ontotext

GATE

Module 13
Semantic Technology and Linked Open Data: Basics, Tools, and Applications

Semantic Technology and Linked Data Annotation

ontotext

GATE

About This Tutorial

- Understand RDF, linked data and SPARQL
- See the semantic technology in practice
- Create semantic annotations
- Index and search semantic annotations with MIMIR

ontotext

GATE

Module 13 Outline

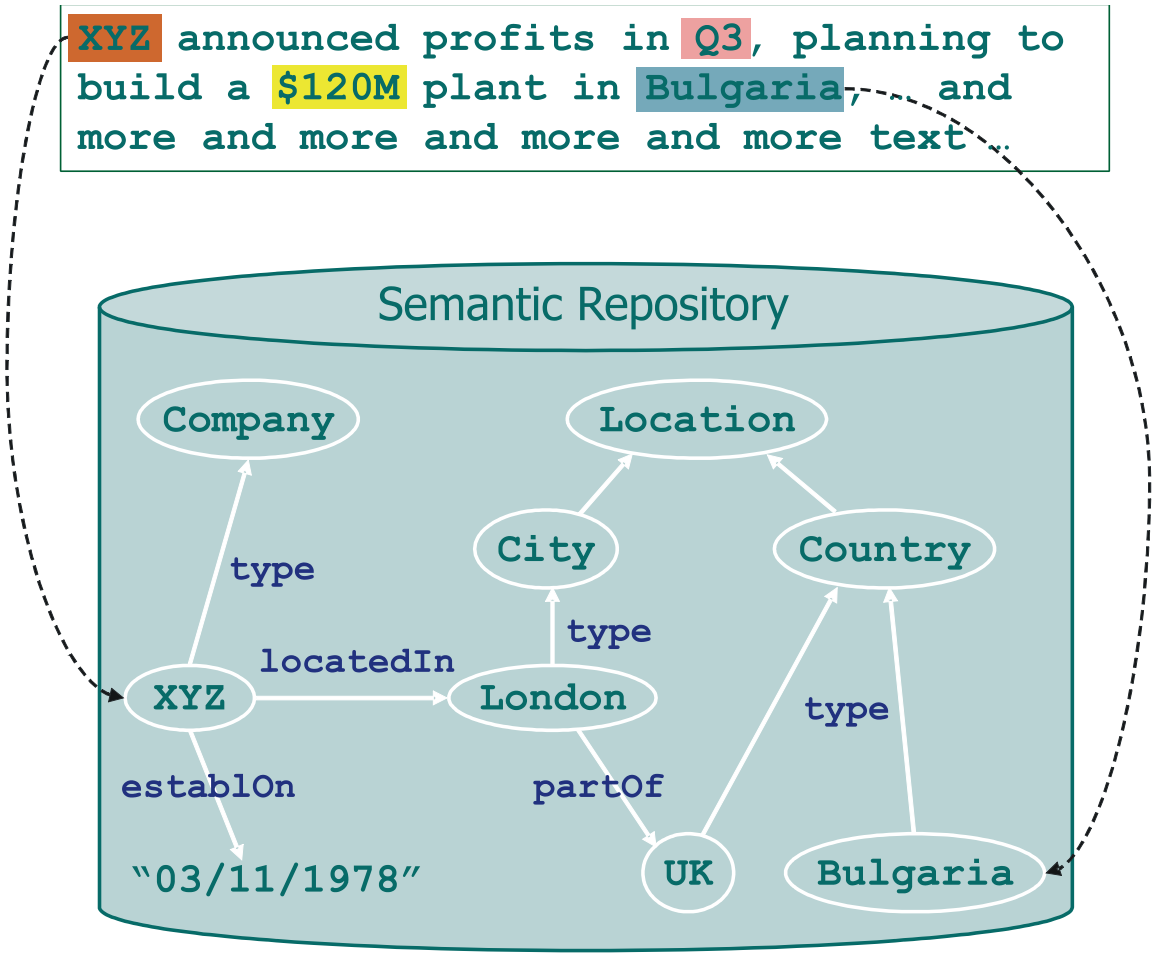
9:00-10.45	1.Introduction - (35 minutes) 2.Processing RDF Data - (35 minutes) 3.Linked Data (30 minutes) 4.SPARQL Query Language (10 minutes)
11:00 – 12:45	5.Query SPARQL Endpoint and Serialize Data (15 minutes) 6.Populate Gazetteer from LLD Endpoint (20 minutes) 7.Semantic Annotations and Linked Data (40 minutes) 8.Query MIMIR and LLD (30 minutes)

ontotext

GATE

Ontotext Company Profile

Interlinking Text and Data



ontotext

GATE

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

ontotext

GATE

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

GATE

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

ontotext

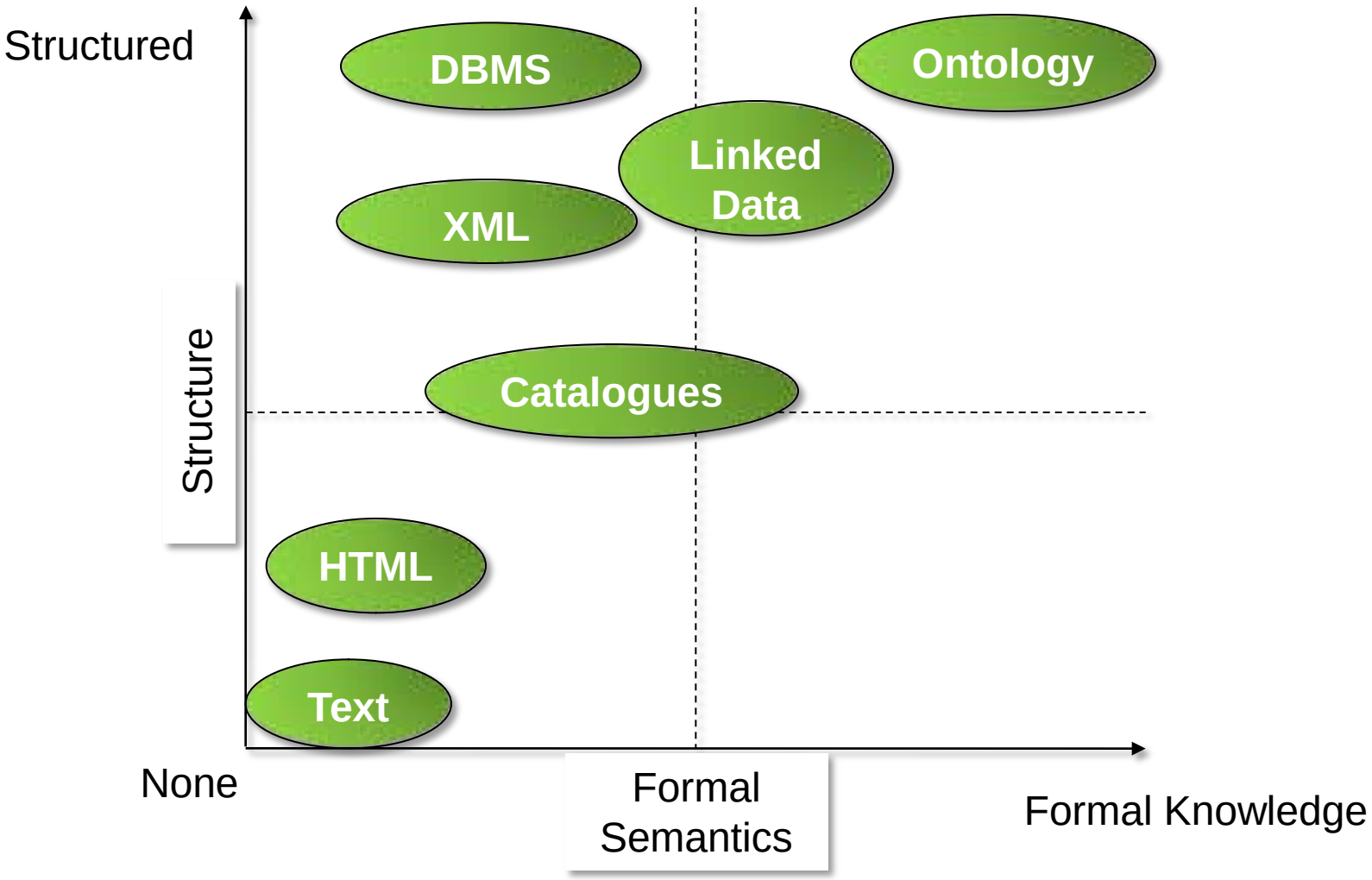
GATE

RDF Introduction

ontotext

GATE

Types of Data



ontotext

GATE

So Why No Just Use XML?

```
<country name="UK">
    <capital name="London">
    <areacode>20</areacode>
    </capital>
</country>
```

```
<nation>
    <name>United Kingdom</name>
    <capital>London</capital>
    <capital_areacode>20
     </capital_areacode>
</nation>
```

No agreement on:
Structure
- is country a:
  - object?
  - class?
  - attribute?
  - relation?
- what nesting mean?

