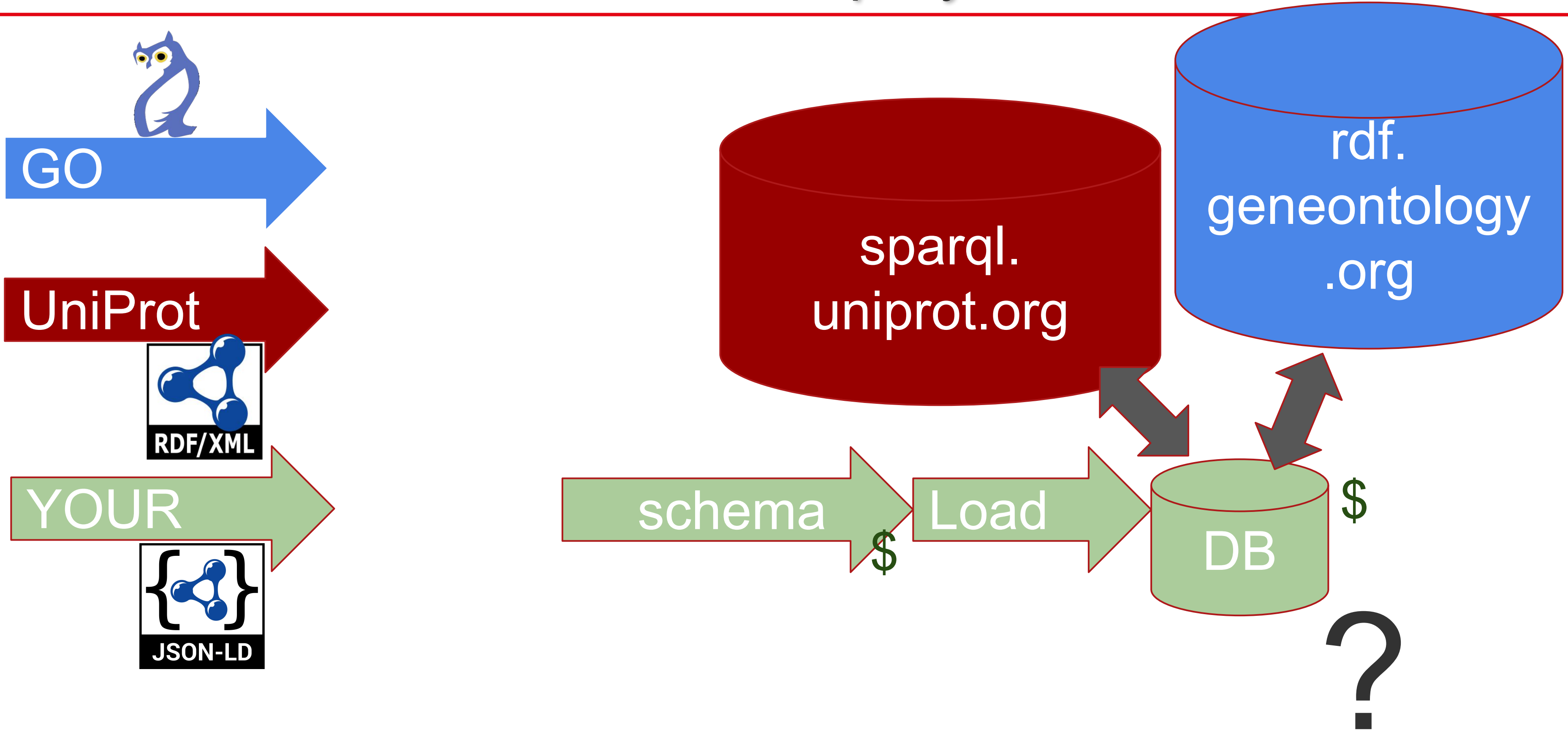
Parse, parse, parse



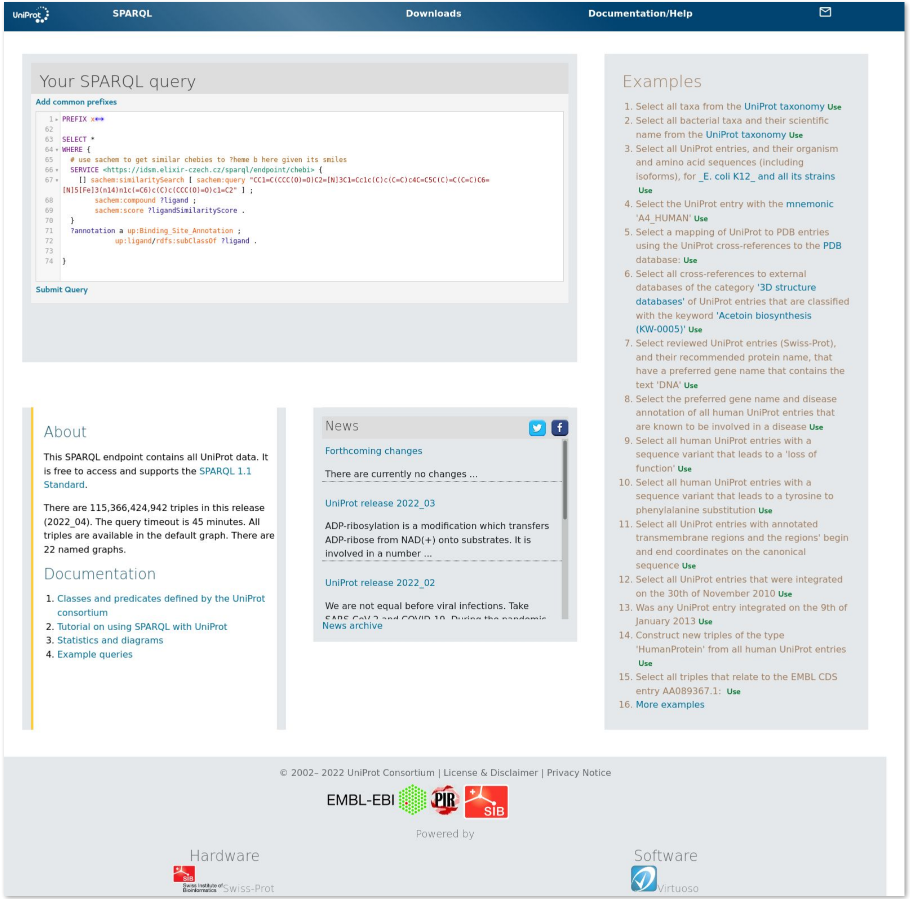
RDF



RDF + SPARQL 1.1 with build in query federation



https://sparql.uniprot.org



Your SPARQL query

Add common prefixes

```
PREFIX x↔

SELECT *
WHERE {
  # use sachem to get similar chebies to ?heme b here given its smiles
  SERVICE <https://idsm.elixir-czech.cz/sparql/endpoint/chebi> {
    [] sachem:similaritySearch [ sachem:query "CC1=C(CCC(O)=O)C2=[N]3C1=Cc1c(C)c(C=C)c4C=C5C(C)=C(C=C)C6=
[N]5[Fe]3(n14)n1c(=C6)c(C)c(CCC(O)=O)c1=C2" ] ;
        sachem:compound ?ligand ;
        sachem:score ?ligandSimilarityScore .
  }
  ?annotation a up:Binding_Site_Annotation ;
            up:ligand/rdfs:subClassOf ?ligand .

}
```

Submit Query

Examples

1. Select all taxa from the UniProt taxonomy Use
2. Select all bacterial taxa and their scientific name from the UniProt taxonomy Use
3. Select all UniProt entries, and their organism and amino acid sequences (including isoforms), for _E. coli K12_ and all its strains Use
4. Select the UniProt entry with the mnemonic 'A4_HUMAN' Use
5. Select a mapping of UniProt to PDB entries using the UniProt cross-references to the PDB database: Use
6. Select all cross-references to external databases of the category '3D structure databases' of UniProt entries that are classified with the keyword 'Acetoin biosynthesis (KW-0005)' Use
7. Select reviewed UniProt entries (Swiss-Prot), and their recommended protein name, that have a preferred gene name that contains the text 'DNA' Use
8. Select the preferred gene name and disease annotation of all human UniProt entries that are known to be involved in a disease Use
9. Select all human UniProt entries with a sequence variant that leads to a 'loss of function' Use
10. Select all human UniProt entries with a sequence variant that leads to a tyrosine to phenylalanine substitution Use
11. Select all UniProt entries with annotated transmembrane regions and the regions' begin and end coordinates on the canonical sequence Use
12. Select all UniProt entries that were integrated on the 30th of November 2010 Use
13. Was any UniProt entry integrated on the 9th of January 2013 Use
14. Construct new triples of the type 'HumanProtein' from all human UniProt entries Use
15. Select all triples that relate to the EMBL CDS entry AA089367.1: Use
16. More examples

About

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

There are 115,366,424,942 triples in this release (2022_04). The query timeout is 45 minutes. All triples are available in the default graph. There are 22 named graphs.

Documentation

1. Classes and predicates defined by the UniProt consortium
2. Tutorial on using SPARQL with UniProt
3. Statistics and diagrams
4. Example queries

News

Forthcoming changes

There are currently no changes ...

UniProt release 2022_03

ADP-ribosylation is a modification which transfers ADP-ribose from NAD(+) onto substrates. It is involved in a number ...

