# UniProt in RDF

Jerven Bolleman, Thomas Kappler



Swiss Institute of
Bioinformatics

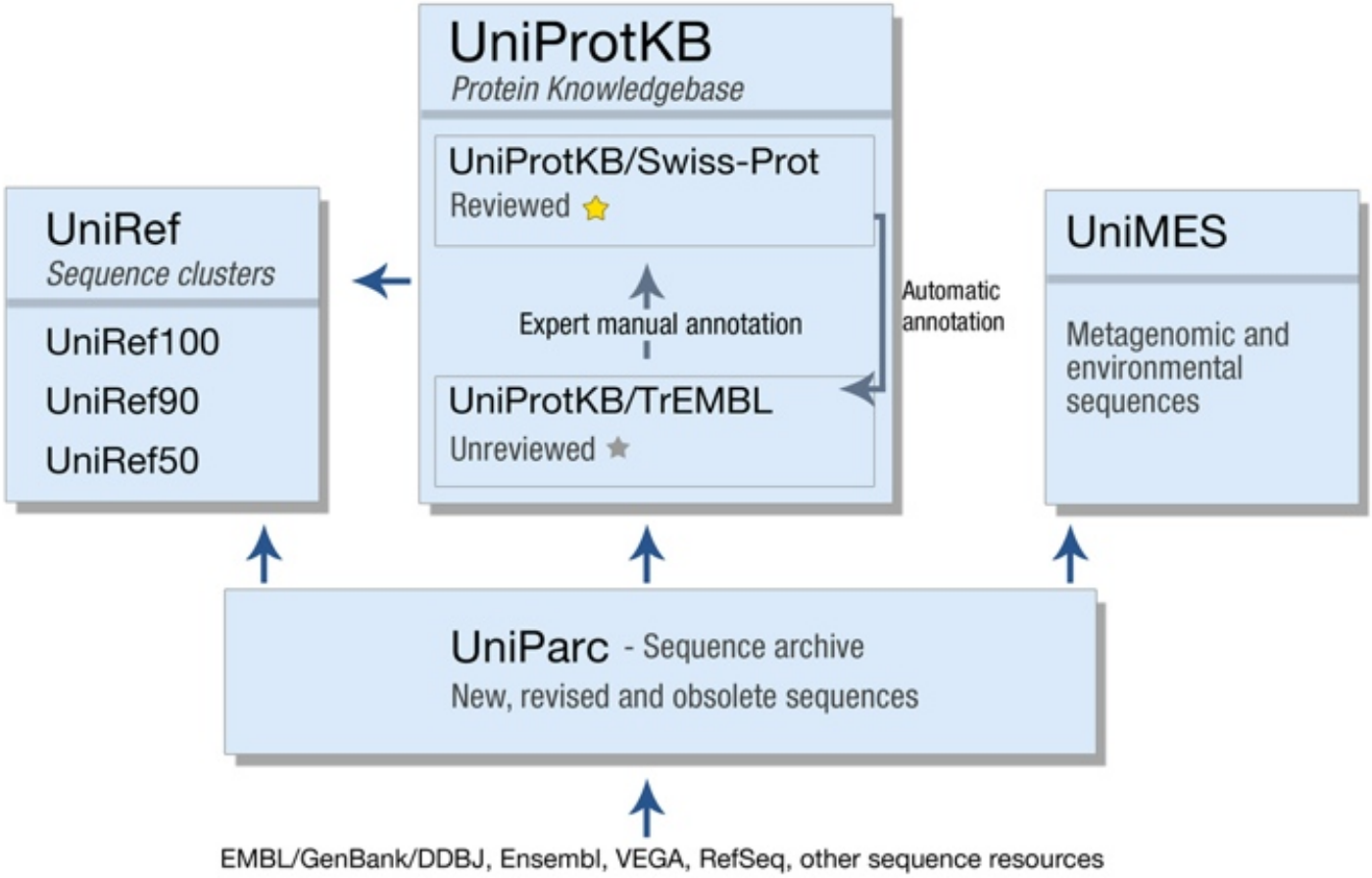
# UniProt in RDF

- Introduction to UniProt
- Practical UniProt RDF
- Experience with RDF
- Challenges with RDF
- What can you do with UniProt RDF

# What is UniProt

- Three data sets
  - **UniProtKB**
    - Swissprot
    - TrEMBL
  - **UniRef:** clusters
  - **UniParc:** archive

# What is UniProt: Databases



# What is UniProt: UniProtKB

- **Swiss-Prot**
  - *Manually annotated and curated* protein sequence records
    - Extracted from literature
    - Evaluated computational analysis
- **TrEMBL**
  - *Computationally analyzed* protein sequence records
    - Quality depends on quality of submission to
      - EMBL/DDBJ/Genbank
      - Ensembl
      - Model organism database
  - Automatic annotation
    - Based on data mining