Vocabulary
- is country same as nation?

Are the above XML documents the same?
Do they convey the same information?
Is that information machine-accessible?

ontotext

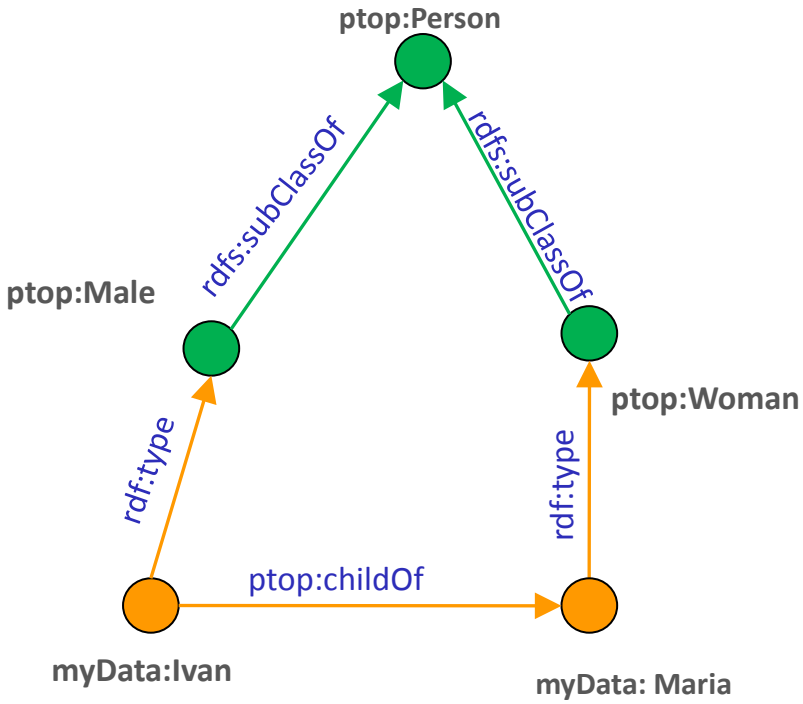
Resource Description Framework

GATE

- A simple data model for
  - describing the semantics of information in a machine accessible way
  - representing meta-data (data about data)
- A set of representation syntaxes
  - XML (standard) but also N3, Turtle, …
- Building blocks
  - Resources (with unique identifiers)
  - Literals
  - Named relations between pairs of resources (or a resource and a literal)

Data representation: XML vs. RDF

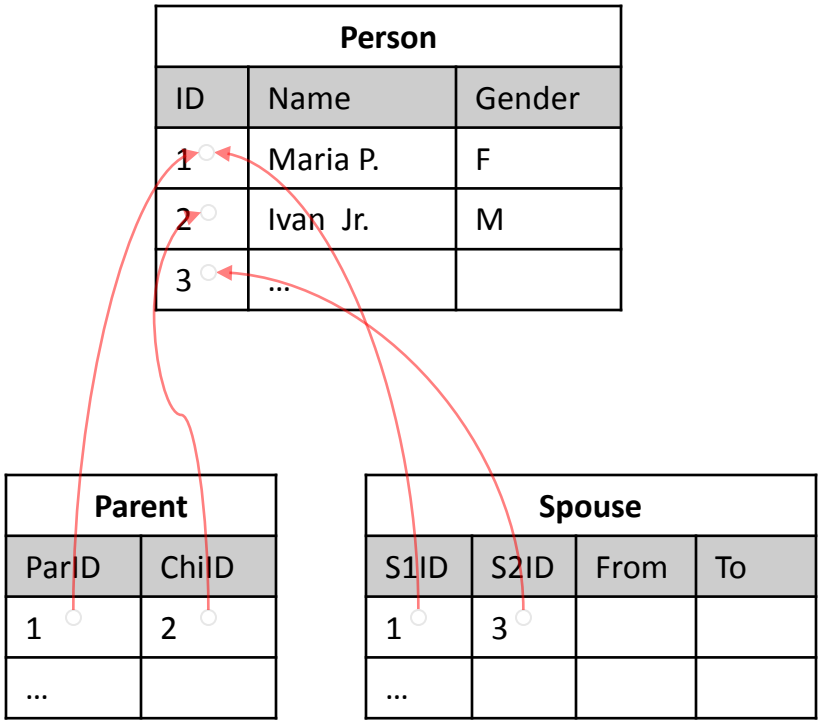
```
<document>
   <person>
        <name>Maria</name>
        <gender>F</gender>
        <relList>
             <rel type="child">Ivan</rel>
       <relLiist>
  </person>
```



XML Documents

RDF Representation

Data representation: RDBMS vs. RDF



Person		
ID	Name	Gender
1	Maria P.	F
2	Ivan Jr.	M
3	…	

Parent	
ParID	ChiID
1	2
…	

Spouse			
S1ID	S2ID	From	To
1	3		
…			

Relational Tables

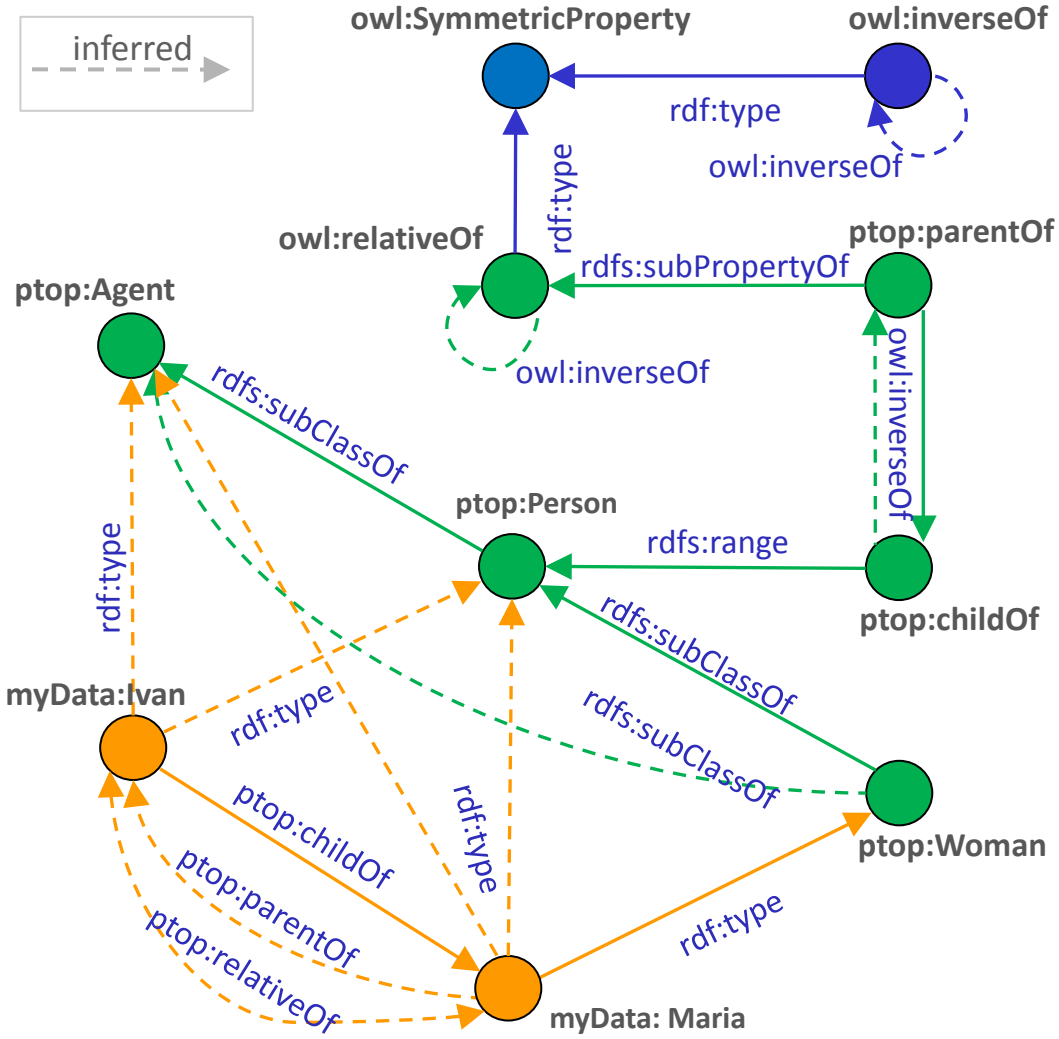
Statement		
Subject	Predicate	Object
myo:Person	rdf:type	rdfs:Class
myo:gender	rdfs:type	rdfs:Property
myo:parent	rdfs:range	myo:Person
myo:spouse	rdfs:range	myo:Person
myd:Maria	rdf:type	myo:Person
myd:Maria	rdf:label	“Maria P.”
myd:Maria	myo:gender	“F”
myd:Maria	rdf:label	“Ivan Jr.”
myd:Ivan	myo:gender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

