



Semantic data integration platform for biomedical domain

Vassil Momtchev

Group Leader, Semantic Life Science Applications

Linked Life Data Features

- **Link data between** different data silos applications
- Put the **information into context**
- Generate **new instance mappings**
- Analyze the knowledge **locked into text**
- Expose the data for access via the **HTTP protocol**

Drugs

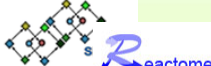


W3C SWEO Community Project

Linked Open Drug Data



Interactions and Pathways






Biological Pathway Exchange

Genes and Proteins




Ontologies and Thesauri





Unified Medical Language System®

The Open Biomedical Ontologies



Text-Analysis




Semantic Annotations

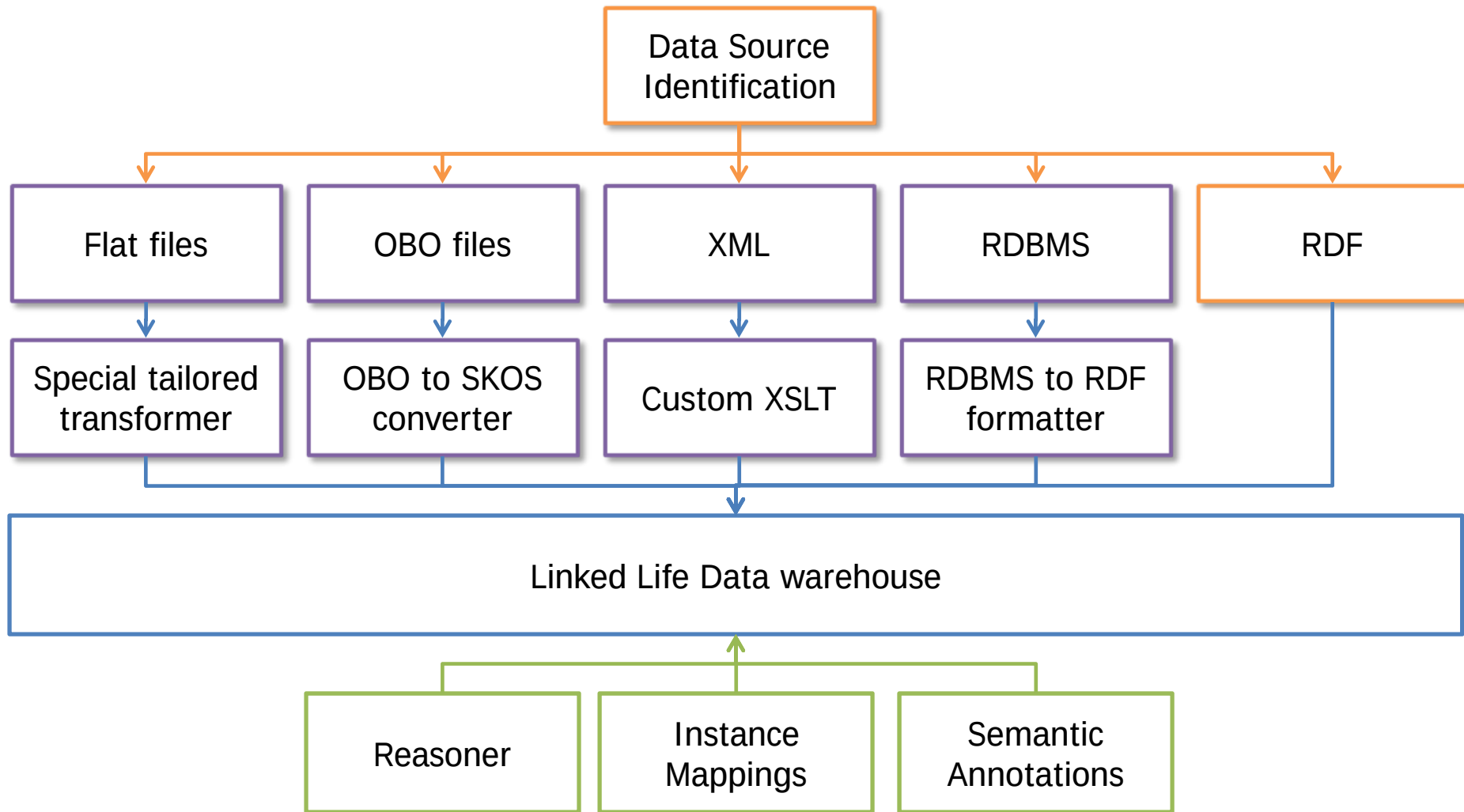
Documents




A service of the U.S. National Institutes of Health

**Over 4 billion
RDF
statements!**

The Warehousing Integration Process



Design Decisions: URI naming convention

