



Linked Life Data

Vassil Momtchev

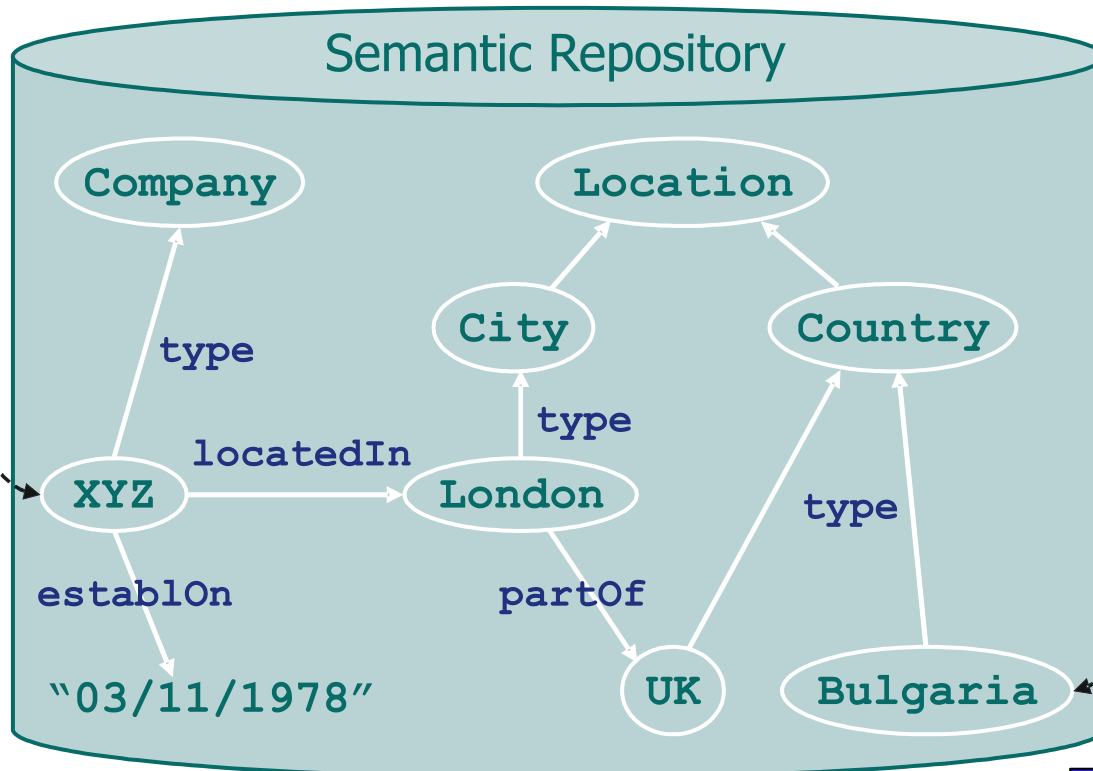
19/04/2011

Outline

- Semantic Data Integration
- Linked Life Data concept
- Integrated datasets
- Behind the scene

Interlinking Text and Data

XYZ announced profits in Q3, planning to build a \$120M plant in Bulgaria, and more and more and more and more text...



Semantic Technologies vs. AI

If It Works, It's Not AI:

A Commercial Look at Artificial Intelligence Startups

Eve M. Phillips, M.Sc. Thesis, 1999 MIT

One can think of “Semantic Technologies” like as AI,
made less abstract and more robust,
predictable and manageable

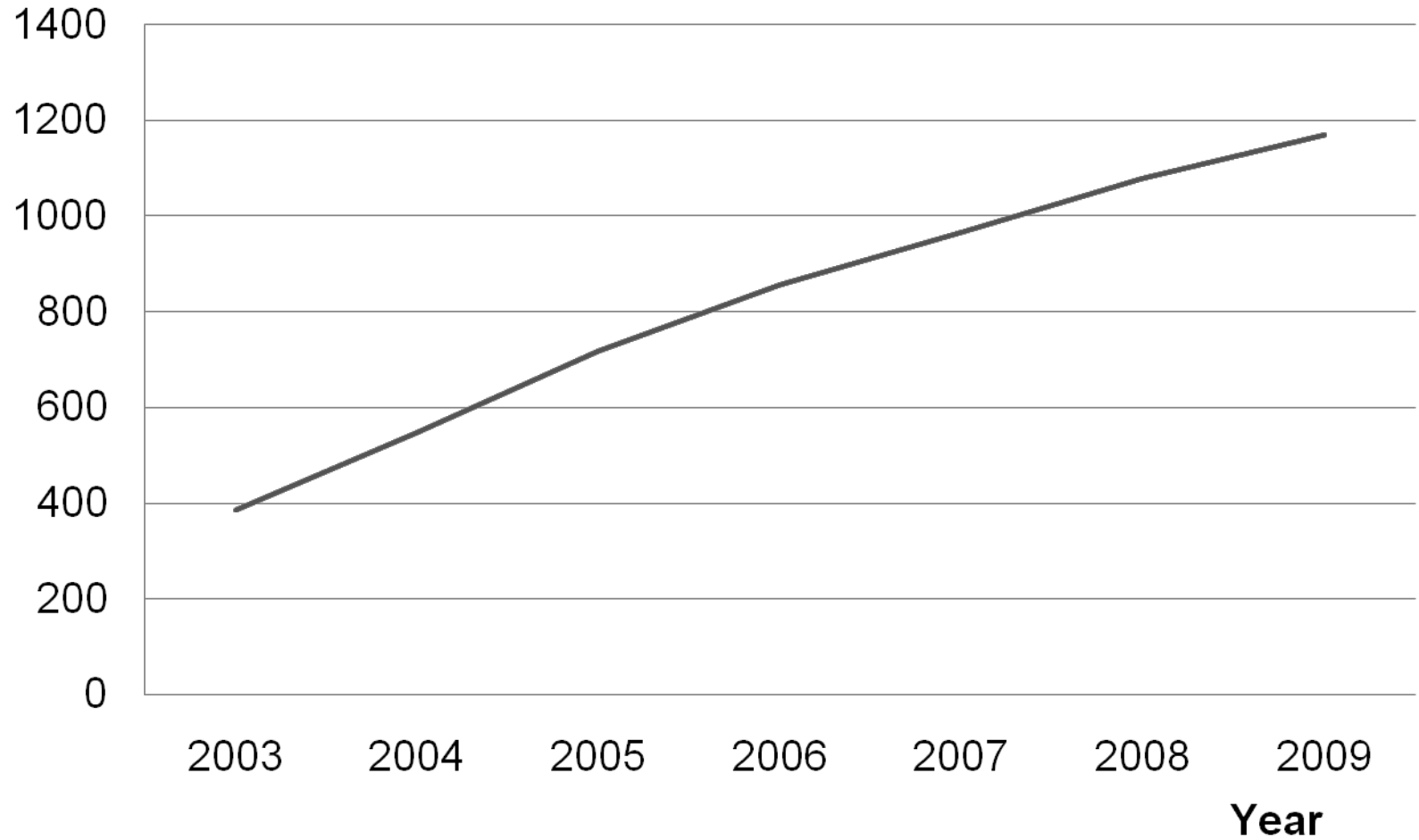
Semantic Technologies

- **“Semantic technologies” (ST)** is a general term for any software that involves **some kind and level of understanding** the meaning of the information it deals with
- Examples:
 - A search engine that can match a query for *“bird”* with a document mentioning *“eagle”*
 - A database that will return Ivan as a result of a query for *“?x relativeOf Maria”*, when the fact asserted was *“Maria motherOf Ivan”*
 - A navigation system that is *more intelligent than what we are already used to*

Ontotext Positioning

- **Leading semantic technology provider**
 - Top-5 core semantic technology developer
 - Supplying engines and components to vendors and solution developers
- **Unique technology portfolio:**
 - **Semantic Databases:** high-performance RDF DBMS, scalable reasoning
 - **Semantic Search:** text-mining (IE), Information Retrieval (IR)
 - **Web Mining:** focused crawling, screen scraping, data fusion
- **Good recognition in the SemTech community**
 - Ontotext pages are ranked #1 for “semantic annotation” and “semantic repository” at GYM

Time to Guess It?

