ontotext
Semantic Technology Lab

Web 3.0: a Network of Semantically Linked Data

Digital Spaces Living Lab

July, 2009

Presentation Outline

- Introduction
- Semantic Databases
- Linked Data
- Reason-able views
- Priming RDF Graphs

What is Ontotext?

- Ontotext is a **semantic technology provider**
- Established in year 2000 as **part of Sirma Group**
  - Sirma is a top-3 software house in Bulgaria, est. 1992, ~300 persons
- **Staff: 40** employees in Sofia and Varna
  - Multiple affiliates and contractors in Western Europe
- Over 150 person-years invested in product development
- Investment acquired in July 2008
  - A financial investor obtained minority share in a deal for 2.5 MEURO
- Ontotext is involved in two joint ventures:
  - **Innonvantage**: online recruitment intelligence provider in UK
  - **Namerimi**: national search engine in Bulgaria

Ontotext Positioning

- Unique technology portfolio:
  - **Semantic Databases**: high-performance RDF DBMS, scalable reasoning
  - **Semantic Search**: text-mining (IE), Information Retrieval (IR)
  - **Web Mining**: focused crawling, wrapping
  - **Knowledge fusion**: identity resolution, record linkage
  - **Web Services and BPM**: WS annotation, discovery, etc.
- **Core business:** core technology development
  - Mostly product development and sales
  - Complemented by professional services
  - Joint ventures for vertical solutions
  - LifeSKIM: solutions for specific domain

Extensive Involvement in Research Projects

- Ontotext has participated in 20+ EC research projects
- > **100 MEuro** is the budget of the projects Ontotext is part of
  - This is above 10% of the EC projects related to semantics
- Ontotext is the **most successful Bulgarian company** in FP6
- Ontotext is part of four **FP7** projects, running until 2011:
  - LARKC**: web-scale reasoning**
  - soa4all: **SOA for the masses** through Semantic Web technology
  - Notube: semantics for personalized **TV guides**
  - Insemtives: **Incentive models** and framework for semantic metadata
- Total income for 2009-2011: **1.4 MEuro**
- **New project proposals:** 2 (subm.Q2), 4 (to be subm. Q4)

Semantic Technologies

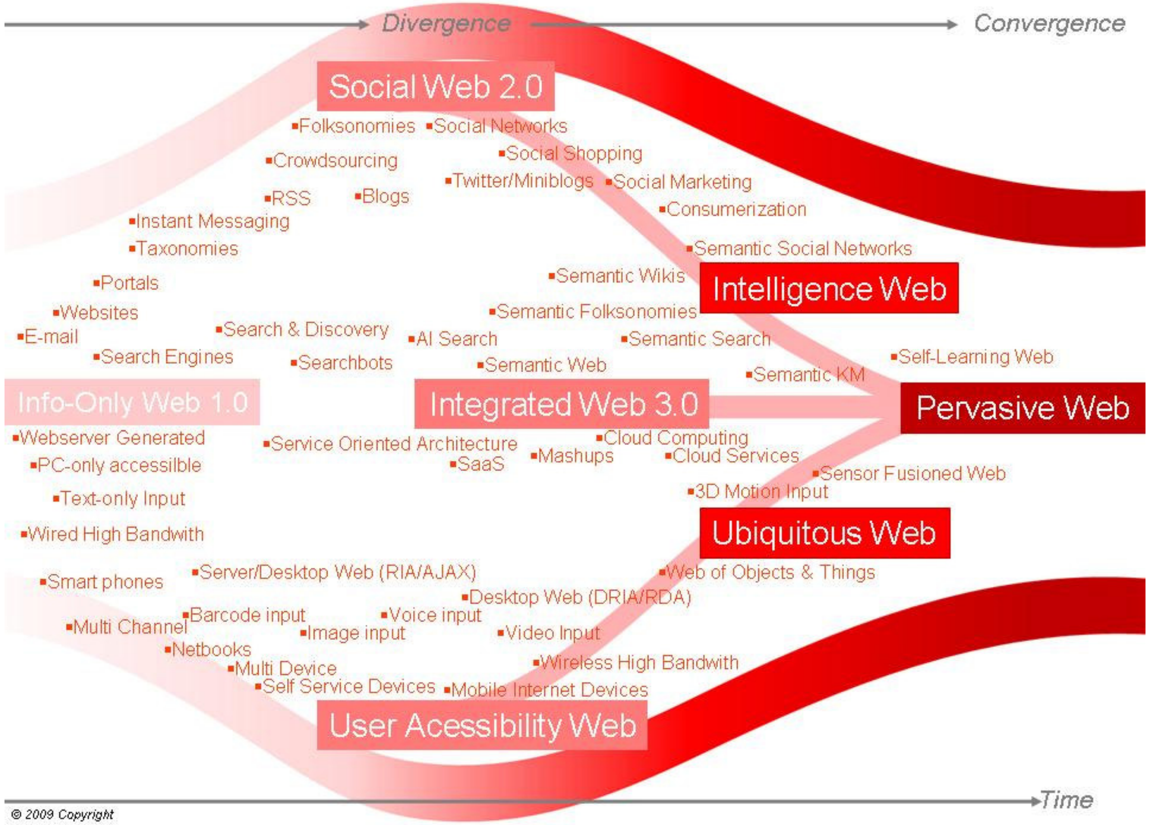
- **“Semantic technologies” (ST)** is a general term for any software, which involves **some kind and level of understanding** of the meaning of the information it deals with
- Examples:
  - A search engine, which can match query for *“bird”* with document mentioning *“eagle”*
  - A database that will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