# What is UniProt: UniRef

- **Clusters** of protein sequences
  - Sequence similarity
  - 100%, 90%, and 50%
  - Used to speed up blast searches
    - Statistics is also improved
    - No redundancy
  - 100% quickly shows conserved proteins

# What is UniProt: UniParc

- **Archive** of all possible protein sequences
  - Includes sequences not thought to exist, i.e., later corrected
- One record per sequence
  - 100% identical over full length
  - Merged across species
  - Tracks in which databases the sequence was/is available
- No annotation

Search | Blast | Align | Retrieve | ID Mapping

**Search in** Protein Knowledgebase (UniProtKB) **Query** Search Clear Fields »

## WELCOME

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

### What we provide

| | |
|---|---|
| UniProtKB | Protein knowledgebase, consists of two sections:<br>★ Swiss-Prot, which is manually annotated and reviewed.<br>☆ TrEMBL, which is automatically annotated and is **not** reviewed.<br>Includes Complete Proteome Sets. |
| UniRef | Sequence clusters, used to speed up similarity searches. |
| UniParc | Sequence archive, used to keep track of sequences and their identifiers. |
| Supporting data | Literature citations, taxonomy, keywords and more. |

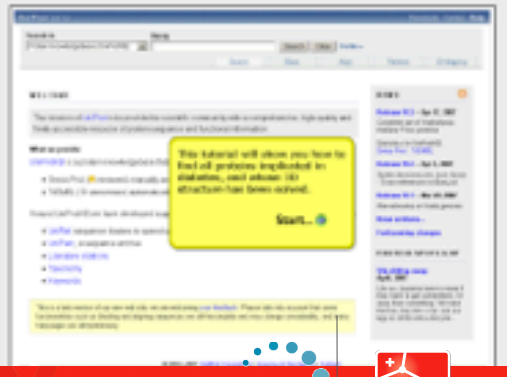
## NEWS

**UniProt release 15.13 – Jan 19, 2010**

*XMRV complete proteome in UniProtKB/Swiss-Prot* · Cross-references to eggNOG · Change to cross-references to HAMAP and HOGENOM

- Statistics for UniProtKB: Swiss-Prot · TrEMBL
- Forthcoming changes
- News archives

## SITE TOUR



# Reviewed, UniProtKB/Swiss-Prot **P99998** (CYC_PANTR)

**Last modified January 19, 2010. Version 58.** History...

Contribute
Send feedback
Read comments (0) or add your own

Clusters with 100%, 90%, 50% identity | Documents (2) | Third-party data | Customize display
TEXT XML RDF/XML GFF FASTA

Names and origin · Protein attributes · General annotation (Comments) · Ontologies · Sequence annotation (Features) · Sequences · References · Web resources · Cross-references · Entry information · Relevant documents

## Names and origin

Hide | Top

| | |
|---|---|
| Protein names | *Recommended name:* **Cytochrome c** |
| Gene names | Name: **CYCS** Synonyms: CYC |
| Organism | **Pan troglodytes (Chimpanzee)** |
| Taxonomic identifier | 9598 [NCBI] |
| Taxonomic lineage | Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorrhini › Catarrhini › Hominidae › Pan |

## Protein attributes

Hide | Top

| | |
|---|---|
| Sequence length | 105 AA |

```
<?xml version='1.0' encoding='UTF-8'?>
<rdf:RDF xmlns="http://purl.uniprot.org/core/"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:owl="http://www.w3.org/2002/07/owl#"
xmlns:skos="http://www.w3.org/2004/02/skos/core#">
<rdf:Description rdf:about="http://purl.uniprot.org/uniprot/P99998">
<rdf:type rdf:resource="http://purl.uniprot.org/core/Protein" />
<reviewed>true</reviewed>
<created>1986-07-21</created>
<modified>2010-01-19</modified>
<version>58</version>
<mnemonic>CYC_PANTR</mnemonic>
<replaces rdf:resource="http://purl.uniprot.org/uniprot/P00001" />
<replaces rdf:resource="http://purl.uniprot.org/uniprot/Q96BV4" />
<citation rdf:resource="http://purl.uniprot.org/citations/12766228" rdf:ID="_1"
/>
<citation rdf:resource="http://purl.uniprot.org/citations/4975694" rdf:ID="_2" />
<rdfs:seeAlso
rdf:resource="http://www.expasy.org/spotlight/back_issues/sptlt076.shtml" />
<rdfs:seeAlso rdf:resource="http://purl.uniprot.org/embl-cds/AAP49489.1" />
<rdfs:seeAlso rdf:resource="http://purl.uniprot.org/pir/CCCZ" />
<rdfs:seeAlso rdf:resource="http://purl.uniprot.org/refseq/NP_001065289.1" />
<rdfs:seeAlso rdf:resource="http://purl.uniprot.org/refseq/XP_001140708.1" />
```