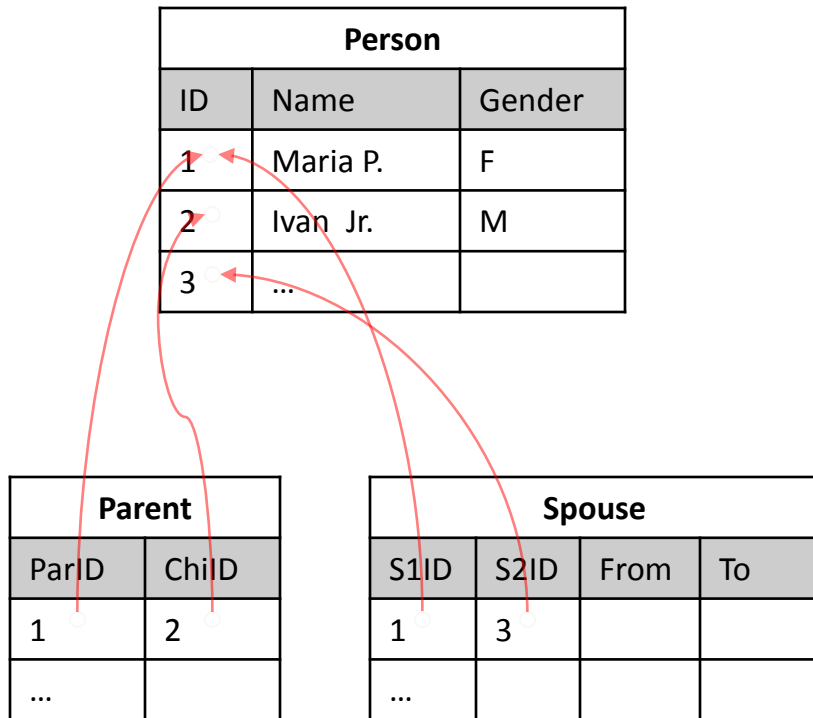


Massive Data Integration Problem

- Extreme amount of data with inconsistent syntax, structure and semantics
- Data is supported by different organizations
- Information is highly distributed and redundant
- Knowledge is locked in vast data silos
- Isolated communities which could not reach cross-domain understanding

Increase the abstraction level of the data!

Data representation: RDBMS vs. RDF



Relational Tables

Statement		
Subject	Predicate	Object
myo:Person	rdf:type	rdfs:Class
myo:gender	rdfs:type	rdfs:Property
myo:parent	rdfs:range	myo:Person
myo:spouse	rdfs:range	myo:Person
myd:Maria	rdf:type	myo:Person
myd:Maria	rdf:label	"Maria P."
myd:Maria	myo:gender	"F"
myd:Maria	rdf:label	"Ivan Jr."
myd:Ivan	myo:gender	"M"
myd:Maria	myo:parent	Myd:Ivan
myd:Maria	myo:spouse	myd:John
...		

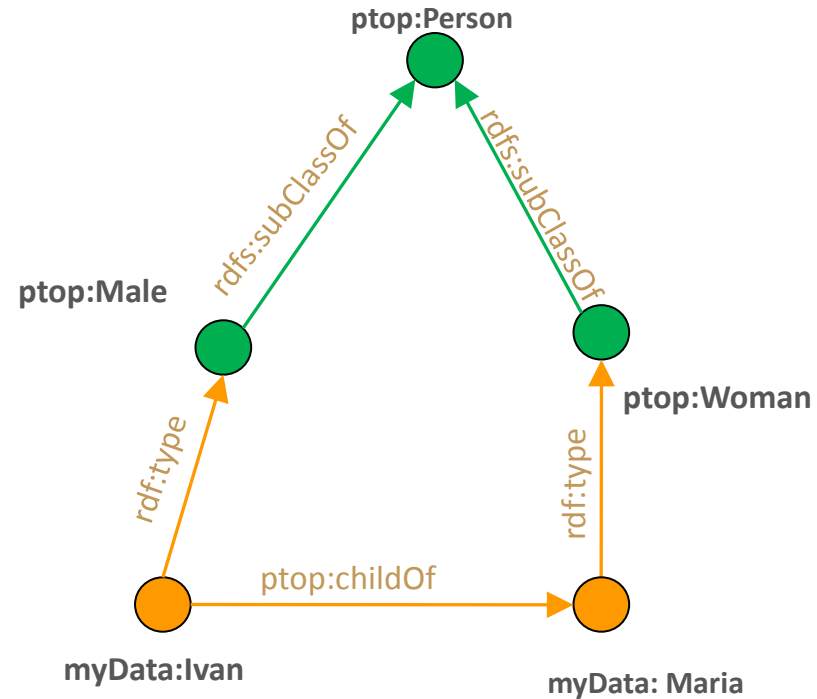
RDF Representation

Data representation: XML vs. RDF

```
<document>
  <person>
    <name>Maria</name>
    <gender>F</gender>
    <relList>
      <rel type="child">Ivan</rel>
    <relList>
  </person>
```

- No agreement over the structure and the vocabulary
- Could not be semantically compared by machine

