

The THEA project

Claude Pasquier

1^{er} séminaire ISIBio, Institut Curie, Paris

30 mars 2005

The THEA project : overview

The THEA project (Tools for High-throughput Experiments Analysis) is dedicated to the elaboration of tools and methods suited for the analysis of post-genomic data.

The first module developed allows to:

- Automatically link gene symbols with selected biological information coming from a knowledge base,
- Manually search and browse through these annotations,
- Automatically generate meaningful generalizations according to statistical criteria.

THEA makes use of ontologies which constitute a popular way to modelize biological concepts and their relationships. Presently THEA includes:

- Gene Ontology (GO), base on a controlled vocabulary which can be used to annotate a gene product in regard to its molecular functions, cellular localizations and biological processes,
- Specific vocabularies describing the developmental stages and the anatomy of *Drosophila melanogaster*,
- The Medical Subject Headings (MeSH), used for indexing biomedical and health-related documents,
- InterPro, a collection of protein families, domains and functional sites.

This example of configuration of THEA's user interface shows six activated plugins.

The **Terms viewer** (top left), allows to graphically browse through the loaded ontologies and select individual terms.

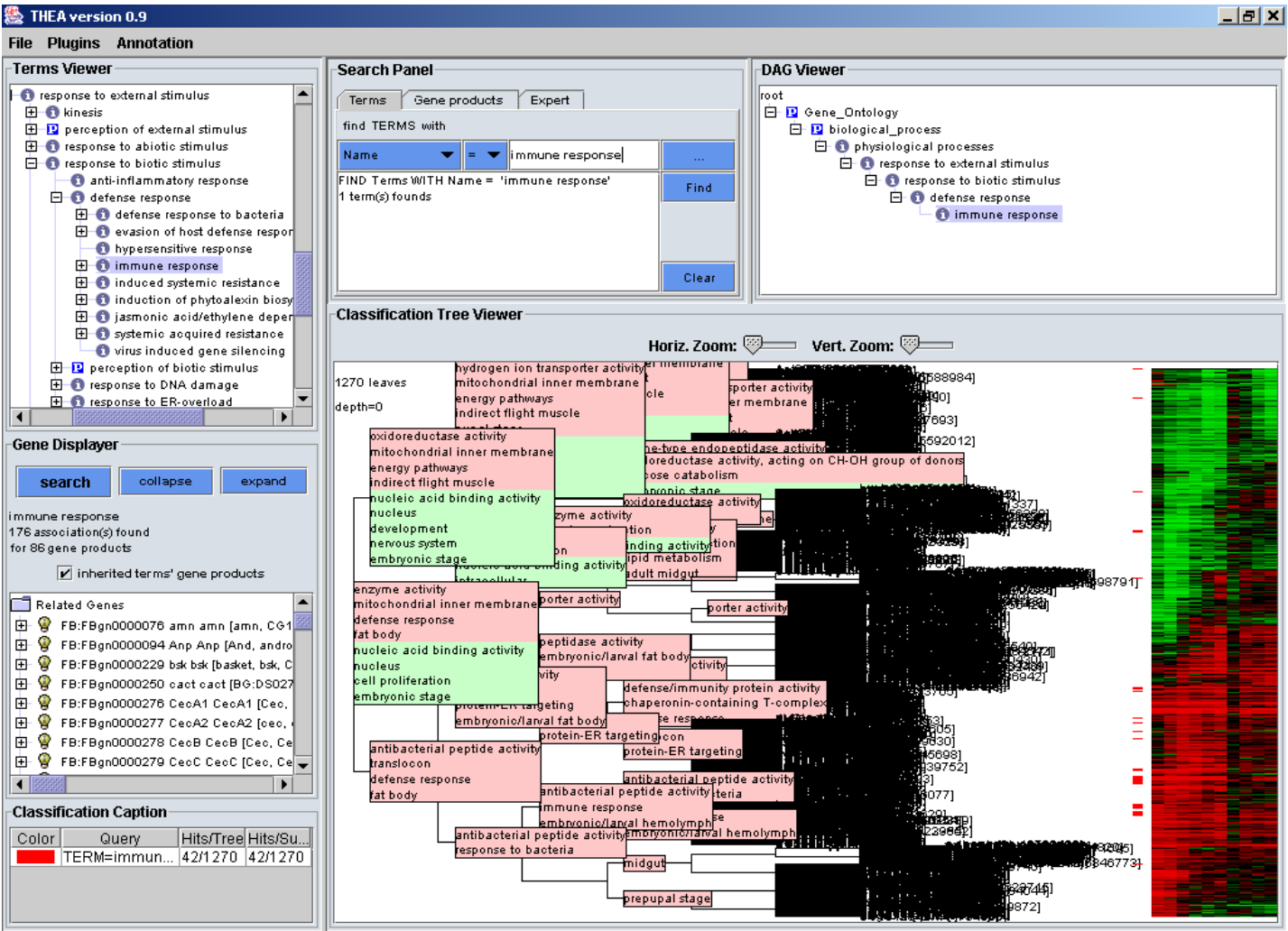
The **Search Panel** (middle top) allows to search for specific terms or character strings among genes or terms. Complex queries are also possible. Here the term “immune response” has been selected.