RDF Representation

ontotext

GATE

RDF Graph



ontotext

GATE

Processing RDF data

ontotext

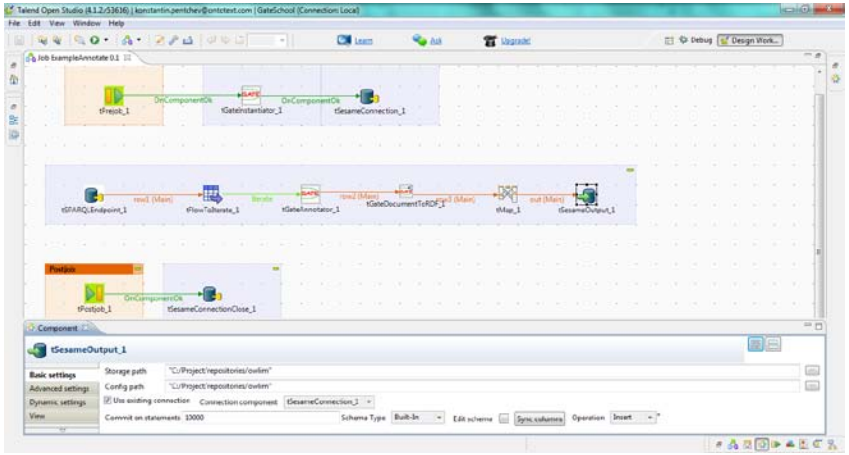
GATE

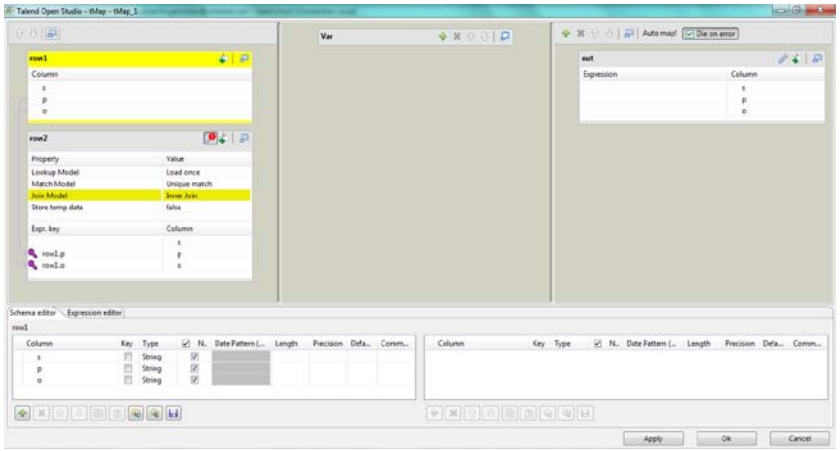
Hands-on Sessions Today

- There are 5 practical examples to be completed
- Hands-on could be downloaded from:
  - ....
- The used software is
  - Sesame (LGPL)
  - Gate Developer (LGPL)
  - MIMIR (GPL v2)
  - OWLIM (Commercial, free for research)
  - Linked Life Data service (free and public)
  - Talend Open Studio (GPL v2)

TalenD Open Studio

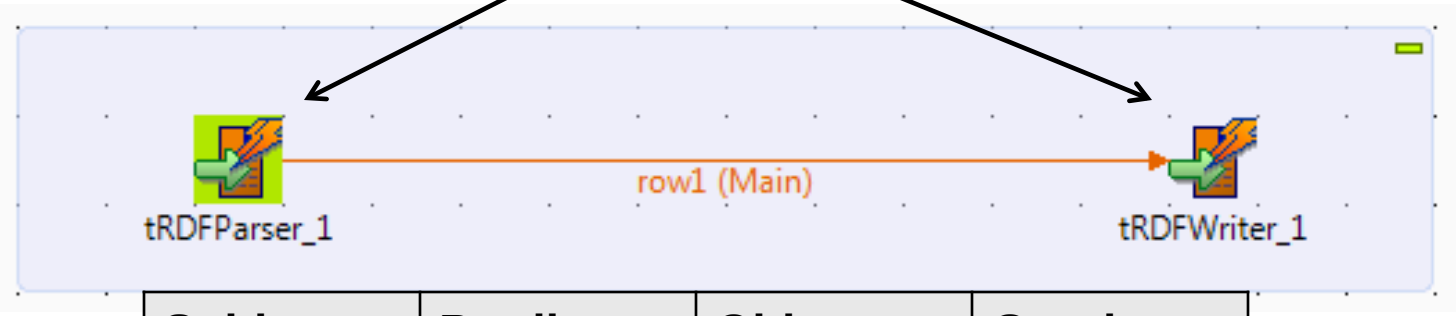
- IDE for development of data transformation jobs
- No programming skills are required
- All used software is integrated as components
- The task will be to select, configure and connect components