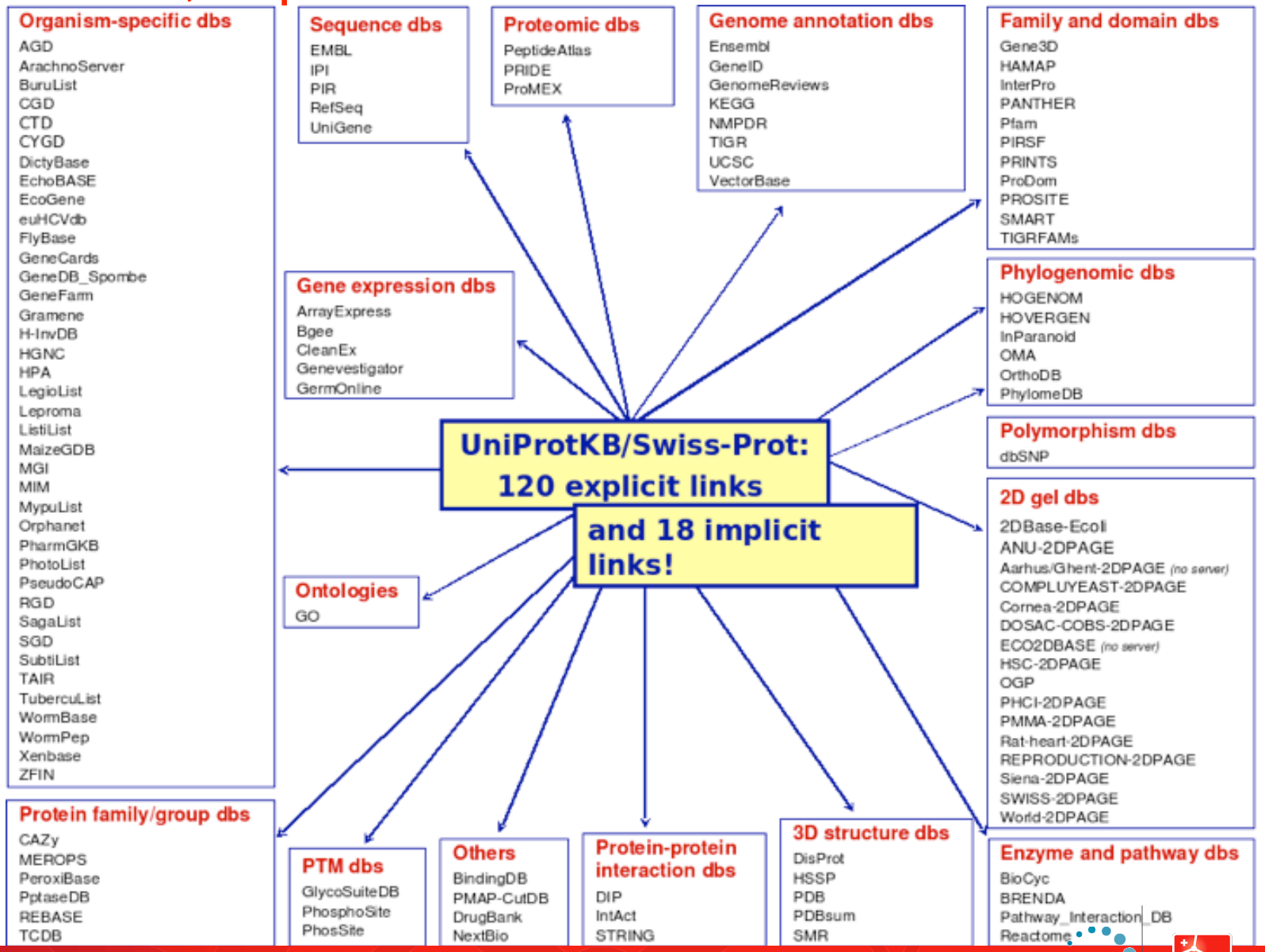
# Download via FTP

- http://www.uniprot.org/downloads

# UniProt: experience with RDF

- Releases all public data as RDF
  - Including supporting data
    - e.g. taxonomy and subset of Medline/Pubmed
- New editor for curators directly reads and writes RDF
- Internal data interchange format, e.g. for uniprot.org
- PURLs
  - dereferencable URIs
  - HTTP content negotiation
- OWL, basic support