UniProt release 2022_02

We are not equal before viral infections. Take SARS CoV 2 and COVID 19. During the pandemic

News archive

© 2002– 2022 UniProt Consortium | License & Disclaimer | Privacy Notice

Powered by

Hardware

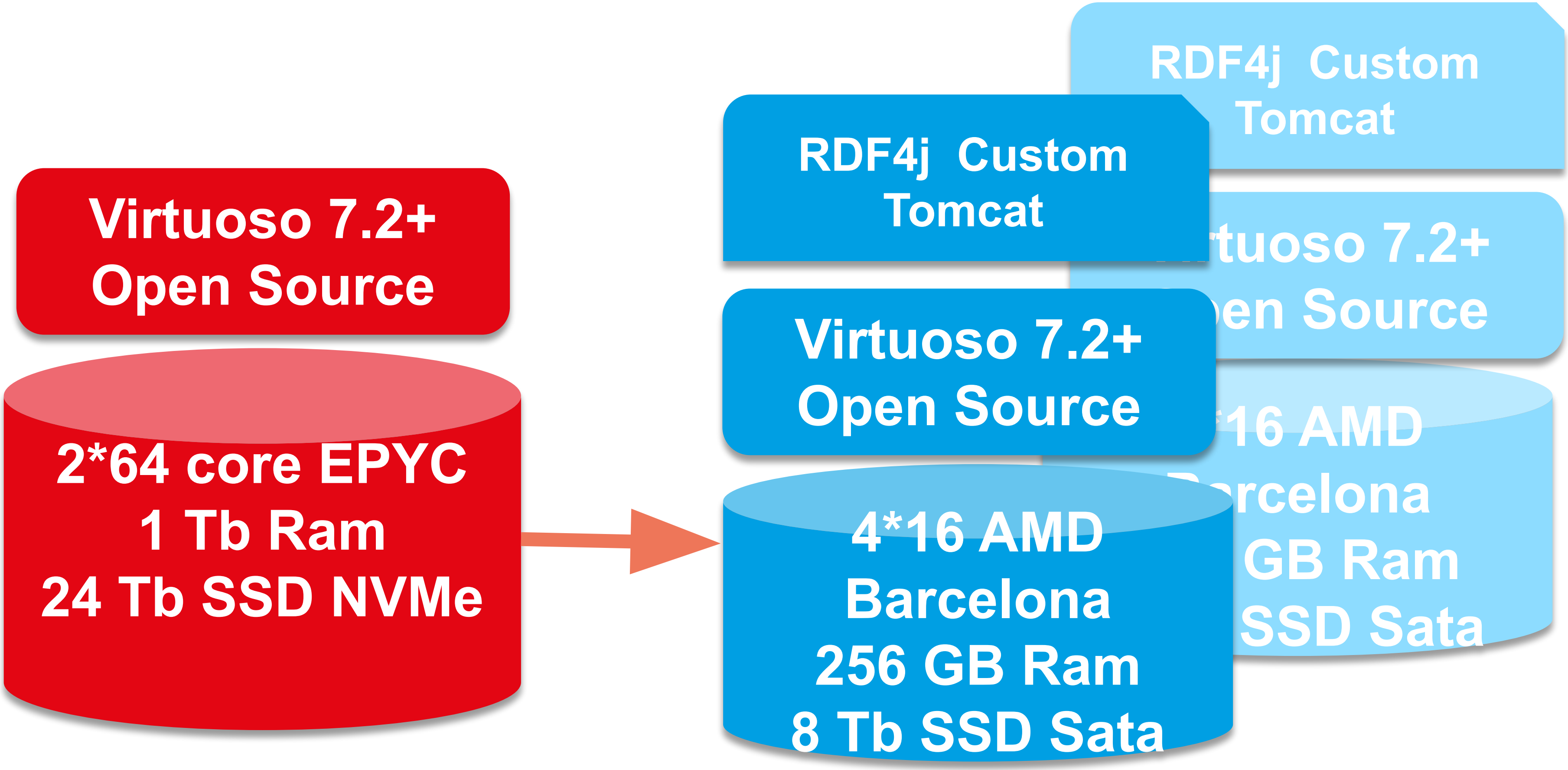
Swiss-Prot

Software

Virtuoso

Query
Examples
Documentation
Tutorials
Limits
Partners
News

Hardware & Software



Virtuoso 7.2+
Open Source

2*64 core EPYC
1 Tb Ram
24 Tb SSD NVMe

RDF4j Custom
Tomcat

Virtuoso 7.2+
Open Source

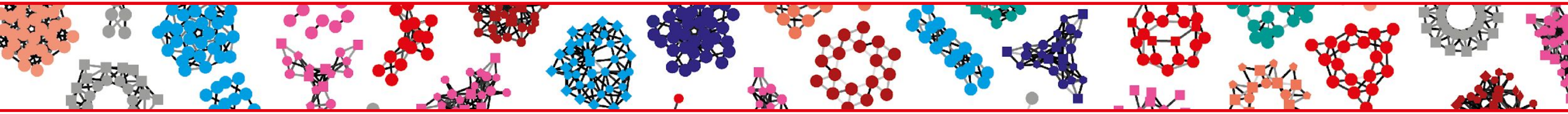
4*16 AMD
Barcelona
256 GB Ram
8 Tb SSD Sata