XML Documents



RDF Representation

Linked Data Design Principles

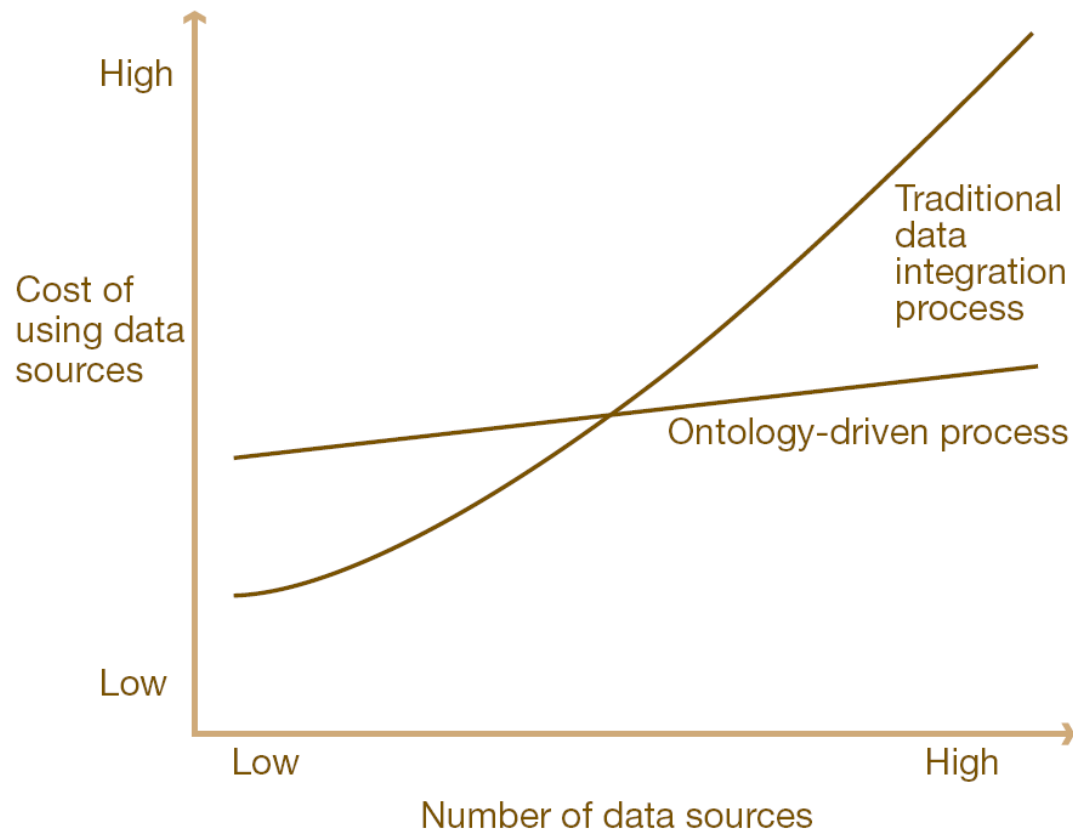
- Unambiguous identifiers for objects (resources)
 - Use URIs as names for things
- Use the structure of the web
 - Use HTTP URIs so that people can look up the names
- Make it easy to discover information about an object (resource)
 - When someone looks up a URI, provide useful information
- Link the object (resource) to related objects
 - Include links to other URIs

PWC on Semantic Technologies

Spring of the data Web

Technology forecast, A quarterly journal, Spring 2009,

<http://www.pwc.com/techforecast/>



There is Nothing You Can Do ...

There is nothing you can do with ontologies
that cannot be done without them

The same holds for language technology:
given unlimited resources, all methods will deliver
comparable results for any text analysis task (Y. Willks)

BTW, there is also nothing you can on Java
than cannot be done on Assembler

Conceptual idea

LINKED LIFE DATA

Semantic Data Integration

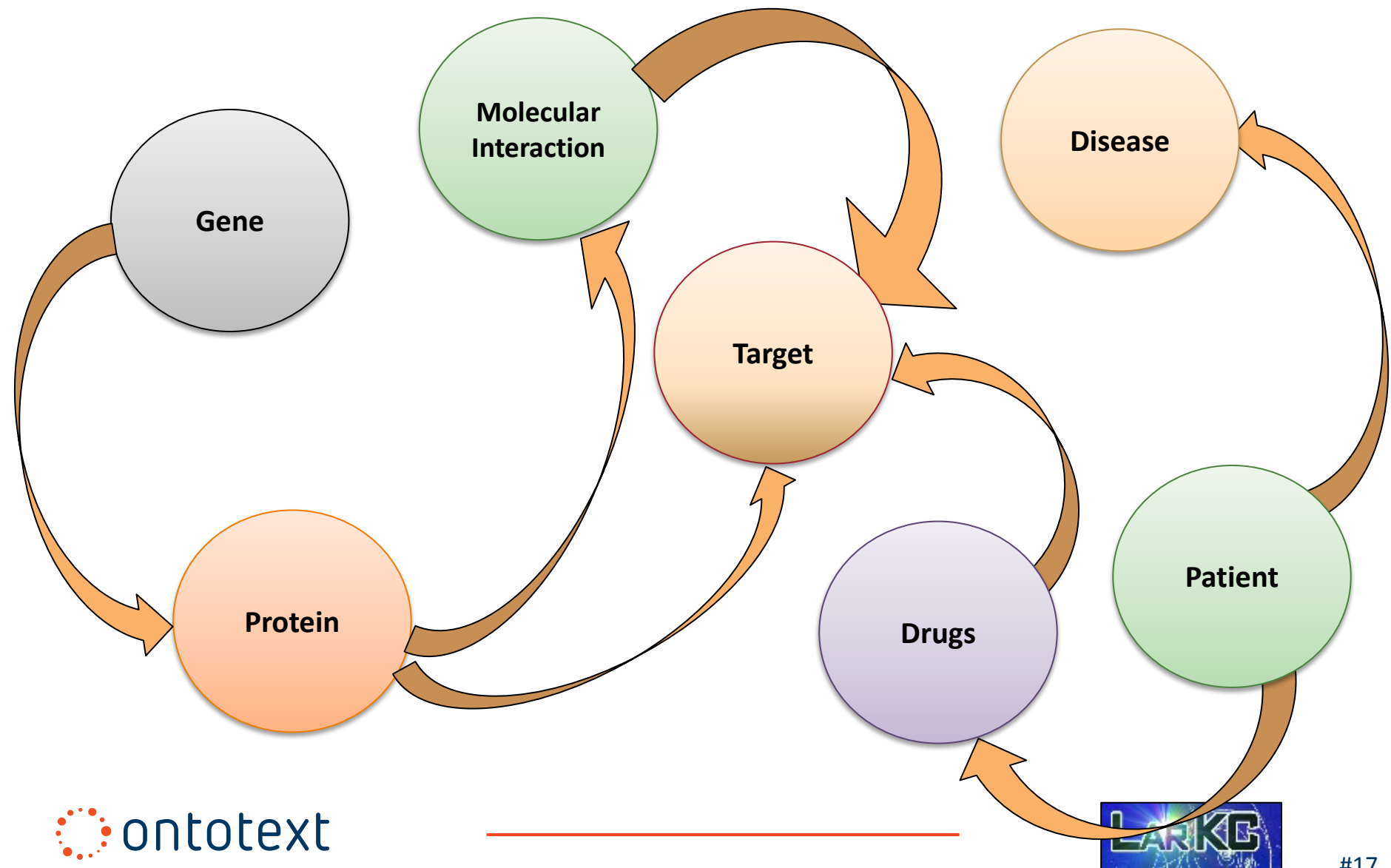
Current

- A lot of biomedical data available on the web and internally
- Very hard to locate the information and put it into context
- Scientists unable to utilize existing information well
- Difficult to automatically combine public domain knowledge with private company expertise