The **DAG Viewer** (top right), displays a trimmed ontology leading to the selected term.

The **Gene Displayer** (middle left) displays the genes associated to a given term and, eventually, their children (here the 86 genes associated to the term ‘immune response’ and their children are displayed) and provides access to the relevant publications and databases *via* Internet.

The **Classification Caption** (bottom left) displays the legends of the color-coded labels as well as some statistical information.

The **Classification Tree Viewer** (bottom right) allows the display of hierarchical clustering results of gene expression profiles. Genes associated to the selected term and its children are labeled with colored boxes. The names automatically assigned to each cluster are displayed in this plugin as well.



Analysis of a dataset consisting of 1270 regulated genes (De Gregorio, E., Spellman, P.T., Rubin, G.M. and Lemaitre, B. (2001) Genome-wide analysis of the Drosophila immune response by using oligonucleotide microarrays. Proc Natl Acad Sci U S A, 98, 12590-12595).

The THEA project : features and facts

Thea v1.2.3, released September 2004, includes 11 plugins.

The application can be used to analyze a single list of genes or complete classification trees in 3 different formats (*Newick*, *Eisen*, *Sota*)

Groups of co-expressed genes can be automatically labelled with terms found to be over- and/or under -represented in the set using several (parametrable) methods: terms count, density of terms, statistical calculation (Binomial or Hypergeometric law).

Calculations can be relative to gene products present in the classification, known genes in a genome or a user-specified list of genes.

The associated database contains:

250 871 genes from 13 organisms (*S. cerevisiae*, *C. briggsae*, *C. elegans*, *A. gambiae*, *D. melanogaster*, *A. mellifera*, *D. rerio*, *G. gallus*, *P. troglodytes*, *H. sapiens*, *M. musculus*, *R. norvegicus* & *T. nigroviridis*),

61 690 genes annotated with at least one ontology term,

1 668 845 known identifiers in 36 databases, including Affymetrix, EMBL/GenBank, LocusLink, RefSeq, UniGene, UniProt, Ensembl, Hugo

Work in progress : THEA2 & AllOnto

THEA2 is a new development of the application built on top of the NetBeans platform. All functionalities of THEA will be available in a new environment with :

- **Plug-and-play concept.** 3rd party tools can be developed independently and “plugged” together.
- **A New Windowing system.** Users can conveniently and flexibly arrange graphical tools on the screen.
- **Zero administration.** Application keeps itself up-to-date with regular updates.
- **Transparent access to the knowledge base.** With lazy loading, data are automatically retrieved when needed.