TalenD Open Studio Basics

components



Subject (string)	Predicate (string)	Object (String)	Graph (String)
…	…	…	…

schema

data

data type

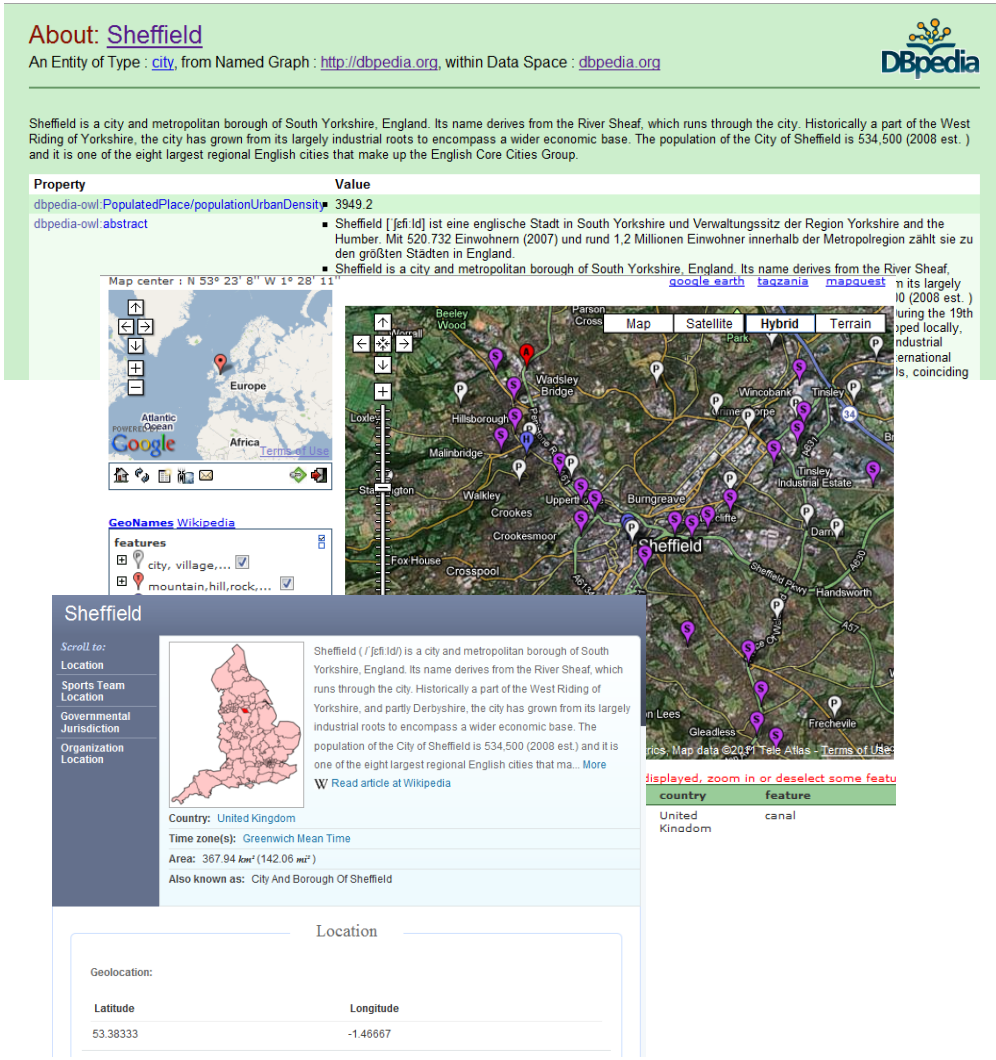
ontotext

GATE

Linked Data

The Web Of Data

- Give to all entities in a data source an URL
- Give to all relationships in a data source an URL
- Link between items in different data sources
- Link between terms from different vocabularies



ontotext

GATE

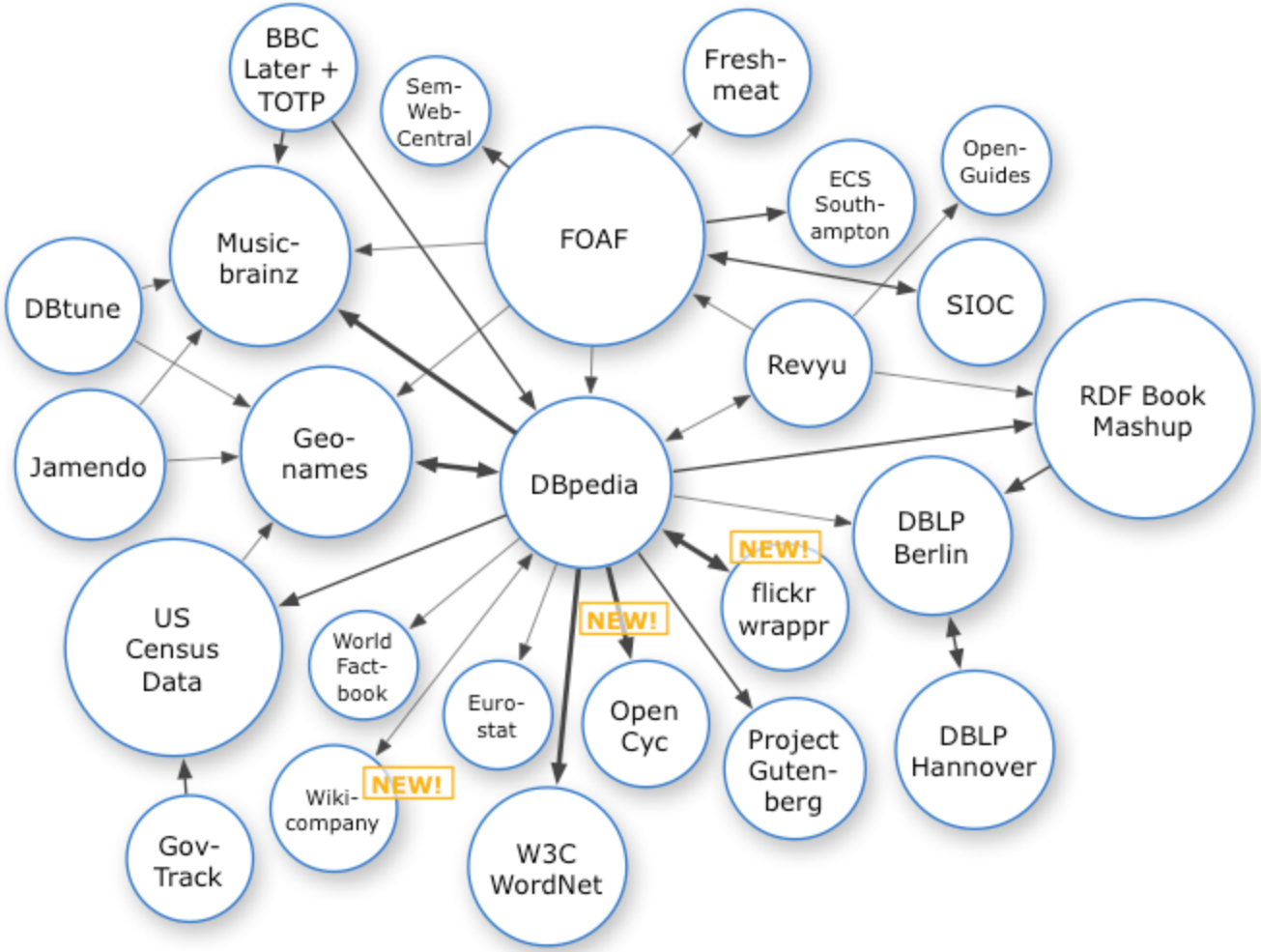
Datasets

- DBpedia
  - Linked Data version of Wikipedia
  - 3.5 million entities, incl. 410K places, 310K persons, 146K species, 140K organisations, 95K music albums, 50K films, 33K buildings, 15K videogames, 5K diseases
  - Descriptions available in 90 languages
  - **1 billion** triples, 10 million links to external RDF datasets
  - Ontology – 260 classes, 1200 properties, 1.5 million instances
    - http://www4.wiwiss.fu-berlin.de/dbpedia/dev/ontology.htm