Accenture on Semantic Technologies

Evolution of the Web

Michael Widjaja, Partner and Senior Executive

July 1st, 2009, http://www.accenture-blogpodium.nl/2point0/michael-widjaja/evolution-of-the-web/



PWC on Semantic Technologies

Spring of the data Web

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

“Semantic Web technologies could revolutionize enterprise decision making and information sharing.”

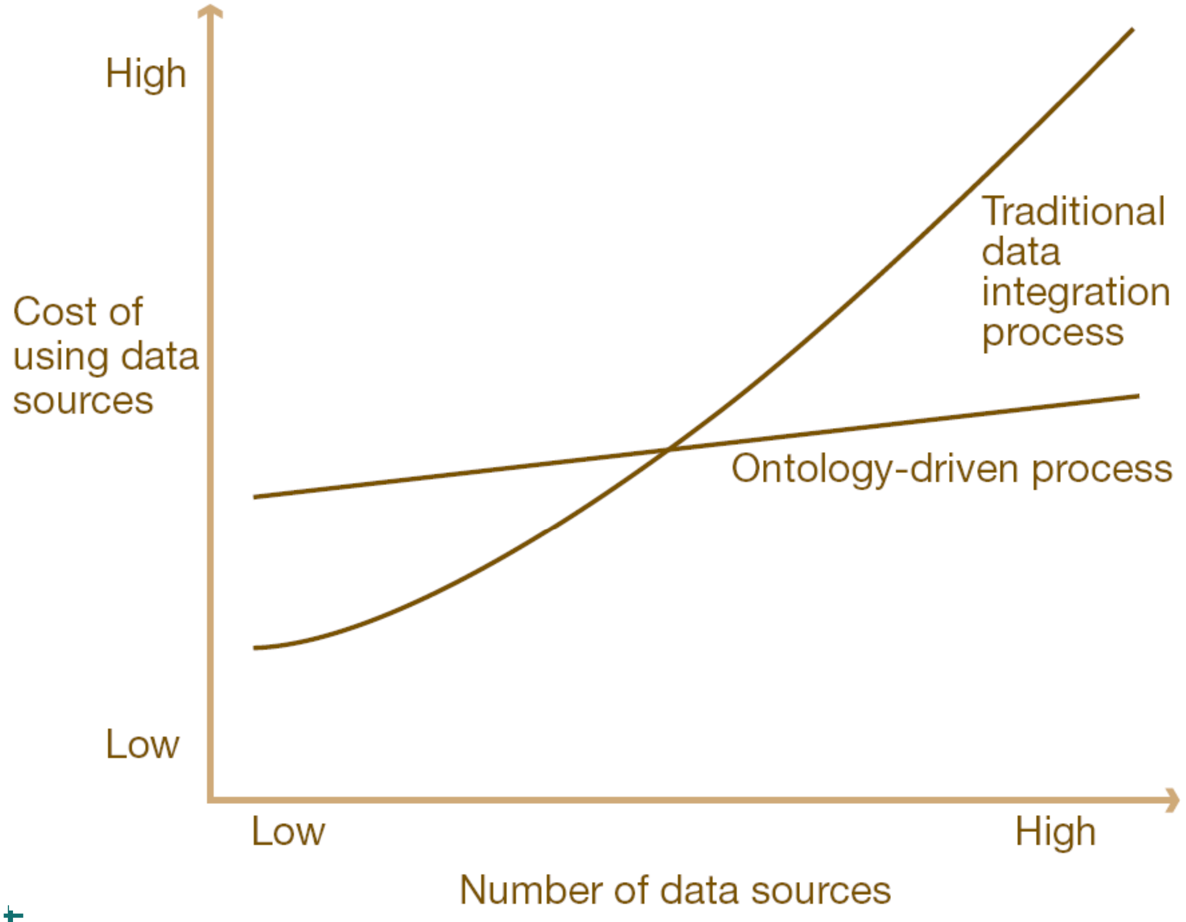
“PricewaterhouseCoopers believes a Web of data will develop that fully augments the document Web of today.