- R1:** Preserve the original RDF structure if distributed by the owner.
- R2:** Use resolvable URIs for the data sources with no RDF distribution
- R3:** Construct the generated URIs in the form of `lld:resource/db/type/id`
- R3:** Identify the graph names with `lld:resource/db`
- R4:** Name all generated predicate URIs `lld:resource/db/predicate`
- R5:** Generate stable new URIs based on unique label that describes the resource (see dataset provenance and updates)

Design Decisions: Dataset Provenance and Updates

Step 1 Regenerate the data source

Step 2 Delete the existing graph information

Step 3 Import the new data

Constraints:

- Every data source has its own named graph

Subject	Predicate	Object	Context
umls-label:A10776677	rdf:type	skos-xl:Label	lld:umls
umls-label:A10776677	skos-xl:literalForm	Chronic Obstructive Airways Disease	lld:umls

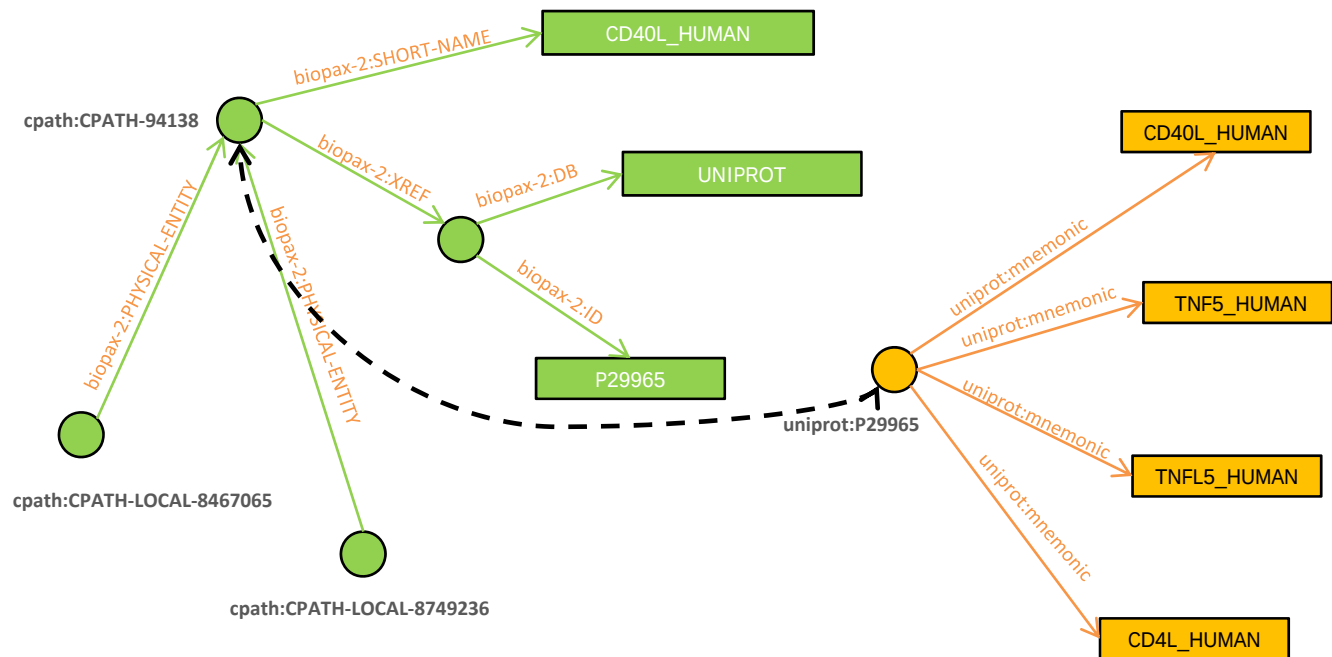
- Only stable URI must be generated

[pubmed-meshheading:Asthma17950641](#)