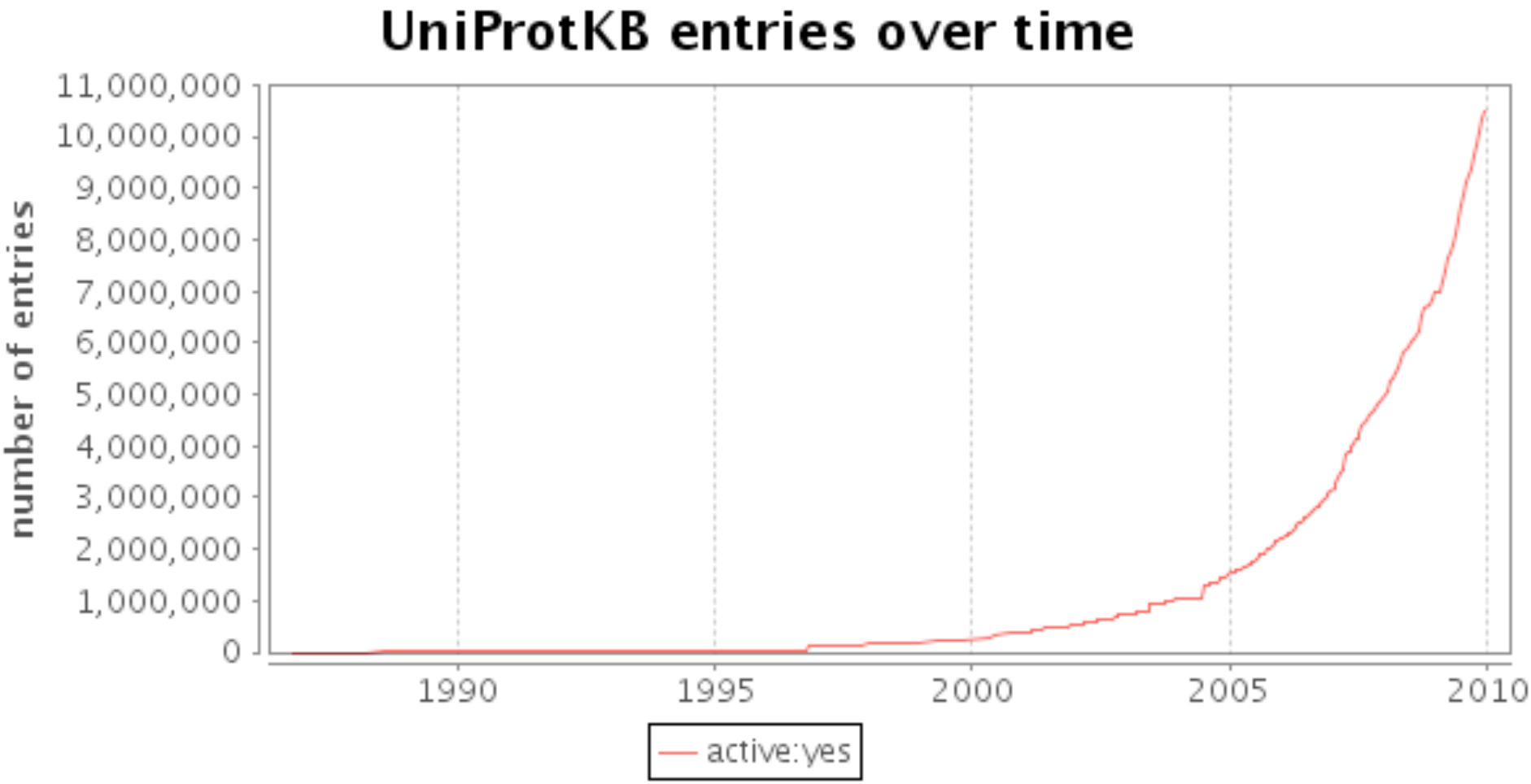
# UniProt; experience with RDF



# Experience: Complex Schema

- Changes as insight in biology improves
  - *All the time*
- Broad range of annotation

# Experience: Volume and growth



# Experience: Legacy

- 20+ year history
- Flat File
  - Slowly migrate users off
- XML
- Legacy programs

# Experience: www.UniProt.org

- Popular site
  - 25,000 - 30,000 users per workday
- 3 mirrors
  - Limited hardware
  - Limited bandwidth
- Update window of 6 days
- 85% searches: simple keywords
  - e.g. kinase, human, p53
  - requires full text indexing
- Low tolerance to latency

# What can you do with UniProt rdf?

- Reduce integration cost
- Query in depth
  - SPARQL at public endpoints
    - http://www.linkedlifedata.com/sparql
    - http://lod.openlinksw.com/sparql
  - Working on official end point
- OWL Reasoning
  - Infer implied knowledge (hackathon goal)

Thank you for listening!

Questions?
Hackathon wishes/ideas?

Feel free to e-mail questions to
jerven.bolleman@isb-sib.ch
thomas.kappler@isb-sib.ch

UniProt

SIB
Swiss Institute of Bioinformatics