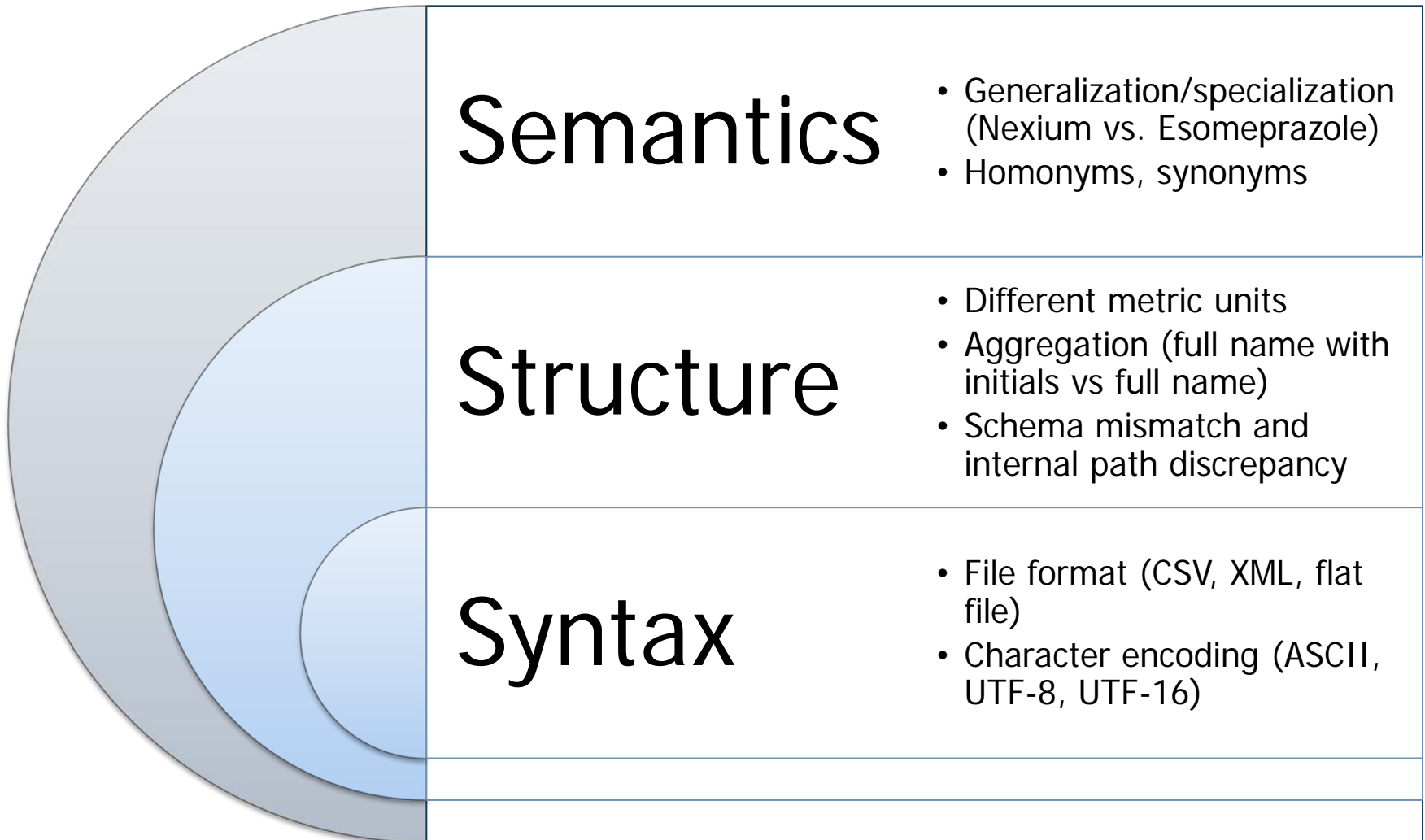
Desired

- Single integration model based on linked data technology and open standards
- Computerized support to interpret the information
- Assists scientists to combine internal data from experiments with external knowledge

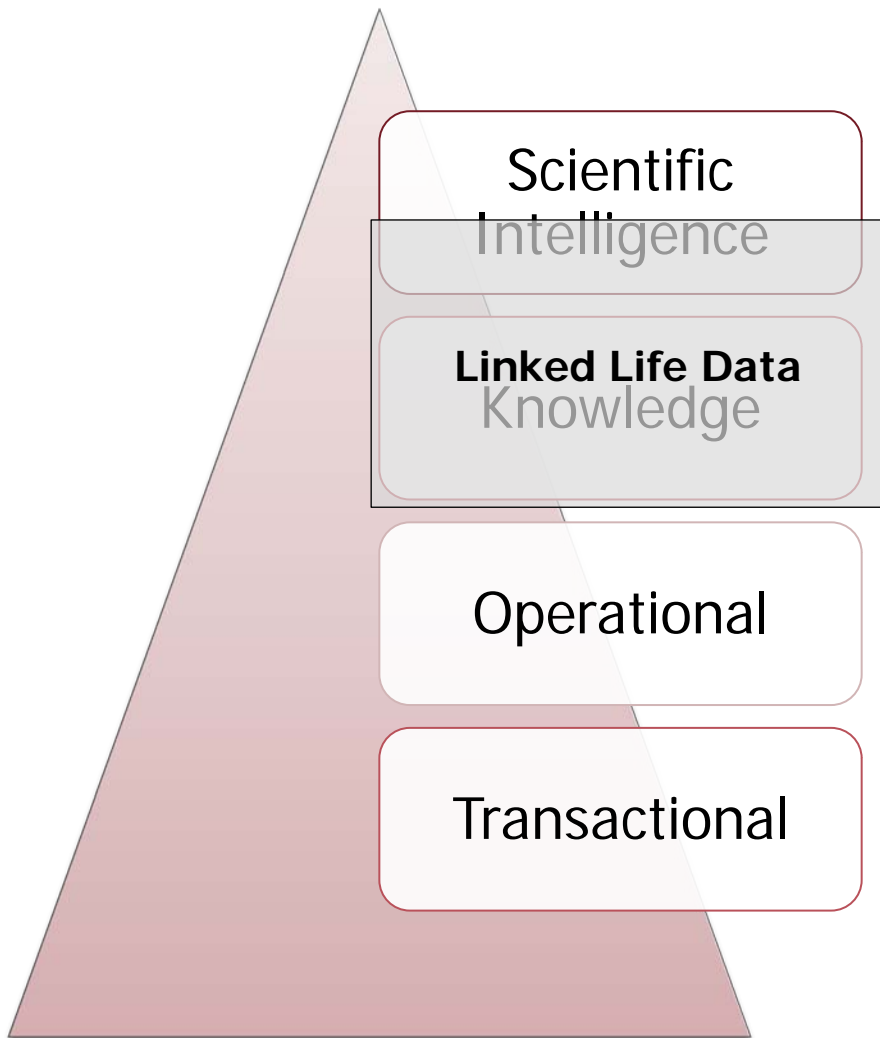
The Original Idea



Data Integration Levels



System Levels in the Knowledge Driven Process



- Advanced visualization and statistical analyzes
- Information extraction
- Schema alignment
- Shared identifiers
- Data silos applications
- Databases
- File system

Syntax and Structure Ambiguity

- RDF data model resolves all syntax level ambiguities
- It helps you express all data in a common data model

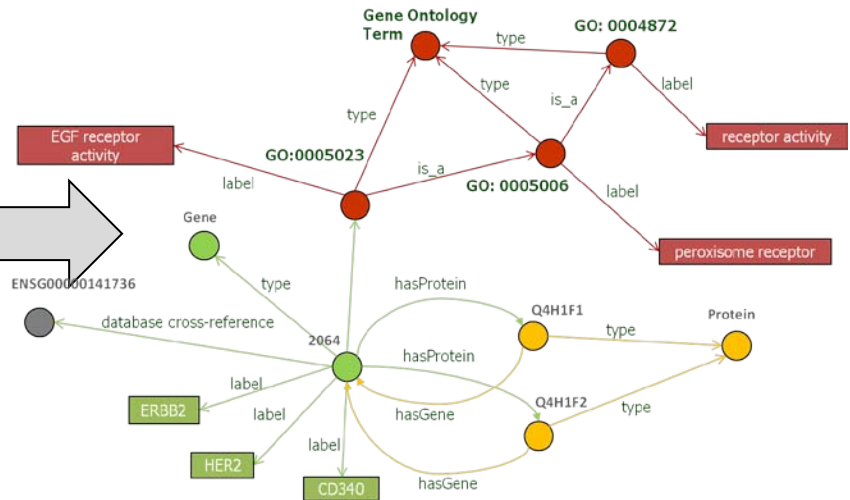
ID GRAA_HUMAN STANDARD; PRT; 262 AA.
AC P12544; DT 01-OCT-1989 (Rel. 12, Created)
DT 01-OCT-1989 (Rel. 12, Last sequence update)