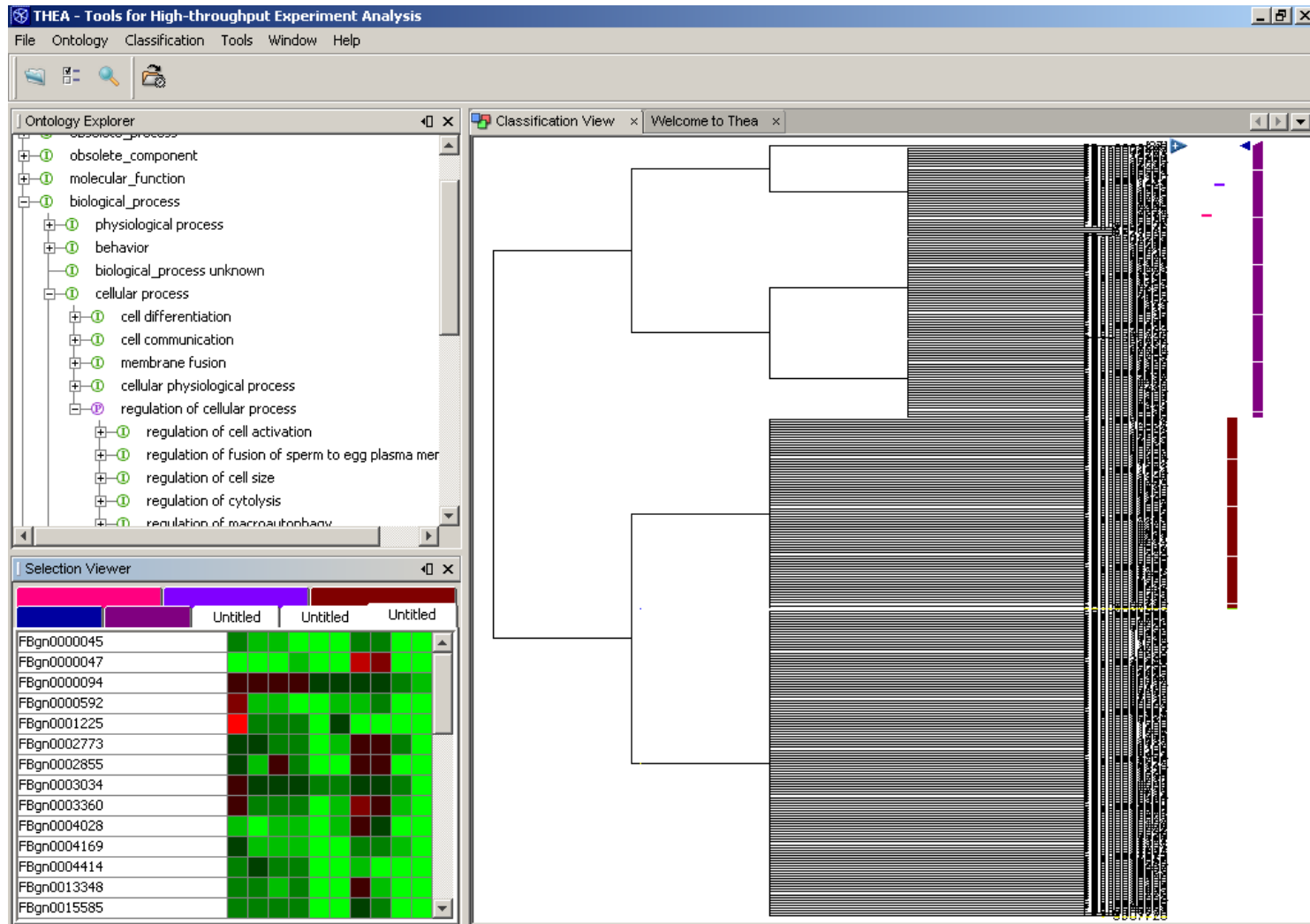
Allonto is a knowledge base for RDF datasets. It will includes :

- RDFS and OWL Reasoning Capabilities (with Pellet OWL reasoner),
- Data persistence in RDBMS,,
- A Query module (using RDQL syntax).
- A possibility to modelize the “context” of assertions

