Weekend Triple Billionaire
Maintaining a Large RDF Data Set in the Life Sciences

Jerven Bolleman, Thomas Kappler, and the UniProt Consortium

Swiss Institute of Bioinformatics
jerven.bolleman@isb-sib.ch, thomas.kappler@isb-sib.ch

Abstract. The UniProt Knowledgebase offers both manually curated and automatically generated information on proteins, and is one of the leading biological databases. While it is one of the largest free data sets that is available in RDF, our infrastructure and website are not based on RDF. We present numbers about the volume and growth of UniProt and show why this volume of data prevents using RDF triple stores and SPARQL with currently available tools.

1 UniProt Data: Nature and Volume

The UniProt Knowledgebase (UniProtKB) [1] consists of two parts: UniProtKB/ Swiss-Prot, containing manually annotated records describing proteins with information from literature and curator-evaluated computational analysis, and UniProtKB/TrEMBL, with automatically annotated records. The UniProt consortium provides several additional data sets: UniRef [2] with clustered sets of sequences from UniProt, the amino acid sequence archive UniParc [3], and supporting data sets such as taxonomy and keywords.

UniProt consists of almost three billion triples (see Table 1), making it one of the largest freely available RDF data sets. The data is maintained at three consortium member sites, and must be kept in sync and integrated into one combined public release[1].

This number of triples consumes considerable harddisk space. We show estimates comparing different RDF stores, based on published LUBM 8000[2] results, with our solution, in Table 2.

2 Performance

UniProt is released every three weeks. In each such period, there is a time window of only six days to prepare all source data for publication. When the curators freeze their work, we need to retrieve all related data from our global partners. We then convert all source data (42 GB as of release 15.8) to RDF/XML

[1] Available for download at http://www.uniprot.org/downloads.

[2] LUBM benchmark, http://swat.cse.lehigh.edu/projects/lubm/.

Table 1. Number of triples in UniProt, releases 15.8 and 15.9

Data set	Triples in 15.8	Triples in 15.9	Difference	Projection for September 2010
Citations	8,612,297	8,616,640	0.05%	
Enzyme	36,461	36,461	0%	
GO	195,249	195,249	0%	
Keywords	7,641	7,649	+0.10%	
Mapped citations[a]	84,308,128	82,605,680	−2.02%	
Locations	4,537	4,532	−0.11%	
Pathways	8,661	8,673	+0.14%	
Taxonomy	3,493,485	3,520,871	+0.78%	
Tissues	6551	6551	0%	
UniParc	640,191,073	691,432,967	+8.00%	2,560,000,000
UniProtKB	1,572,097,256	1,610,723,778	+2.46%	2,433,000,000
UniRef	487,209,989	499,947,698	+2.61%	775,000,000
Total	2,796,164,777	2,897,100,198	+3.61%	5,768,000,000

[a] Automatically generated, not public.

(180 GB, 18 GB compressed), while validating it. This conversion took around 16 hours for release 15.8 of September 2009.

The UniProt.org [4] website and its query engine run on an internally developed solution using BDB/je[3] and Lucene[4]. On a release, data is loaded into its store in RDF form; the store and query engine themselves, however, are not RDF-based. Loading and full text indexing all UniProt data sets took 29 hours for release 15.8, with 2 GB of memory on a dual core 2.8 GHz Xeon$^{\text{TM}}$and a single hard disk. More than 20,000 unique users visit UniProt.org each workday, averaging 130,000 queries and 1,710,000 direct lookups, while consuming just under a terabyte of bandwidth a month. On average, 99.9% of queries finish in less than 0.8 seconds, including the transfer over the internet.

UniProt.org runs on three mirrors, each of which needs to be provisioned with all data in time for a release. Data size is a major factor in our deployment as we are limited by upload speed. UniProt 15.8 as needed for the website consumes 39 GB when gzip compressed, and takes four hours to upload. Larger data sizes increase the risk of transfer failure.

We also have an internal website for our curators and internal software tools. Because it needs to reflect current work, we rebuild the manually curated UniProtKB/Swiss-Prot every night with a time window of 3.5 hours for converting, loading and validating about 160 million triples.

UniProtKB/TrEMBL with supporting data is converted, loaded, and validated every Sunday. At this occasion we generate about 1.4 billion triples while checking the data against our validation rules, making us indeed weekend triple

[3] http://www.oracle.com/database/berkeley-db/je/index.html

[4] http://lucene.apache.org/java/docs/index.html

billionaires. Current generic RDF stores are unable to handle this amount of data on the limited hardware budget available.