Source: <http://linkedlifedata.com/resource/pubmed/meshheading/Asthma17950641>

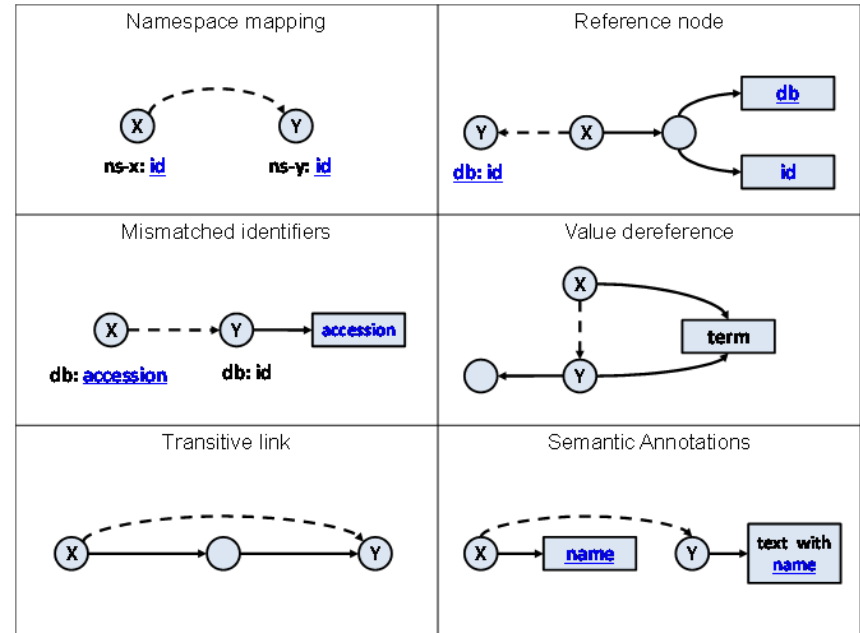
Design Decisions: Linked Data Mapping

- How well interlinked is the linked data cloud?
 - Many interesting queries are difficult to be expressed in SPARQL
 - String functions could not be index
 - Often there are misplaced identifiers



Design Decisions: Linked Data Mapping #2

- Identified 6 linked data integration patterns
- Define meta-rules to connect resources with various predicates
- Manually controlled process



The blue lines and the blue text of the captions (used either as part of the URI or literals) designate the criteria for linking the information