DT 15 <PubmedArticle> <MedlineCitation Owner="NLM"
DE G Status="In-Process"> <PMID
lymph Version="1">21500419</PMID> <DateCreated>
DE 1) <Year
(CTL <Day
DE (F PubM
GZMA IssnT
(Hum <Jou
<Vol
<Pub
<Mor
</Pu

PK.FK1	DATA	DOCUMENTSOURCE
PK.FK2	DATABASE	PK.FK1
PK.FK3	COMPRESSED	PK.FK1
PK.FK4	DOCUMENTKEY	PK.FK1
PK.FK5	DOCUMENTKEY	PK.FK1

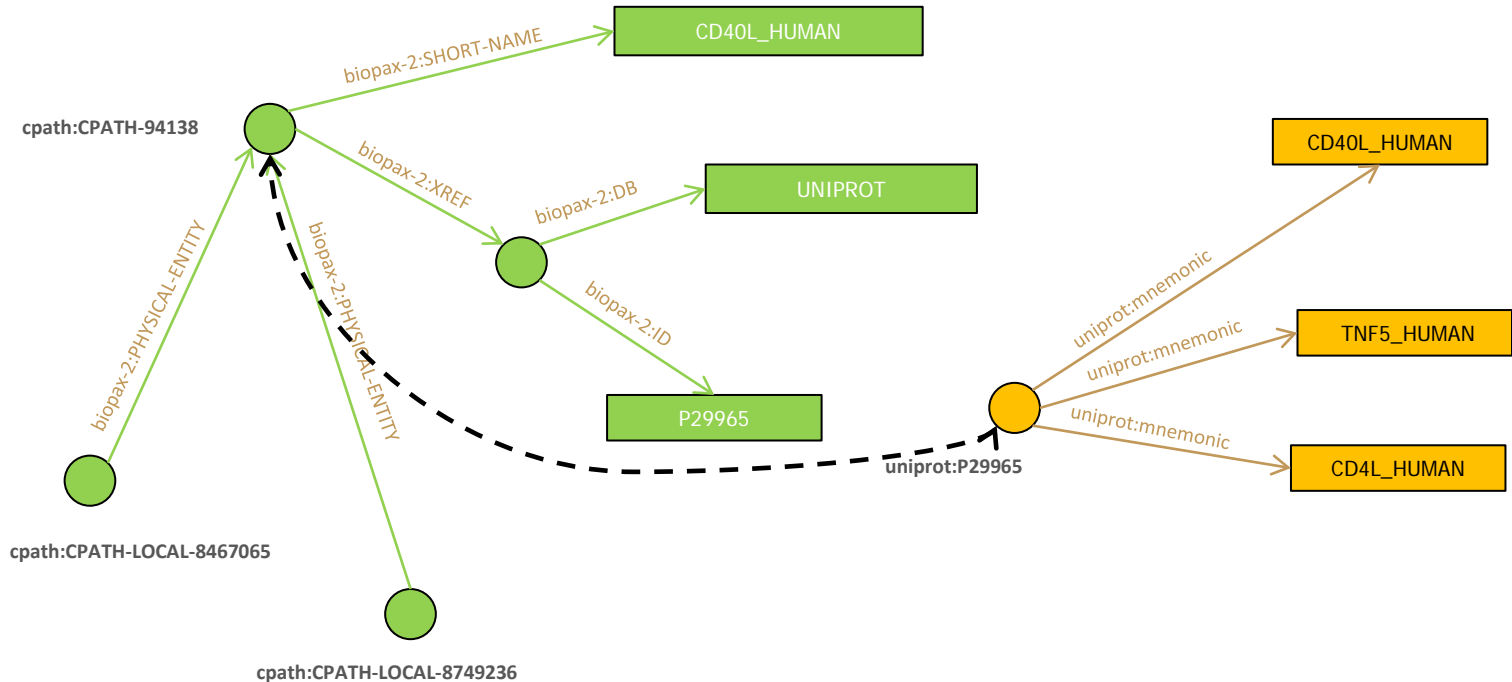
PK.FK1	DOCUMENT
PK.FK2	FRAMES
PK.FK3	STATE
PK.FK4	STEP
PK.FK5	ACTION

PK.FK1	TRANSFORM
PK.FK2	STATUS
PK.FK3	TRANSFORM
PK.FK4	FRAME
PK.FK5	LOGS
PK.FK6	GROUP



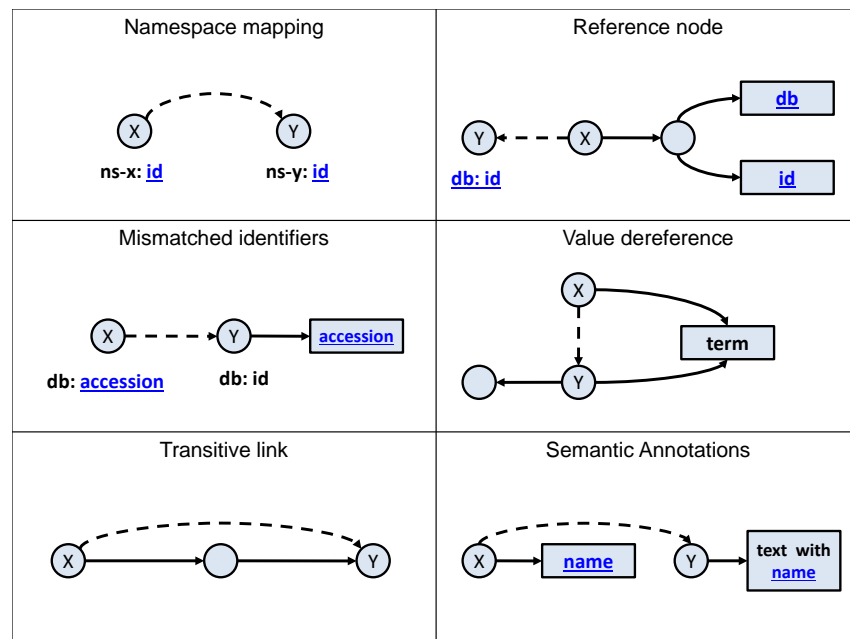
Linked Data Mapping

- 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



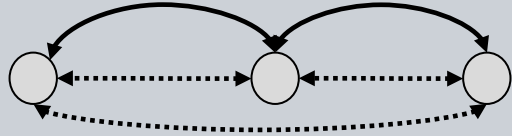
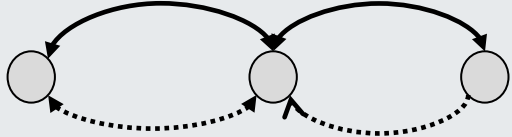
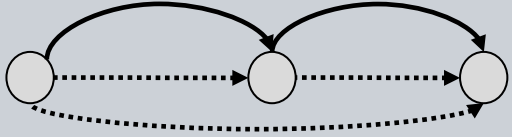

Linked Data Mappings

- 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

Instance Level Identify Alignment

Relationship	Semantics	Example
Exact match	Transitive equivalence	
Close match	Equivalent only for search purposes	
Broader match	Generalization of a concept	
Narrower match	Specialization of a concept	Inverse of broader match
Related	Unspecified relation (no real semantics)	

Quick Facts!