ontotext

GATE

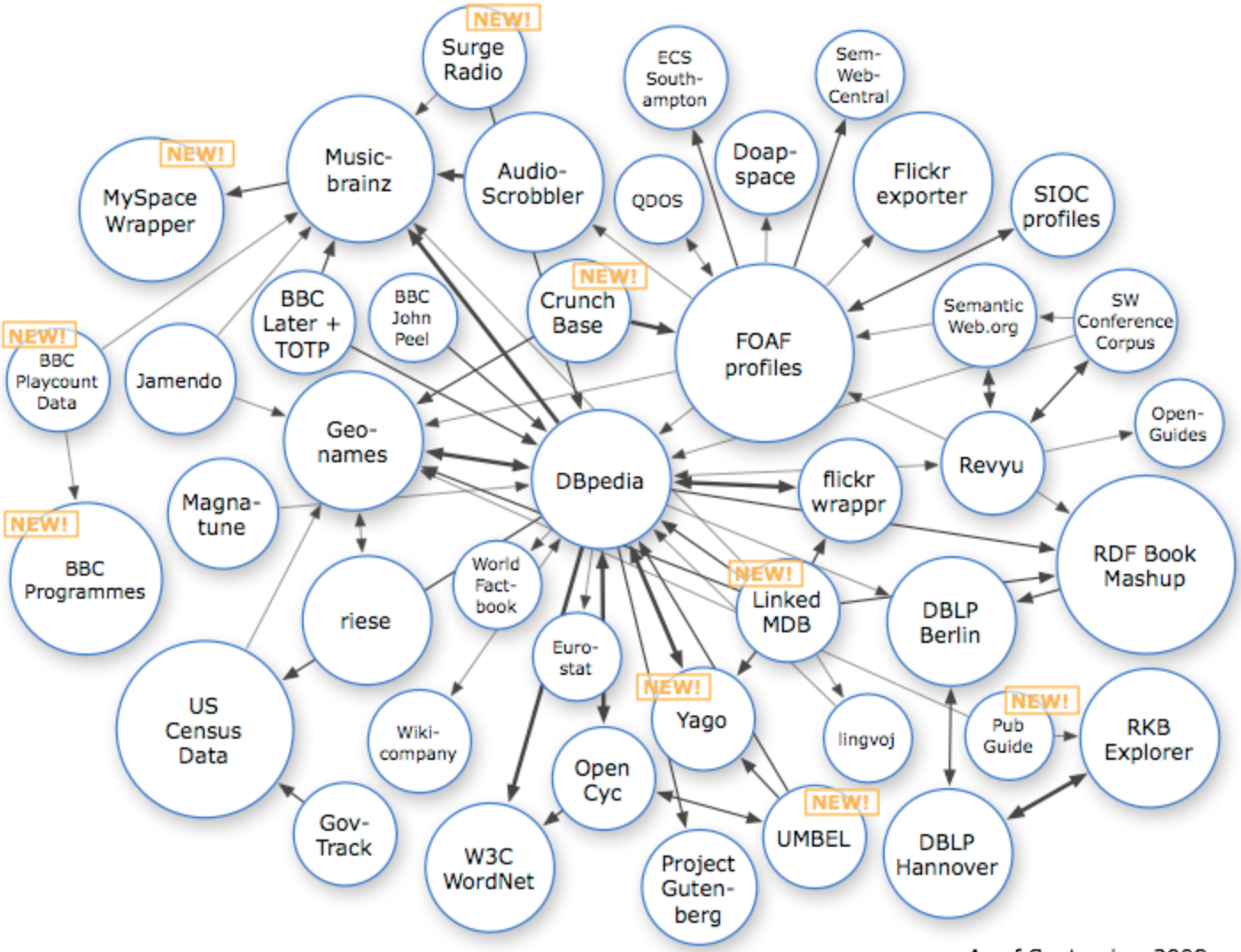
Linked Data Evolution – Oct 2007



ontotext

GATE

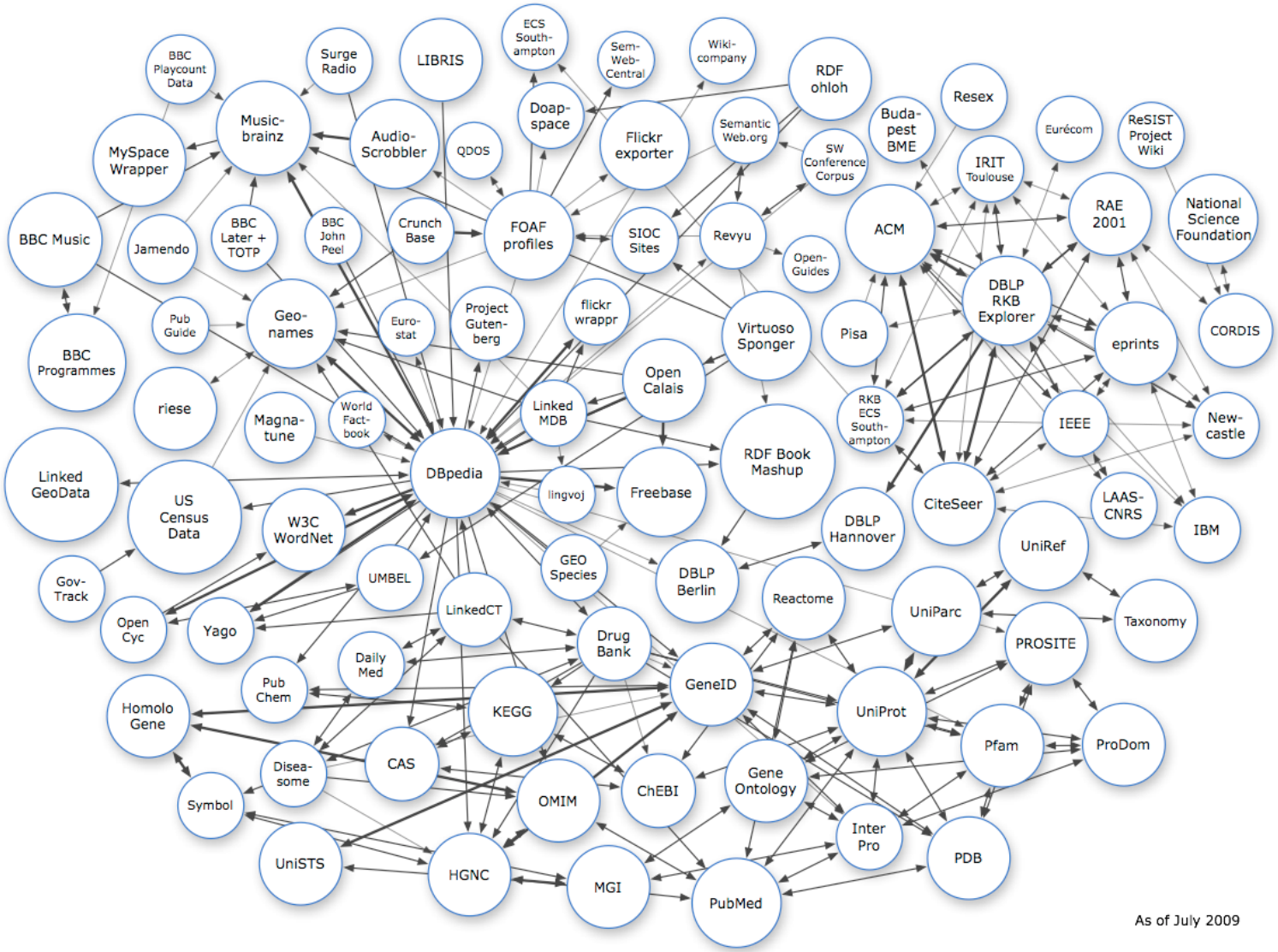
Linked Data evolution – Sep 2008



ontotext

GATE

Linked Data evolution – Jul 2009



As of July 2009

ontotext

GATE

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 is easy to discover information about an object (resource)
  - When someone lookups a URI, provide useful information
- Link the object (resource) to related objects
  - Include links to other URIs

ontotext

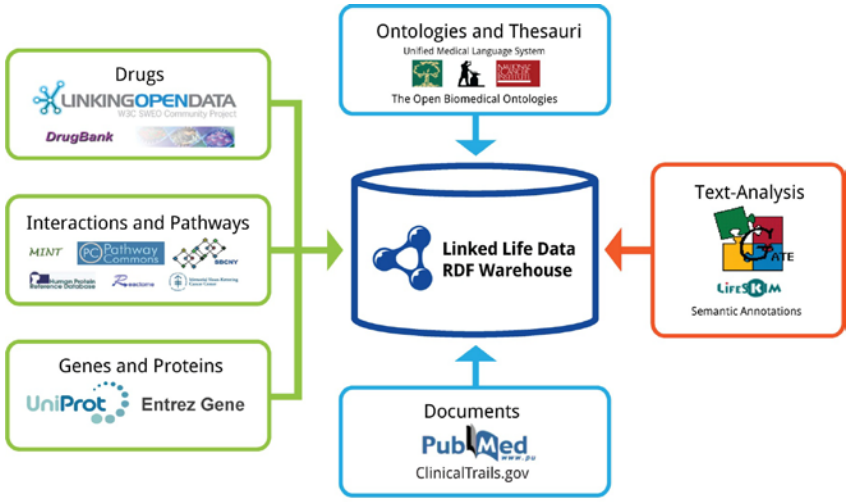
GATE

Reason-able Views to the Web of Data

- *Reason-able views* represent an approach for reasoning and management of linked data
  - Integrate selected datasets and ontologies in one dataset
    - Clean up, post-process and enrich the datasets if necessary
  - Load the compound dataset in a single RDF repository
  - Perform inference with respect to tractable OWL dialects
  - Define sample queries against the integrated dataset