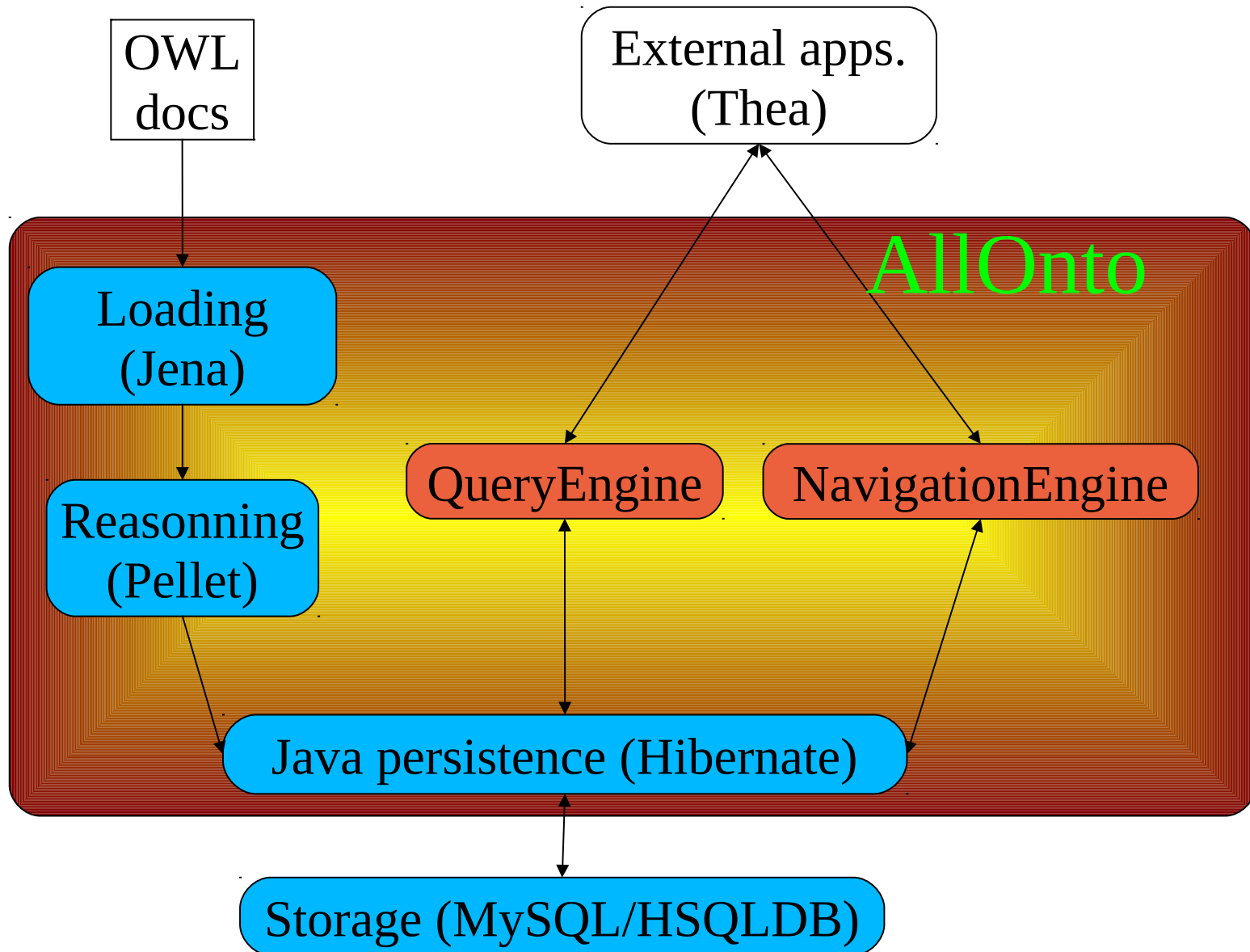
The objective is double :

- To be able to store several tens of millions of statements (GO may be modeled in OWL with about 300 000 statements),
- To be able to execute queries in a reasonable time (about a few seconds)

Snapshot of THEA2 accessing a knowledge base storing “Gene Ontology”,
all genes from *Drosophila Melanogaster* and their annotations
(550 000 resources ; 2 000 000 statements)



AllOnto architecture



Modelling biological data in RDF format.

Some examples (YeastHub)

```
<rdf:Description rdf:about="http://db.yeastgenome.org/cgi-bin/locus.pl?locus=S000007287">
<ns0:DB>SGD</ns0:DB> <ns0:DB_Object_ID>S000007287</ns0:DB_Object_ID>
<ns0:DB_Object_Symbol>15S_RRNA</ns0:DB_Object_Symbol> <ns0:Qualifier_ />
<ns0:GO_ID>GO:0005763</ns0:GO_ID>
<ns0:DB_Reference>SGD_REF:S000073642</ns0:DB_Reference>
<ns0:DB_Reference>PMID:6261980</ns0:DB_Reference>
<ns0:Evidence_>ISS</ns0:Evidence_> <ns0:With_Or_From /> <ns0:Aspect>C</ns0:Aspect>
<ns0:DB_Object_Name /> <ns0:DB_Object_Synonym>15S_rRNA</ns0:DB_Object_Synonym>
<ns0:DB_Object_Synonym>15S_RRNA_2</ns0:DB_Object_Synonym>
<ns0:DB_Object_Type>gene</ns0:DB_Object_Type> <ns0:Taxon>taxon:4932</ns0:Taxon>
<ns0>Date>20040202</ns0>Date> <ns0:Assigned_by>SGD</ns0:Assigned_by> </rdf:Description>
```

=> Equivalent to a tabular format

Modelling biological data in RDF format.