RDF4j Custom
Tomcat

rtuoso 7.2+
en Source

*16 AMD
rcelona
GB Ram
SSD Sata

Overview



01 Background

02 Why RDF ?

03 How to be FAIR

04 **Challenges**

05 Questions (and hopefully answers)

Complexity of schema

- Entity Relation Schema Diagram
  - Large
- Biology has many concepts
  - Simplicity is misleading
  - Compare search to analysis

Service description

- Drives documentation
- Used to rewrite hard queries
  - e.g. How many IRIs?
  - Taking care of variable and prefix use
- Note funders and database links

Query speed

- Aggregates
- Distinct
- Order by

Caching

- ETag: W/"2022_03"
- SERVICE
    - ETag code parses a query
    - Checks last modified of remote endpoint
    - Combined header

Large literals

- DNA as strings
  - Multi-gigabase
  - Max int as an index is a problem
- Proteins
  - 70,000+ (tintin)

Ecosystem

- Out of date tools/reports
- Push instead of pull
- Schema.org

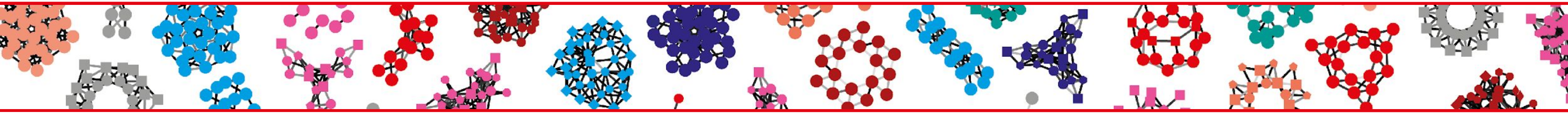
Fast?

- ZSTD
  - Facebook says its fast
  - Still takes a day to decompress 4TB

- Linked Data Fragments
  - SELECT ?x WHERE {?x a ?c }

- IO & Network

Overview



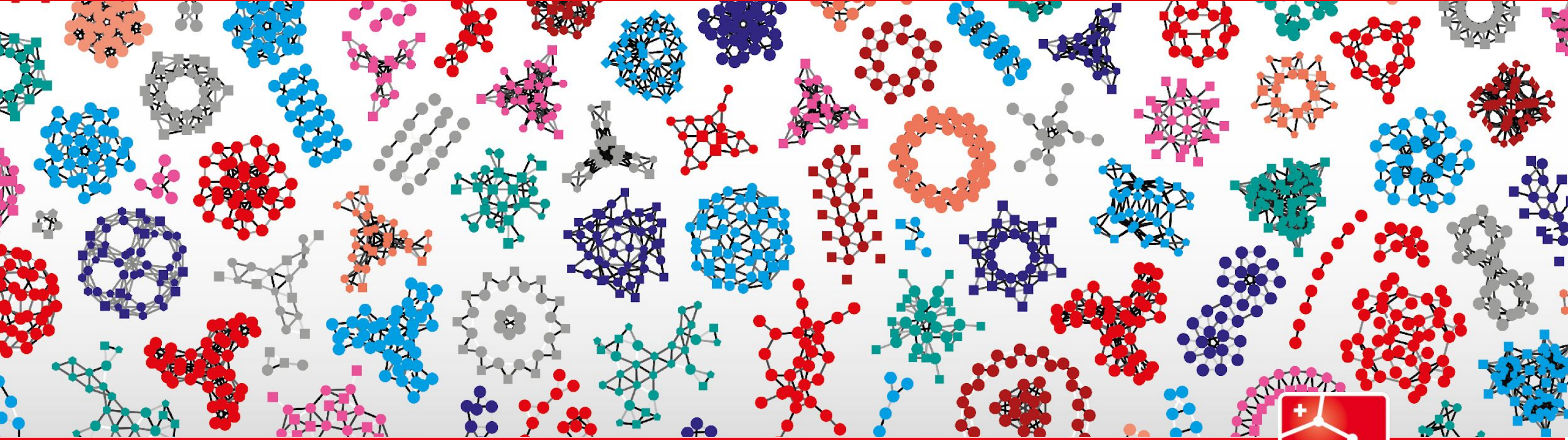
01 Background

02 Why RDF ?

03 How to be FAIR

04 Challenges

05 Questions (and hopefully answers)



SIB

Swiss Institute of Bioinformatics

UniProt: is a team effort!