Linked Life Data Service

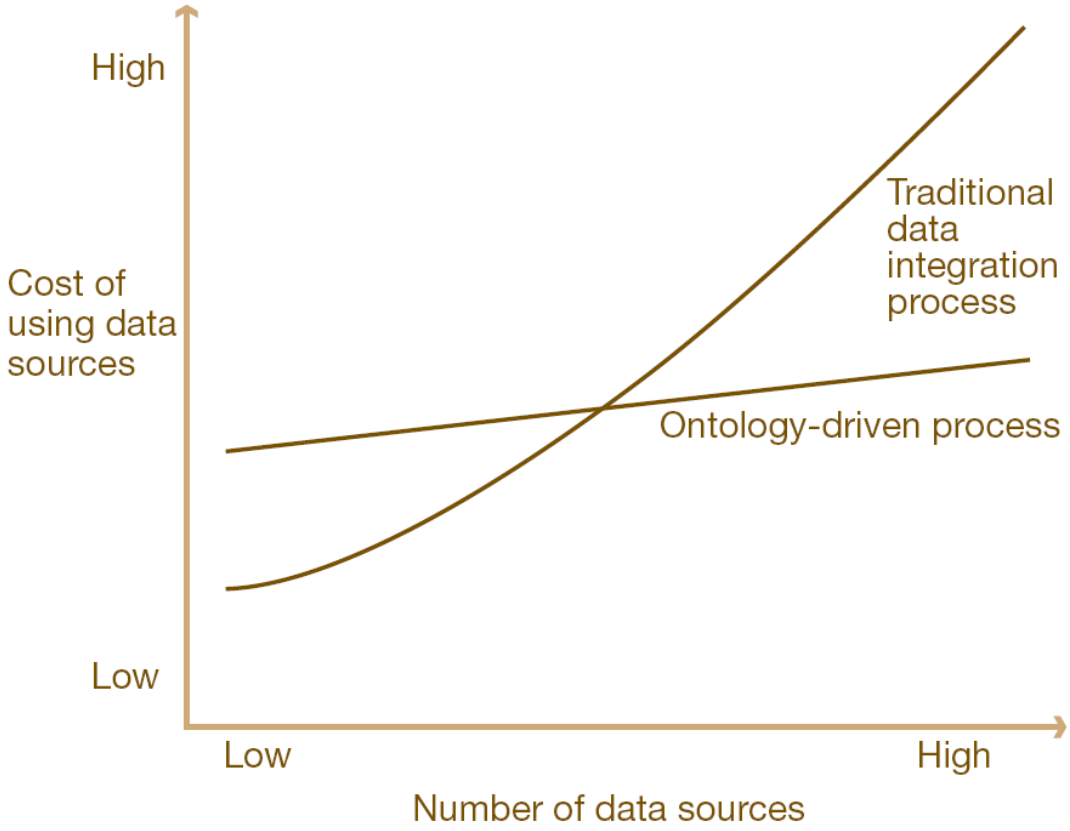
- Linked Life Data is a public RDF warehouse service
- Integrates more than 25 popular biomedical data sources
- Specifies many cross data sources semantic mappings
- Exposes massive amounts of linked data



PWC on Semantic Technologies

Spring of the data Web

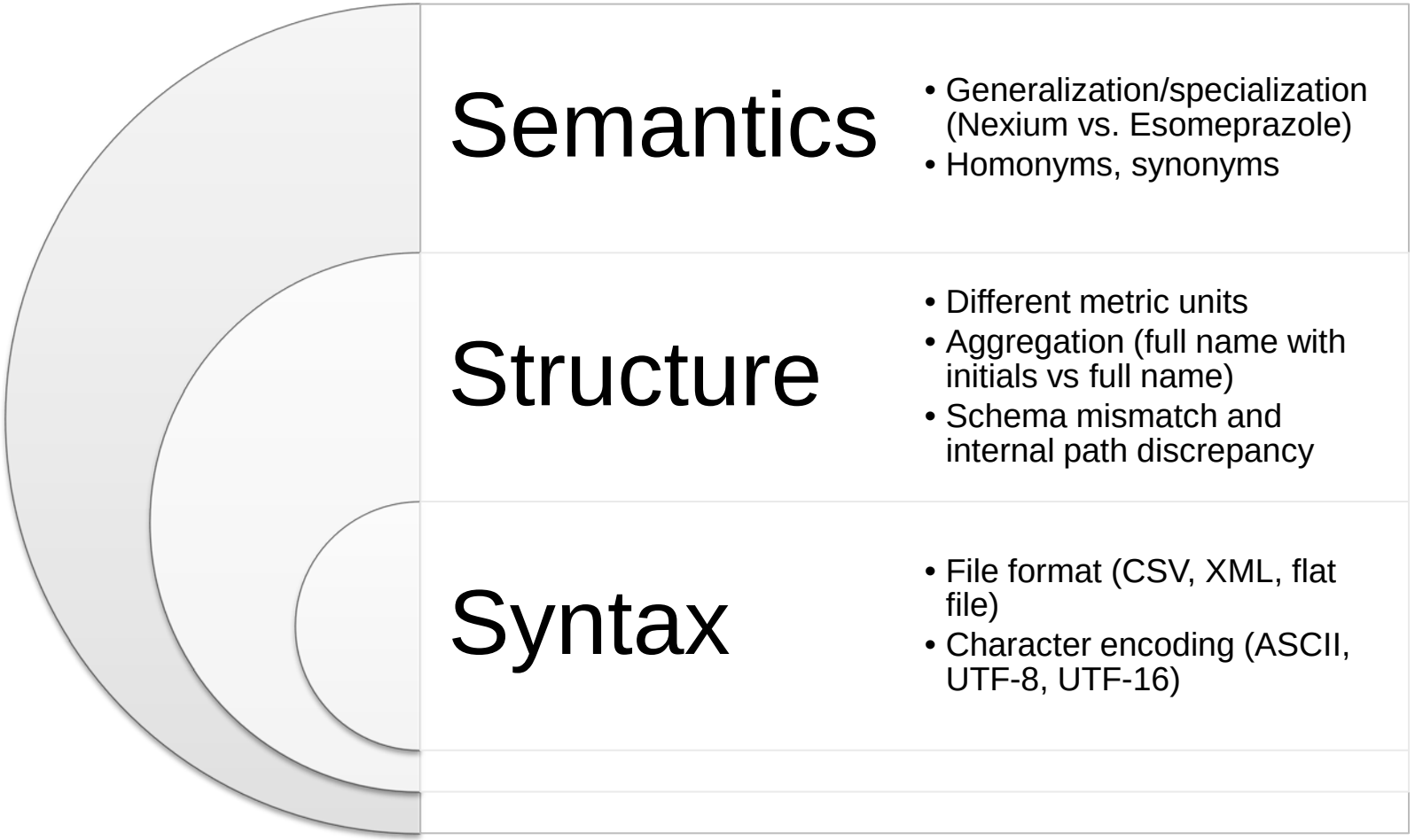
Technology forecast, A quarterly journal, Spring 2009,
http://www.pwc.com/techforecast/