Some examples (Reactome)

```
<protein
rdf:ID="UniProt_P48023_1_Tumor_necrosis_factor_ligand_superfamily_member_6__FAS_antigen_ligand__Apoptosis_antig
en_ligand__APTL__CD178_antigen_">
  <ORGANISM rdf:resource="#Homo_sapiens" />
  <XREF rdf:resource="#UniProt_P48023" />
</protein>
<unificationXref rdf:ID="UniProt_P48023">
  <DB rdf:datatype="http://www.w3.org/2001/XMLSchema#string">UniProt</DB>
  <ID rdf:datatype="http://www.w3.org/2001/XMLSchema#string">P48023</ID>
</unificationXref>
<bioSource rdf:ID="Homo_sapiens">
  <NAME rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Homo sapiens</NAME>
  <TAXON-XREF rdf:resource="#NCBI_taxonomy_9606" />
</bioSource>
<xref rdf:ID="NCBI_taxonomy_9606">
  <DB rdf:datatype="http://www.w3.org/2001/XMLSchema#string">NCBI_taxonomy</DB>
  <ID rdf:datatype="http://www.w3.org/2001/XMLSchema#string">9606</ID>
</xref>
```

The diagram illustrates the relationships between different biological data sources. A red arrow points from the `<XREF rdf:resource="#UniProt_P48023" />` line in the first XML block to the `<unificationXref rdf:ID="UniProt_P48023">` line in the second block. A blue arrow points from the `<ORGANISM rdf:resource="#Homo_sapiens" />` line in the first block to the `<bioSource rdf:ID="Homo_sapiens">` line in the second block. A green arrow points from the `<TAXON-XREF rdf:resource="#NCBI_taxonomy_9606" />` line in the second block to the `<xref rdf:ID="NCBI_taxonomy_9606">` line in the third block.

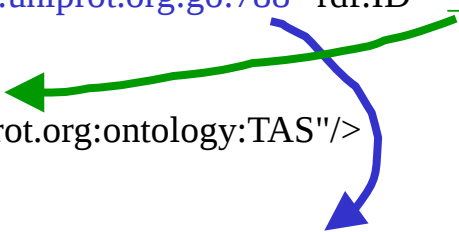
=> External resources identified by literals

Modelling biological data in RDF format.

Some examples (UniProt)

```
<rdf:Description rdf:about="urn:lsid:uniprot.org:uniprot:P02309">
<rdf:type rdf:resource="urn:lsid:uniprot.org:ontology:Protein"/>
<organism rdf:resource="urn:lsid:uniprot.org:taxonomy:4932"/>
<rdfs:seeAlso rdf:resource="urn:lsid:uniprot.org:pir:HSBY4"/>
<rdfs:seeAlso rdf:resource="urn:lsid:uniprot.org:sgd:S000000213"/>
<classifiedWith rdf:resource="urn:lsid:uniprot.org:go:788" rdf:ID="_216"/>
</rdf:Description>
<rdf:Description rdf:about="#_216">
<evidence rdf:resource="urn:lsid:uniprot.org:ontology:TAS"/>
</rdf:Description>

<rdf:Description rdf:about="urn:lsid:uniprot.org:go:0000788">
<rdf:type rdf:resource="urn:lsid:uniprot.org:ontology:Concept"/>
<rdfs:label>nuclear nucleosome</rdfs:label>
<rdfs:comment>A complex comprised of DNA wound around [...] </rdfs:comment>
<rdfs:subClassOf rdf:resource="urn:lsid:uniprot.org:go:0000786"/>
<rdfs:subClassOf rdf:resource="urn:lsid:uniprot.org:go:0000790"/>
</rdf:Description>
```



=> Closed world modelling

Modelling biological data in RDF format.