New Type of Possible Query #1

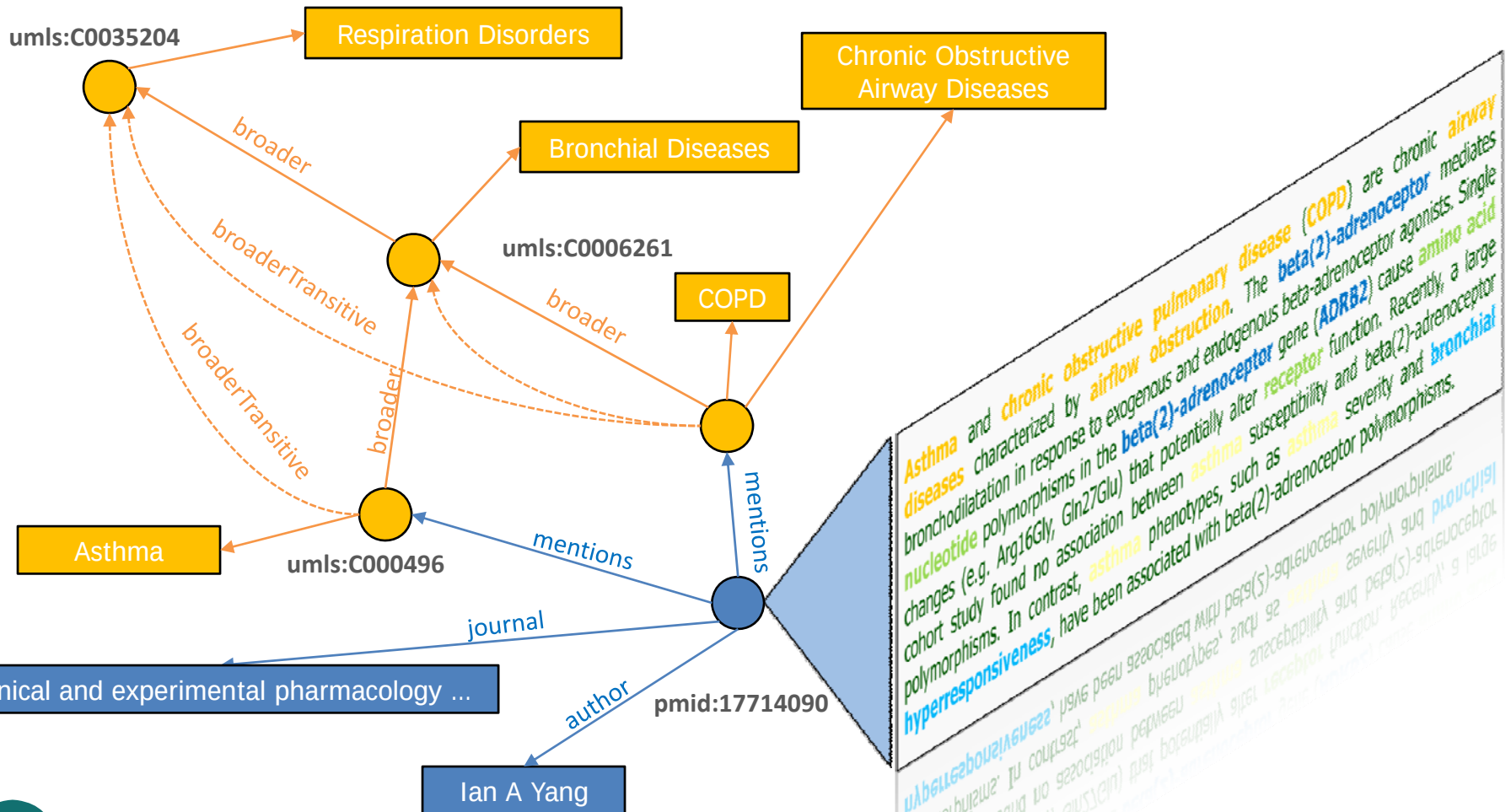
Select drugs related to asthma that are linked to a curated molecular interaction in the literature where the protein is known to cause inflammatory response

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX biopax2: <http://www.biopax.org/release/biopax-level2.owl#>
PREFIX uniprot: <http://purl.uniprot.org/core/>
PREFIX drugbank: <http://www4.wiwi.fu-berlin.de/drugbank/resource/drugbank/>

SELECT DISTINCT ?fullname ?drugname
WHERE {
  ?interaction rdf:type biopax2:physicalInteraction .
  ?interaction biopax2:PARTICIPANTS ?participant .
  ?participant biopax2:PHYSICAL-ENTITY ?physicalEntity .
  ?physicalEntity skos:exactMatch ?protein .
  ?protein uniprot:classifiedWith <http://purl.uniprot.org/go/0006954> .
  ?protein uniprot:recommendedName ?name .
  ?name uniprot:fullName ?fullname .
  ?target skos:exactMatch ?protein .
  ?drug drugbank:target ?target .
  ?drug drugbank:genericName ?drugname .
  ?drug drugbank:indication ?indication .
}
```

The **red** graph patterns indicate the usage of mapping rules.

Design Decisions: Semantic Annotations



New Type of Possible Query #2

Select all located in Y-chromosome, human genes with known molecular interactions, which are analysed with 'Transfection'

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX gene: <http://linkedlifedata.com/resource/entrezgene/>
PREFIX core: <http://purl.uniprot.org/core/>
PREFIX biopax2: <http://www.biopax.org/release/biopax-level2.owl#>
PREFIX lifeskim: <http://linkedlifedata.com/resource/lifeskim/>
PREFIX umls: <http://linkedlifedata.com/resource/umls/>
PREFIX pubmed: <http://linkedlifedata.com/resource/pubmed/>
```

```
SELECT distinct ?genedescription ?prefLabel ?pmid
WHERE {
    ?interaction rdf:type biopax2:interaction .
    ?interaction biopax2:PARTICIPANTS ?p .
    ?p biopax2:PHYSICAL-ENTITY ?protein .
    ?protein skos:exactMatch ?uniprotaccession .
    ?uniprotaccession core:organism
    <http://purl.uniprot.org/taxonomy/9606> .
    ?geneid gene:uniprotAccession ?uniprotaccession .
    ?geneid gene:description ?genedescription .
    ?geneid gene:pubmed ?pmid .
    ?geneid gene:chromosome 'Y' .
    ?pmid lifeskim:mentions ?umlsid .
    ?umlsid skos:prefLabel 'Transfection' .
    ?umlsid skos:prefLabel ?prefLabel .
}
```