ontotext

GATE

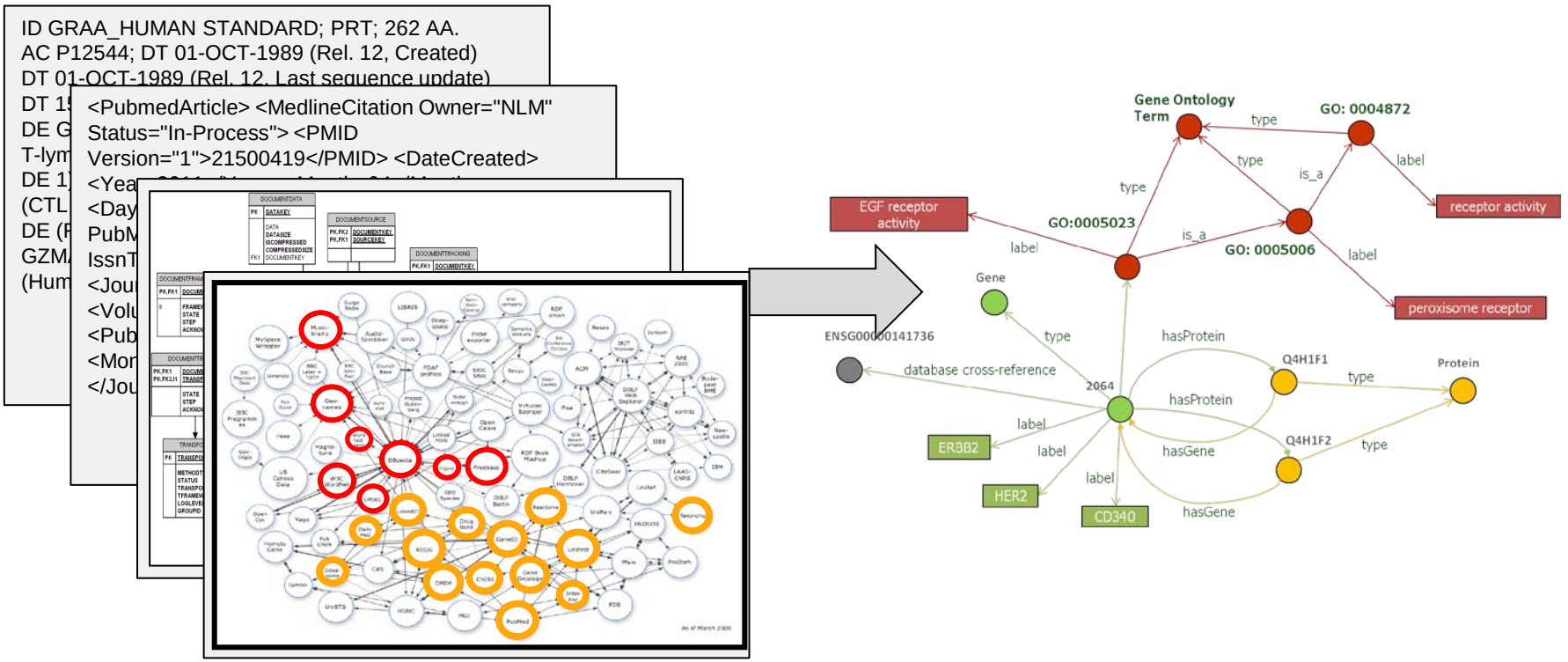
Data Integration Levels



Syntax and Structure Ambiguity

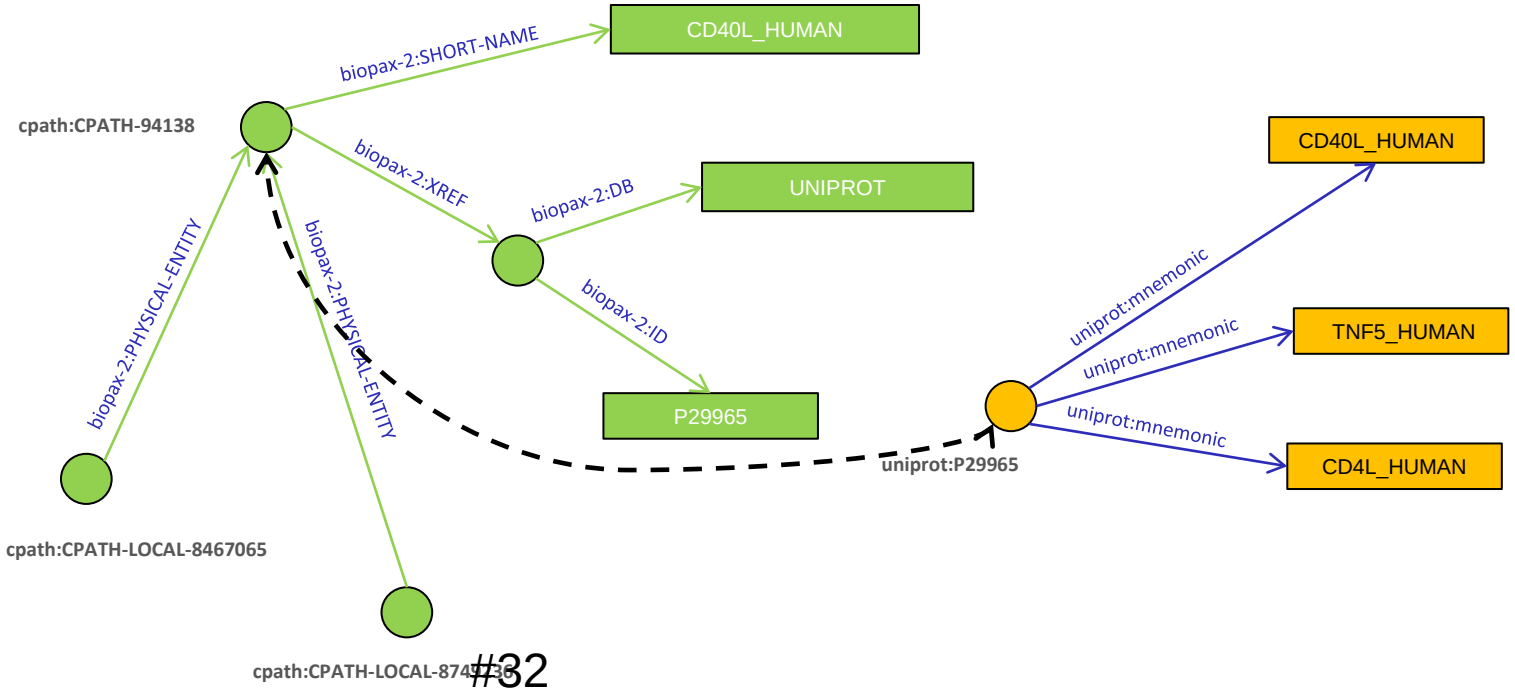
- RDF data model resolves all syntax level ambiguities

It helps you express all data in a common data model



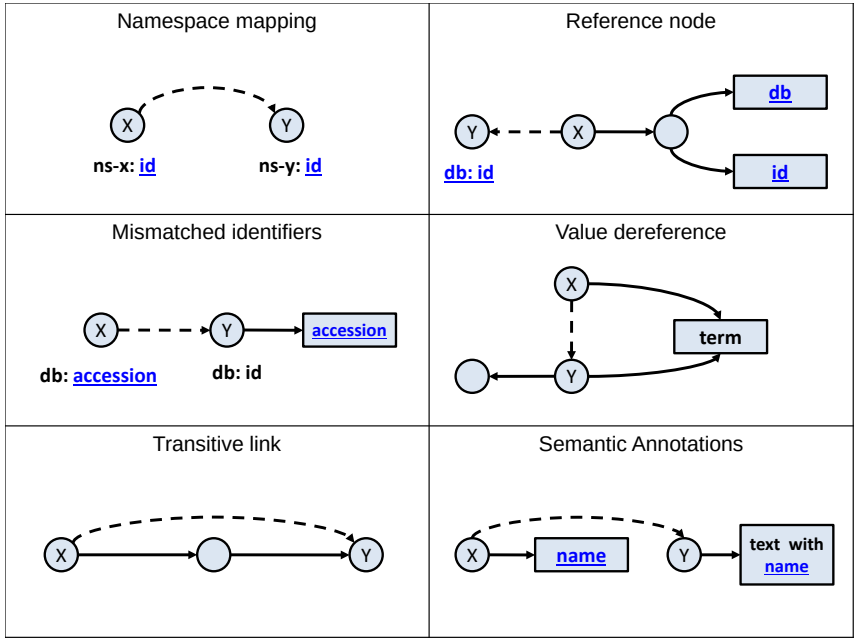
Linked Data Mapping

- 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



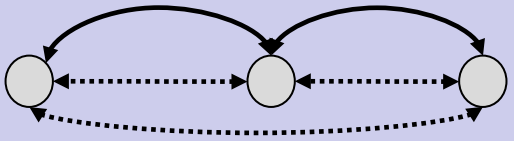
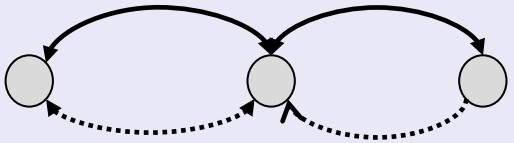
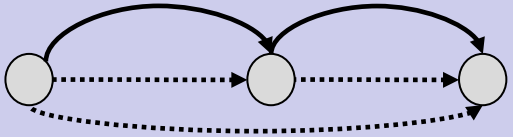
Linked Data Mappings