Some examples (Gene Ontology)

```
<owl:Class xmlns:owl=http://www.w3.org/2002/07/owl# rdf:ID="GO_0000788">
  <rdfs:label>nuclear nucleosome</rdfs:label>
  <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string">A complex comprised of DNA wound around a multisubunit core
and associated proteins, which forms the primary packing unit of DNA
in the nucleus into higher order structures.</rdfs:comment>
<!-- nucleosome -->
  <rdfs:subClassOf rdf:resource="#GO_0000786"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty>
        <owl:ObjectProperty rdf:about="#part_of"/>
      </owl:onProperty>
      <owl:someValuesFrom rdf:resource="#GO_0000790"/>
    </owl:Restriction>
  </rdfs:subClassOf>
<!-- nuclear chromatin -->
</owl:Class>
```

=> Terms are represented by classes

Problems of data integration

Gene Ontology

```
<owl:Class xmlns:owl="http://www.w3.org/2002/07/owl#" rdf:ID="GO_0005515">  
  <rdfs:label>protein binding</rdfs:label>  
  <rdfs:subClassOf rdf:resource="#GO_0005488"/>  
</owl:Class>
```

Uniprot

```
<rdf:Description rdf:about="urn:lsid:uniprot.org:uniprot:P31946">  
  <classifiedWith rdf:resource="urn:lsid:uniprot.org:go:5515" rdf:ID="_25"/>  
  <organism rdf:resource="urn:lsid:uniprot.org:taxonomy:9606"/>  
</rdf:Description>
```

```
<protein  
rdf:ID="UniProt_P31946_14_3_3_protein_beta_alpha__Protein_kinase_C_inhibitor_protein_1__KCIP_1__Protein_1054_"  
>  
  
  <ORGANISM rdf:resource="#Homo_sapiens" />  
  
  <NAME rdf:datatype="http://www.w3.org/2001/XMLSchema#string">UniProt:P31946 14-3-3 protein beta/alpha (Protein  
kinase C inhibitor protein-1) (KCIP-1) (Protein 1054)</NAME>  
  
  <SYNONYMS rdf:datatype="http://www.w3.org/2001/XMLSchema#string">YWHAB</SYNONYMS>  
  
  <XREF rdf:resource="#UniProt_P31946" />  
  
</protein>
```