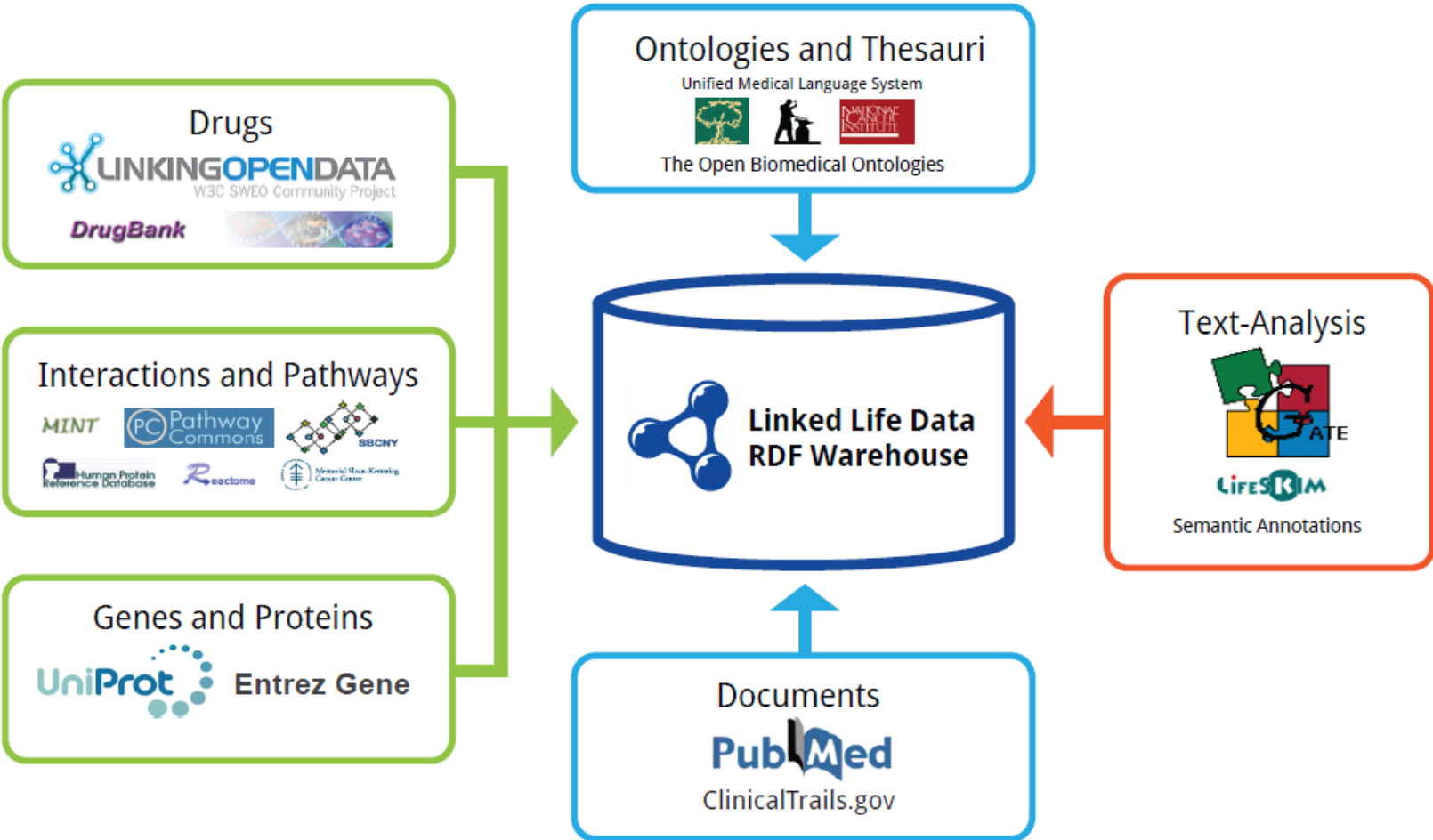
- Public and free RDF warehouse service
- Integrates more than 25 popular data sources
- Apply text mining technology to link the text with entities
- Computer friendly API to access the information