Table 2. Rough disk consumption estimates for UniProt release 15.8. The factor 1.4 is the ratio between the larger number of long literals in UniProt compared to a LUBM 8000 data set.

RDF Store	Data provided	Estimate		Source
Allegro Graph	155GB/1.1B LUBM 8000	552 GB	$\frac{155}{1.1} \times 2.8 \times 1.4$	([a])
OWLIM	92GB/1.85B LUBM 8000	195 GB	$\frac{92}{1.85} \times 2.8 \times 1.4$	([b])
Oracle 11G	154GB/1.1B LUBM 8000	549 GB	$\frac{154}{1.1} \times 2.8 \times 1.4$	([c])
Virtuoso	120GB/1.1B various	305GB	$\frac{120}{1.1} \times 2.8$[d]	([e])
UniProt.org[f]	36GB Store, 36GB Text index	72GB (real)		Internal

[a] www.franz.com/agraph/allegrograph/

[b] www.ontotext.com/owlim/OWLIMPres.pdf

[c] www.oracle.com/technology/tech/semantic_technologies/htdocs/performance.html

[d] Storage cost of Virtuoso was not multiplied as they used a dataset with a higher number of triples with large literals in comparison to LUBM8000.

[e] virtuoso.openlinksw.com/Whitepapers/html/Virt6FAQ.html#StorageCostPerTriple

[f] Our current custom solution, described in Section 2, not a SPARQL engine

3 Data Growth

We have to deal with ever growing data volumes (Fig. 3). When evaluating tools, we need to take into account not just performance on today's data, but also on the expected data in five years. The yearly growth rate of the core UniProtKB data is currently 51%. This is a doubling time of 17 months, faster than Moore's Law[5] that predicts a doubling time of 24 months.

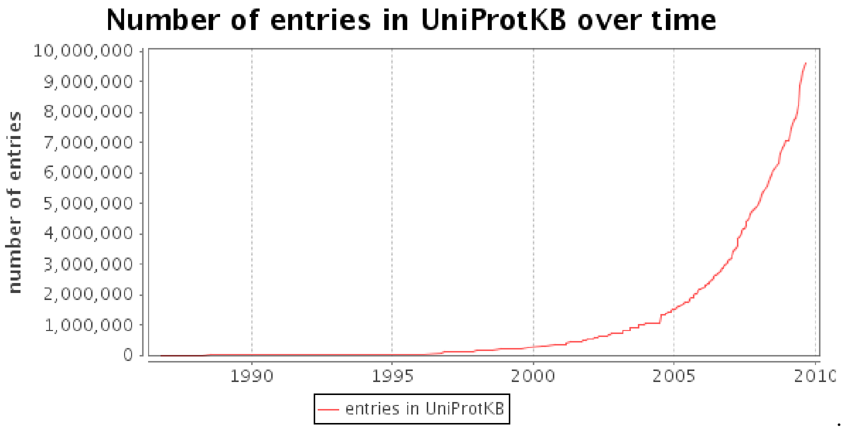
Assuming that the current growth rate of UniParc does not accelerate from the current 270% a year (doubling every 7 months), we will see at least 10 billion sequences in UniParc in five years. This estimate translates to 3×10^{11} triples, around 470 times the current size.

4 Summary

We have met the requirements for data processing speed and query performance for the near future, unfortunately on a non-RDF technology stack. Currently the available tools do not meet our performance needs for UniProt.org, especially since any RDF solution must preserve the current user interface and full text search. We do however aim to deploy a public SPARQL endpoint once it becomes feasible.

[5] http://en.wikipedia.org/wiki/Moores_law

Fig. 1. UniProtKB is growing faster over time. This graph does not show the growth of existing entries as more information becomes available.



5 Acknowledgments

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 contracts FELICS (021902RII3) and SLING (226073). PIR activities are also supported by the NIH grants and contracts HHSN266200400061C, NCI-caBIG, and 5R01GM080646-04, and the Department of Defense grant W81XWH0720112. This work has benefited from Oracle funding support.

References

1. The UniProt Consortium. The Universal Protein Resource (UniProt). Nucl. Acids Res. 37: D169-D174 (2009).
2. Suzek B.E., Huang H., McGarvey P., Mazumder R., and Wu C.H. UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters. Bioinformatics 23:1282-1288 (2007).
3. Apweiler R., Bairoch A., and Wu C.H. Protein sequence databases. Curr. Opin. Chem. Biol. 8:76-80 (2004).
4. Jain E., Bairoch A., Duvaud S., Phan I., Redaschi N., Suzek B.E., Martin M.J., McGarvey P., and Gasteiger E. Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics 10:136 (2009).