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



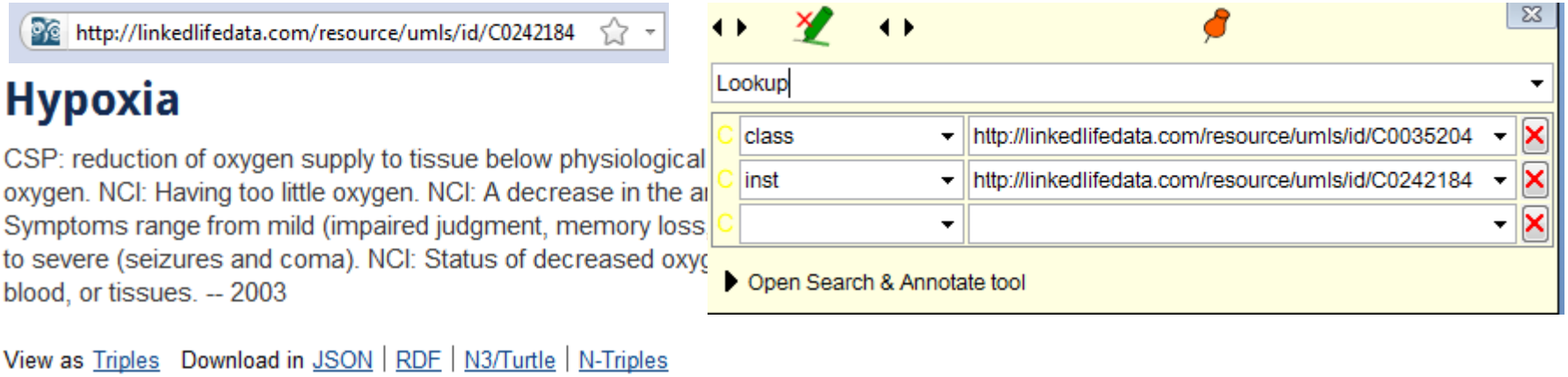
The blue lines and the blue text of the captions (used either as part of the URI or literals) designate the criteria for linking the information

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)	

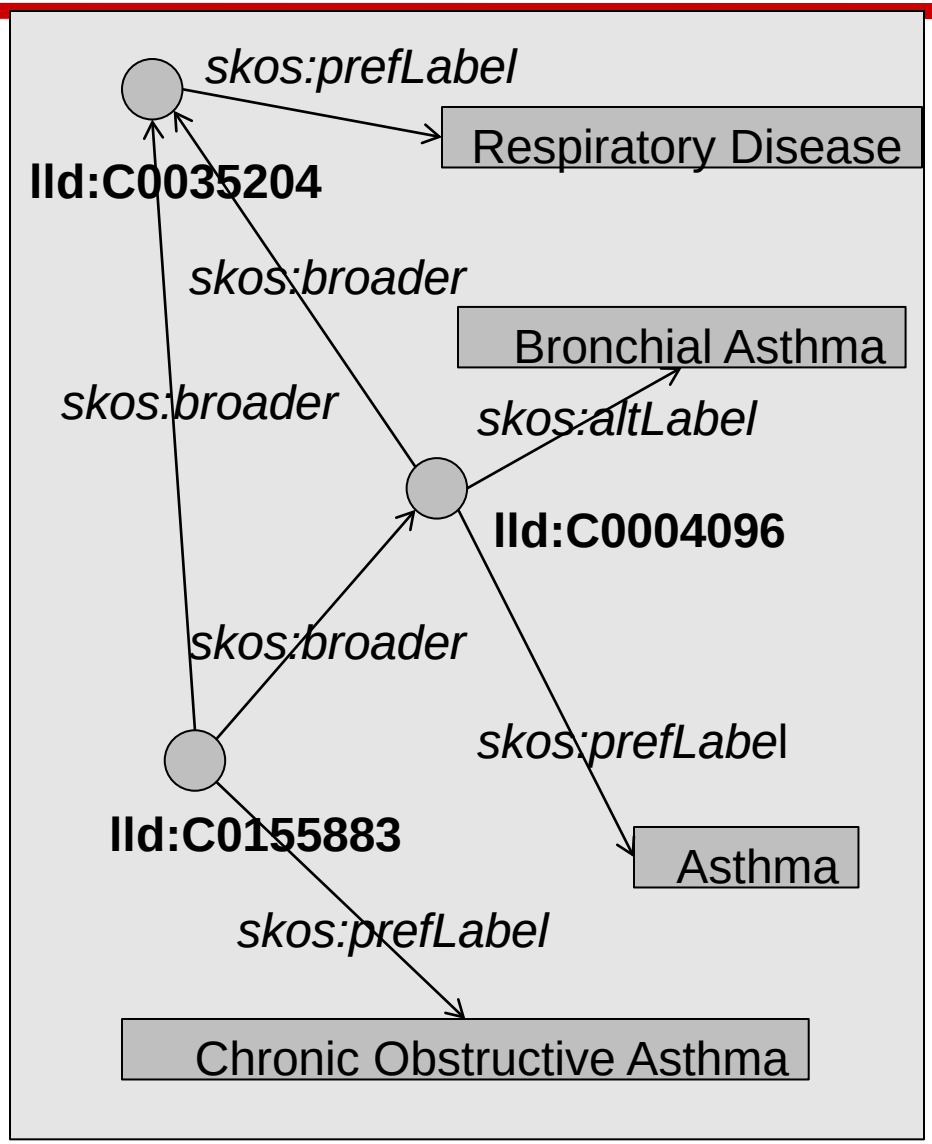
What is the Relation with Text Mining?

The molecular basis for the renal compensation to respiratory acidosis and specifically the role of pendrin in this condition are unclear. Therefore, we studied the adaptation of the proximal tubule and the collecting duct to respiratory acidosis. Male Wistar-Hannover rats were exposed to either hypercapnia and hypoxia [8% CO(2) and 13% O(2) (hypercapnic, n = 6) or normal air (controls, n = 6)] in an environmental chamber for 10 days and were killed under the same atmosphere.



Simple Knowledge Organisation Schema (SKOS)

- SKOS is a common linked data vocabulary
- Serialized as RDF graph
- Published on the web in a to be shared between applications
- Efficient structuring of terms in thesauri



ontotext

GATE

SPARQL Query Language

ontotext

SPARQL Protocol and RDF Query Language (SPARQL)

GATE

- SQL-like query language for RDF data
- Simple protocol for querying remote databases over HTTP
- Query types
  - *select* – projections of variables and expressions
  - *construct* – create triples (or graphs) based on query results
  - *ask* – whether a query returns results (result is true/false)
  - *describe* – describe resources in the graph

ontotext

GATE

Anatomy of a SPARQL a SELECT query

- List of namespace prefixes
  - PREFIX xyz: <URI>
- List of variables
  - ?x, $y
- Graph patterns + filters
  - Group, alternative, optional
- Modifiers
  - ORDER BY, DISTINCT, OFFSET/LIMIT

Querying SKOS Data

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX lld: <http://linkedlifedata.com/resource/>

SELECT DISTINCT ?label ?concept ?top
WHERE {
    ?top skos:prefLabel "Respiration Disorders".
    ?concept skos:broader ?top.
    ?concept skos:inScheme lld:umls.
    ?concept rdfs:label ?label.
}
```

Namespace prefixes

concept with this name

child concepts

part of UMLS

all their synonyms

Return all “Respiration Disorder” concepts in LLD and all their IDs and labels