Query Results

Results for [PREFIX rdf: <http://www.w3....](#) (14)

View as [Exhibit](#) Download in [JSON](#) | [SPARQL Results in XML](#) | [SPARQL Results in JSON](#)

genedescription	prefLabel	pmid
interleukin 9 receptor	Transfection	pubmed-citation:1376929
RNA binding motif protein, Y-linked, family 1, member A1	Transfection	pubmed-citation:11149922
vesicle-associated membrane protein 7	Transfection	pubmed-citation:18362137
CD99 molecule	Transfection	pubmed-citation:16421247
colony stimulating factor 2 receptor, alpha, low-affinity (granulocyte-macrophage)	Transfection	pubmed-citation:12504125
interleukin 3 receptor, alpha (low affinity)	Transfection	pubmed-citation:12504125
sex determining region Y	Transfection	pubmed-citation:18454134
jumonji, AT rich interactive domain 1D	Transfection	pubmed-citation:9143681
zinc finger, BED-type containing 1	Transfection	pubmed-citation:12663651
sex determining region Y	Transfection	pubmed-citation:9346931
short stature homeobox	Transfection	pubmed-citation:12960152
CD99 molecule	Transfection	pubmed-citation:16984917
thymosin beta 4, Y-linked	Transfection	pubmed-citation:15557202
solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 6	Transfection	pubmed-citation:14746803

New Type of Possible Query #3

Select all participating in
interactions human genes
which are drug target and
are analysed with
'Transfection'

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX gene: <http://linkedlifedata.com/resource/entrezgene/>
PREFIX core: <http://purl.uniprot.org/core/>
PREFIX biopax2: <http://www.biopax.org/release/biopax-level2.owl#>
PREFIX lifeskim: <http://linkedlifedata.com/resource/lifeskim/>
PREFIX umls: <http://linkedlifedata.com/resource/umls/>
PREFIX pubmed: <http://linkedlifedata.com/resource/pubmed/>
PREFIX drugbank: <http://www4.wiwiw.fu-berlin.de/drugbank/resource/drugbank/>
```