Reactome

≠

≠

≠

≠

A possible solution

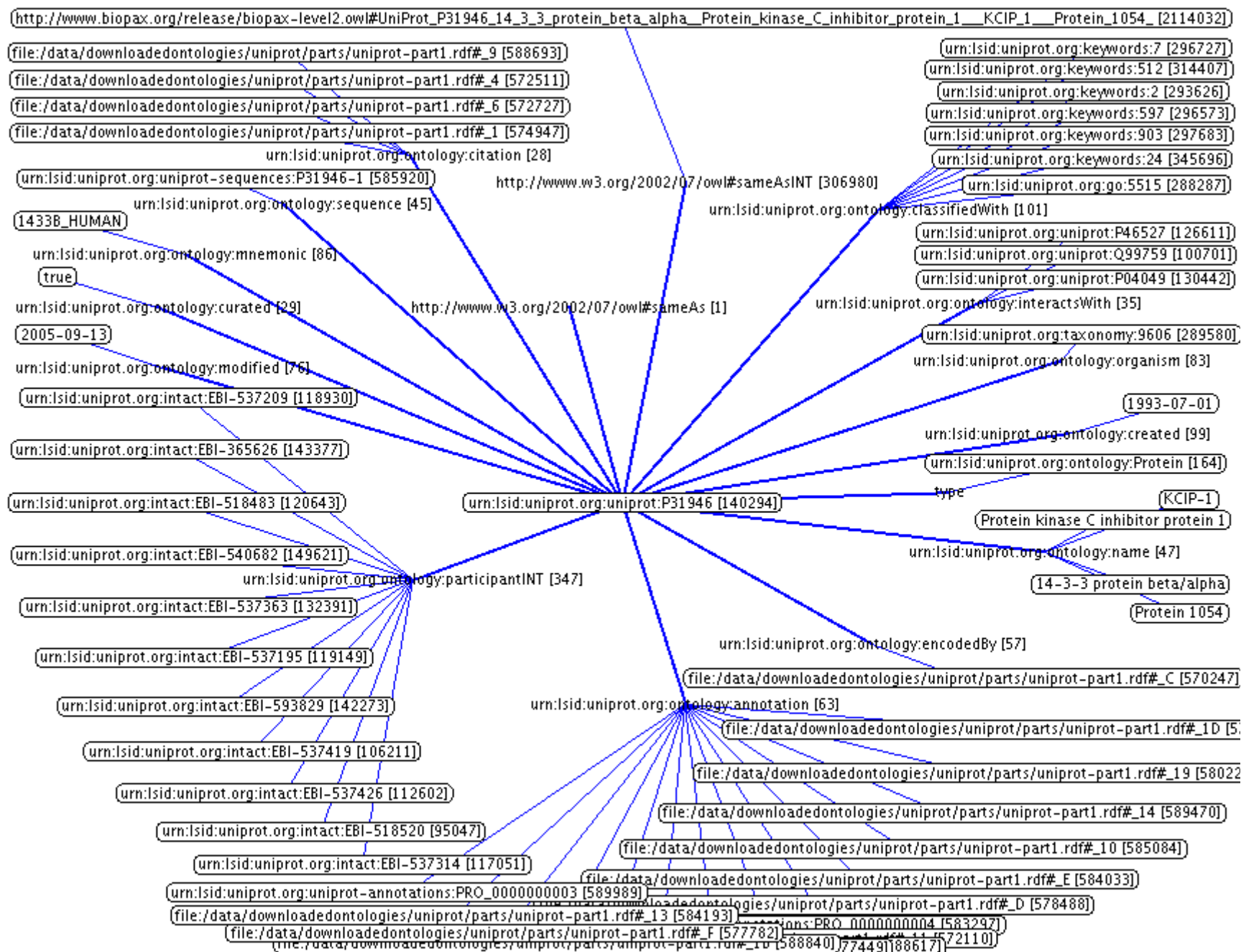
Uniprot

```
<rdf:Description rdf:about="urn:lsid:uniprot.org:uniprot:P31946">  
<classifiedWith rdf:resource="urn:lsid:uniprot.org:go:5515" rdf:ID="_25"/>  
<organism rdf:resource="urn:lsid:uniprot.org:taxonomy:9606"/>  
</rdf:Description>
```

```
<rdf:Description  
rdf:about="#UniProt_P31946_14_3_3_protein_beta_alpha__Protein_kinase_C_inhibit  
or_protein_1__KCIP_1__Protein_1054_">  
  <owl:sameAs rdf:resource="urn:lsid:uniprot.org:uniprot:P31946"/>  
</rdf:Description>
```

```
<protein  
rdf:ID="UniProt_P31946_14_3_3_protein_beta_alpha__Protein_kinase_C_inhibitor_protein_1__KCIP_1__Protein_1054_"  
>  
  <ORGANISM rdf:resource="#Homo_sapiens" />  
  <NAME rdf:datatype="http://www.w3.org/2001/XMLSchema#string">UniProt:P31946 14-3-3 protein beta/alpha (Protein  
kinase C inhibitor protein-1) (KCIP-1) (Protein 1054)</NAME>  
  <SYNONYMS rdf:datatype="http://www.w3.org/2001/XMLSchema#string">YWHAB</SYNONYMS>  
  <XREF rdf:resource="#UniProt_P31946" />  
</protein>
```

Reactome



and tryptophan hydroxylases in the presence of Ca(2+)/calmodulin-dependent protein kinase II, and strongly activates protein kinase C. Is probably a multifunctional regulator



Example of simple RDQL queries

Retrieve all gene products annotated with the term “aging” (GO:0007568)

```
SELECT ?gene "  
WHERE (<go:GO_0007568>, <biowl:annotates>, ?gene )  
USING go FOR <http://www.geneontology.org/owl#>  
      , biowl FOR <http://www.unice.fr/bioinfo/owl/biowl#>
```

Execution with AllOnto : 5 results found in 1,6s

Retrieve all gene products with their corresponding GO annotation

```
SELECT ?go, ?gene "  
WHERE (?go, <biowl:annotates>, ?gene )  
USING go FOR <http://www.geneontology.org/owl#>  
      , biowl FOR <http://www.unice.fr/bioinfo/owl/biowl#>
```

Execution with AllOnto : 36773 results found in 25s

Example of contextual RDQL queries

Retrieve all articles describing an annotation of “chico” with the term “aging”

```
SELECT ?db, ?acc
WHERE (<go:GO_0007568>, <biowl:annotates>, <biowl:CG5686>, ?c )
    , (?c, <biowl:xref>, ?x)
    , (?x, <biowl:acc>, ?acc);
    , (?x, <biowl:db", ?db);
USING go FOR <http://www.geneontology.org/owl#>
    , biowl FOR <http://www.unice.fr/bioinfo/owl/biowl#>
```

Execution with AllOnto : 1 result found in 1,4s