Type of possible questions, analysis and interpretation

INTEGRATED DATASET

Linked Life Data Datasets



Rest API

SPARQL
endpoint

Co-
Occurrence

Relation
Finder

New Type of Possible Query #1

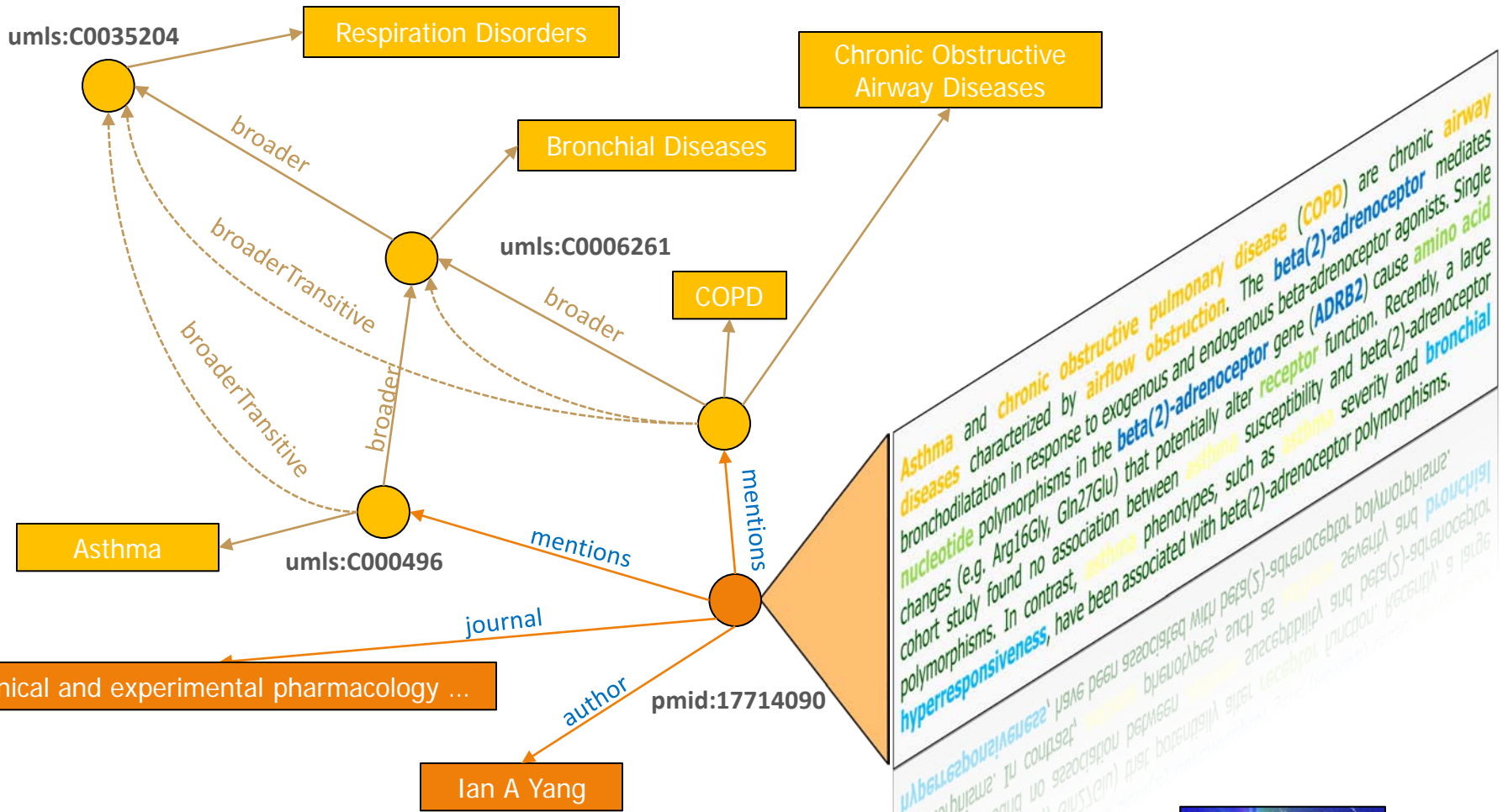
Select drugs related to asthma that are linked to a curated molecular interaction in the literature where the protein is known to cause inflammatory response

```
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.wiwiw.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.

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 a 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.wiwi.wiwi.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)

gene description ▼	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

gene description

- 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

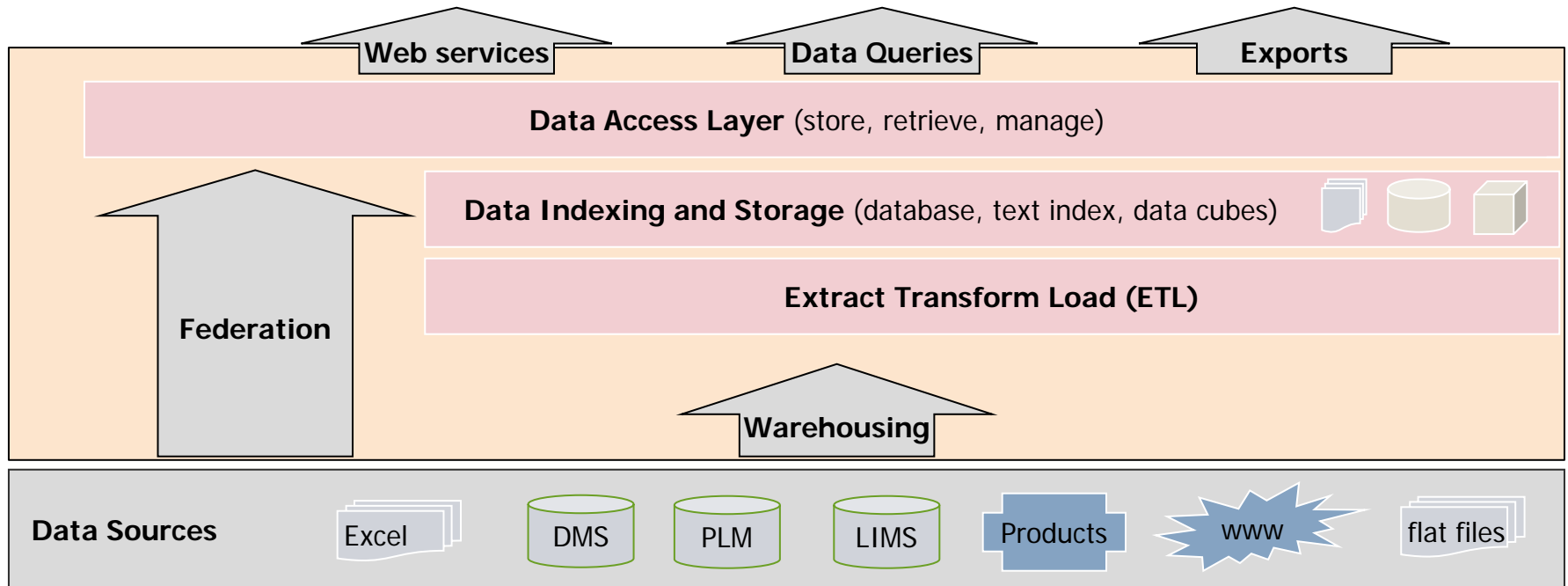
- Lucene based index
- Special predicate to execute full-text queries
- Multiple retrieval modes
 - Literal
 - RDF molecules