```
SELECT distinct ?genedescription ?prefLabel ?drugname ?pmid
WHERE {
    ?interaction rdf:type biopax2:interaction .
    ?interaction biopax2:PARTICIPANTS ?p .
    ?p biopax2:PHYSICAL-ENTITY ?protein .
    ?protein skos:exactMatch ?uniprotaccession .
    ?uniprotaccession core:organism
    <http://purl.uniprot.org/taxonomy/9606> .
    ?geneid gene:uniprotAccession ?uniprotaccession .
    ?geneid gene:description ?genedescription .
    ?geneid gene:pubmed ?pmid .
    ?pmid lifeskim:mentions ?umlsid .
    ?umlsid skos:prefLabel 'Transfection' .
    ?umlsid skos:prefLabel ?prefLabel .
    ?target skos:closeMatch ?geneid .
    ?drug drugbank:target ?target .
    ?drug rdfs:label ?drugname .
}
```

Query Results

SPARQL Query

7 record(s)

genedescription▼	prefLabel	drugname	pmid
S100 calcium binding protein P	Transfection	Cromoglicate	http://linkedlifedata.com/resource/pubmed/id/16061848
retinoic acid receptor, gamma	Transfection	SR11254, Tazarotene, Alitretinoin, Dodecyl-Alpha-D-Maltoside, CD564, Tretinoin, Etretinate, Adapalene, BMS184394, Acitretin, and 4-[3-Oxo-3-(5,5,8,8-Tetramethyl-5,6,7,8-Tetrahydro-Naphthalen-2-Yl)-Propenyl]-Benzoic Acid	http://linkedlifedata.com/resource/pubmed/id/1318502
progesterone receptor	Transfection	Progesterone, Mifepristone, Methyltrienolone, Megestrol, Dydrogesterone, Norgestimate, Norgestrel, Tanaproget, Desogestrel, Norethindrone, Levonorgestrel, Drospirenone, Etonogestrel, Medroxyprogesterone, and Ethynodiol Diacetate	http://linkedlifedata.com/resource/pubmed/id/12101239 , http://linkedlifedata.com/resource/pubmed/id/16647340 , http://linkedlifedata.com/resource/pubmed/id/15084343 , and http://linkedlifedata.com/resource/pubmed/id/15084345
interleukin 8	Transfection	Ketoprofen, Simvastatin, Zileuton, and Salbutamol	http://linkedlifedata.com/resource/pubmed/id/14645117 , http://linkedlifedata.com/resource/pubmed/id/17035306 , and http://linkedlifedata.com/resource/pubmed/id/15039334
gonadotropin-releasing hormone receptor	Transfection	Nafarelin, Leuprolide, Danazol, Cetrorelix, Abarelix, and Gonadorelin	http://linkedlifedata.com/resource/pubmed/id/16613990
aldo-keto reductase family 1, member C4 (chlordecone reductase; 3-alpha hydroxysteroid dehydrogenase, type I; dihydrodiol dehydrogenase 4)	Transfection	NADH	http://linkedlifedata.com/resource/pubmed/id/11158055
3-hydroxy-3-methylglutaryl-Coenzyme A reductase	Transfection	1,4-Dithiothreitol, Pravastatin, Rosuvastatin, Adenosine-5'-Diphosphate, Simvastatin, NADH, 2'-Monophosphoadenosine 5'-Diphosphoribose, Atorvastatin, and Lovastatin	http://linkedlifedata.com/resource/pubmed/id/14697242

genedescription

- 3-hydroxy-3-methylglutaryl-Coenzyme A reductase
- aldo-keto reductase family 1, member C4 (chlordecone reductase; 3-alpha hydroxysteroid dehydrogenase, type I; dihydrodiol dehydrogenase 4)
- gonadotropin-releasing hormone receptor

prefLabel

- Transfection

drugname

- 1,4-Dithiothreitol
- 2'-Monophosphoadenosine 5'-Diphosphoribose
- 4-[3-Oxo-3-(5,5,8,8-Tetramethyl-5,6,7,8-Tetrahydro-Naphthalen-2-Yl)-Propenyl]-Benzoic Acid
- Abarelix
- Acitretin

pmid

- <http://linkedlifedata.com/resource/pubmed/id/11158055>
- <http://linkedlifedata.com/resource/pubmed/id/14697242>

Classical Information Retrieval Queries

- Lucene based index
- Special predicate to execute full-text queries
- Multiple retrieval modes
 - Literal
 - RDF molecules

```
select * where {
```

```
?article <http://www.ontotext.com/luceneQuery> "+lung  
COPD^5 asthma^3".
```

```
?article <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>  
<http://linkedlifedata.com/resource/pubmed/Citation>.
```

```
?article  
<http://linkedlifedata.com/resource/pubmed/articleTitle>  
?title.
```

```
} limit 1000
```

Wrap-up

- LLD is free and actively developed public service available:
<http://linkedlifedata.com>
- We have new improved UI
- LLD uses OWLIM engine which is experimentally proven to scale up to:
 - 20 billion RDF statements (15 billions explicit)
 - On a computer that costs less than 10'000\$
- Why not give it a try?

Acknowledgement

- AstraZeneca
- LODD
- BioRDF
- HCLSIG



The development of Linked Life Data is partially funded by
FP7 215535