```
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/Citatio  
n>.  
  ?article  
<http://linkedlifedata.com/resource/pubmed/article  
Title> ?title.  
} limit 1000
```

Loading procedure, testing environment, used environment

BEHIND THE SCENE






Linked Life Data Architecture



LLD Current Statistics

Repository overview

Repository	
ID:	LLD 0.6.2
Description:	Linked Life Data is a semantic data integration platform for the biomedical domain.
Number of statements:	5,512,293,218
Number of expl. statements:	5,155,779,890
Number of entities:	1,009,497,196

Data source	Named graph	Load date	Number of statements	Instances type
Disease Ontology  Reference License	http://linkedlifedata.com/resource/diseaseontology	21.10.10	144,541	diseaseontology:DiseaseOntologyConcept
LinkedCT  Reference License	http://linkedlifedata.com/resource/linkedct	14.10.10	7,027,372	linkedct:condition
Reactome  Reference License	http://linkedlifedata.com/resource/reactome	24.09.10	698,567	biopax-2:entity
HPRD  Reference License	http://linkedlifedata.com/resource/hprd	24.09.10	1,917,107	biopax-2:entity
DBPedia  Reference License	http://linkedlifedata.com/resource/dbpedia	20.07.10	496,360,473	skos:Concept

Maintain Two Parallel and Independent Processes

- Data source updates and transformation
 - Download the data source
 - Apply ETL script that generates RDF data
 - Load and index the RDF data with a global Entity Pool
 - Do a local inference for the data source graph
- LLD release builds
 - Merge previously loaded repositories
 - Execute post processing instance mappings
 - Do a global inference across all graphs

Node	Id
URI1	1
URI2	2
Literal1	3
URI3	4

Entity Pool

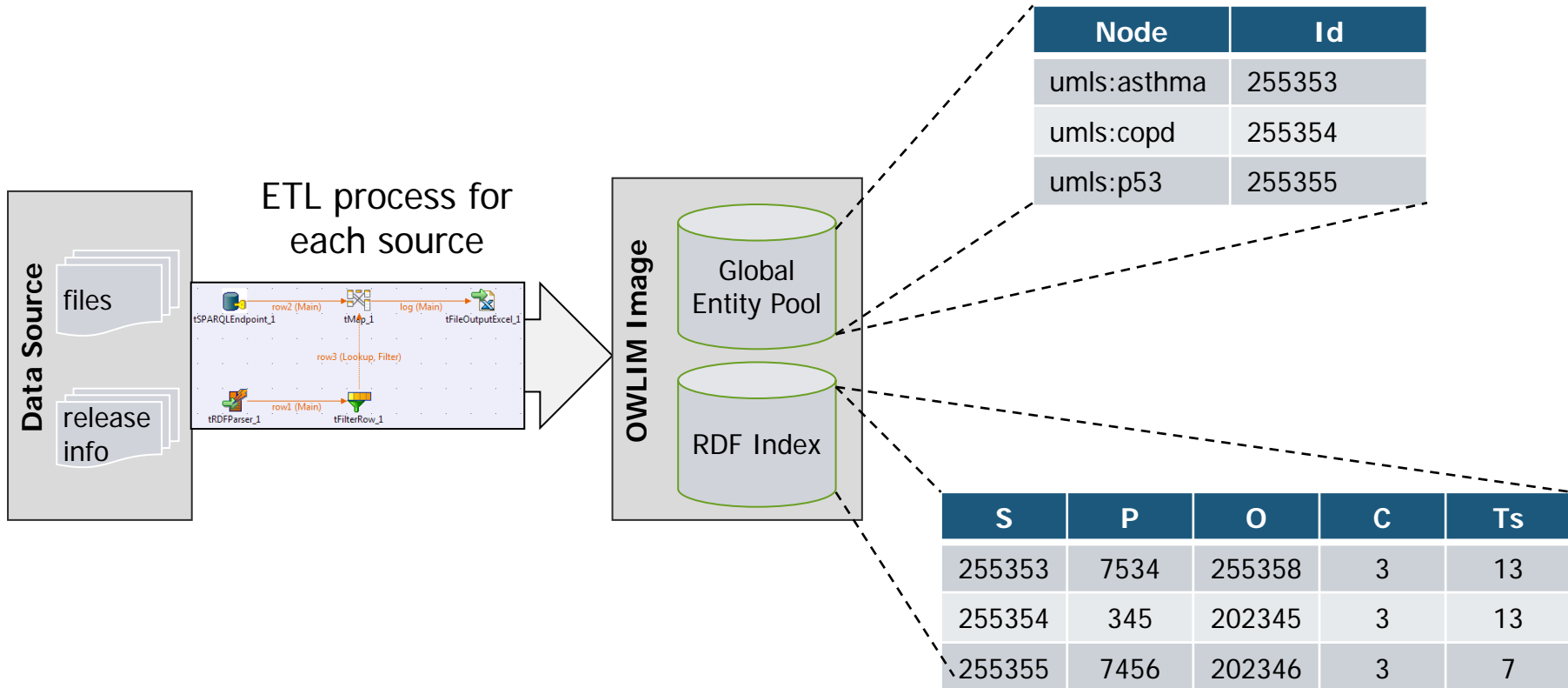
RDF data

S	P	O	C
URI1	URI2	Literal1	URI3
URI1	URI2	Literal2	URI3

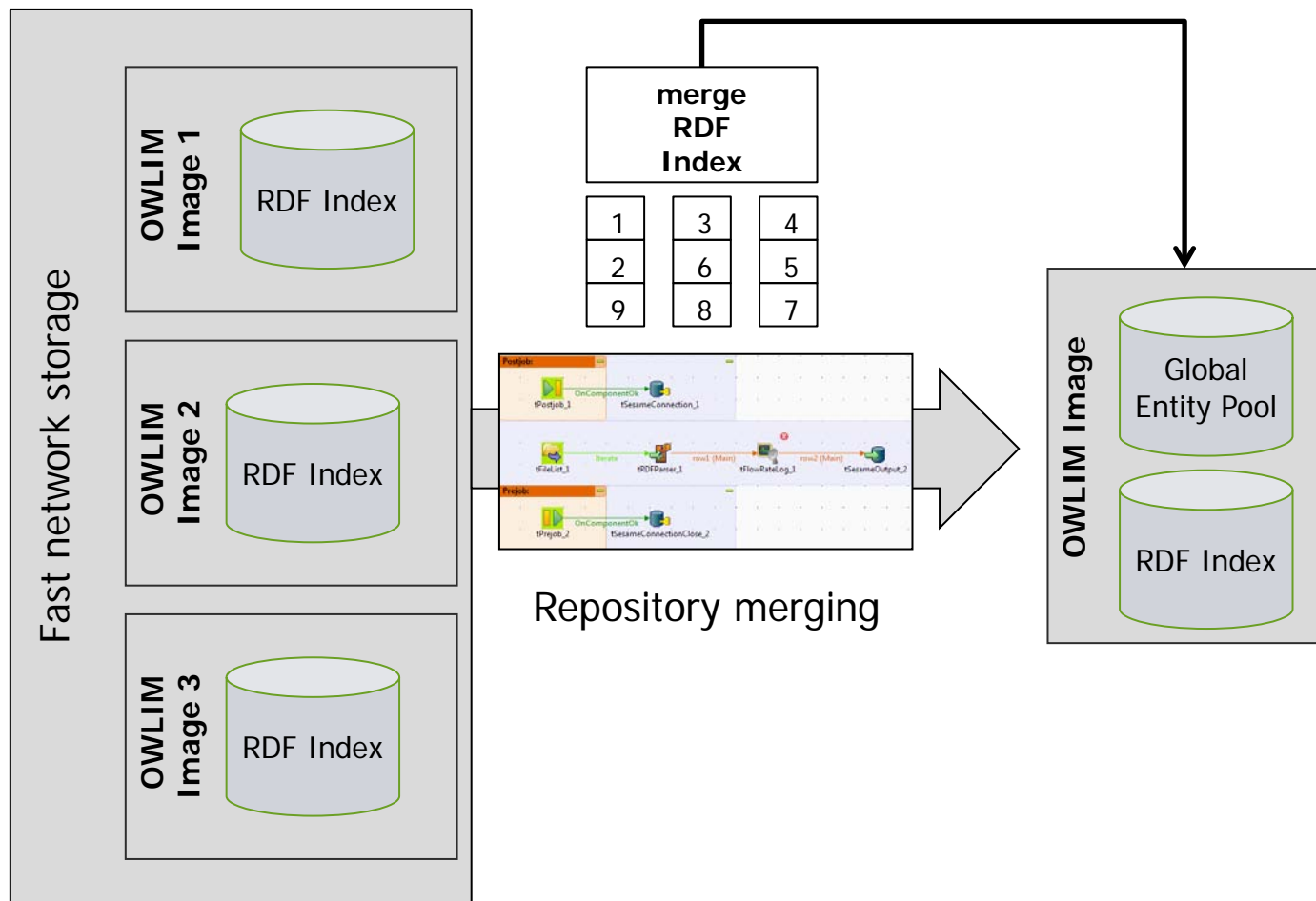
S	P	O	C
255353	7534	255358	3
255354	345	202345	3
255355	7456	202346	3

RDF index

Life Cycle of a LLD Data Source



Combining Arbitrary Data Sources



Advantages of the LLD Approach

- Each data source has a consistent query-able repository
 - All repositories are compatible and could be efficiently combined
 - You can maintain multiple versions of the repository
 - Fixes in the RDF schema are very quick
- The data source updates are absolutely independent from the production releases
 - We can maintain multiple LLD versions optimized for different needs
 - The extension with new data sources is trivial
 - You have the capability to support global reasoning

Wrap-up

- Free and actively developed public service available:
<http://linkedlifedata.com>
- 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\$
- Warehouse methodology that scales for tens of data sources
- Integrate information-extraction algorithms