*You’ll be able to find and **take pieces of data sets from different places**, aggregate them without warehousing, and **analyze them in a more straightforward, powerful way than you can now**. …; the underlying technology also applies to internal information and non-Web-based external information. In fact, it can **bridge data from anywhere—including your data warehouse and your business partners**”*

PWC on Semantic Technologies (2)

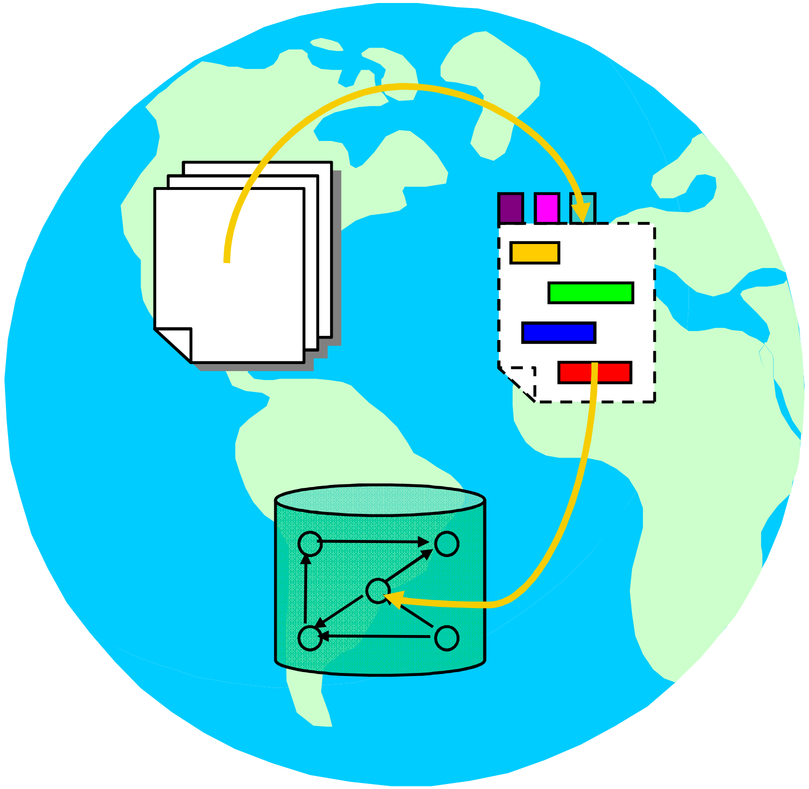
Spring of the data Web

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

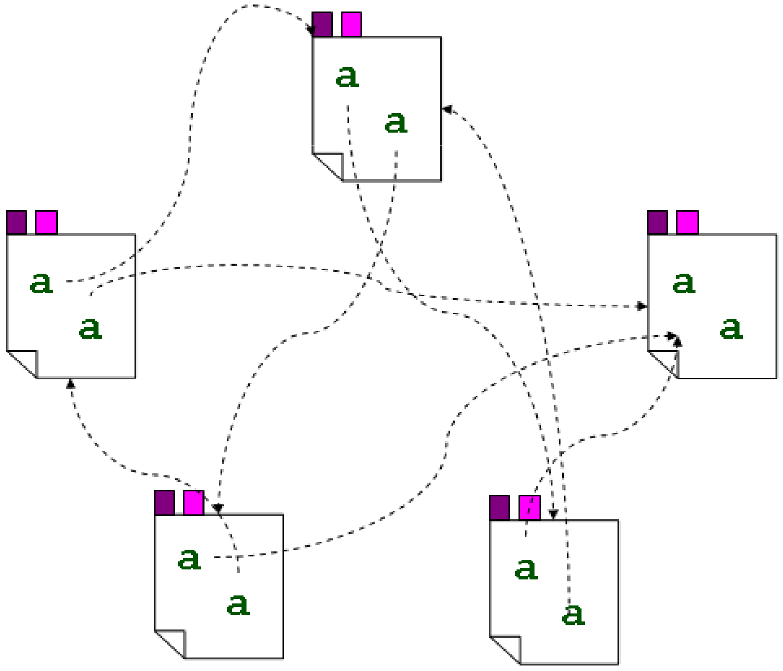


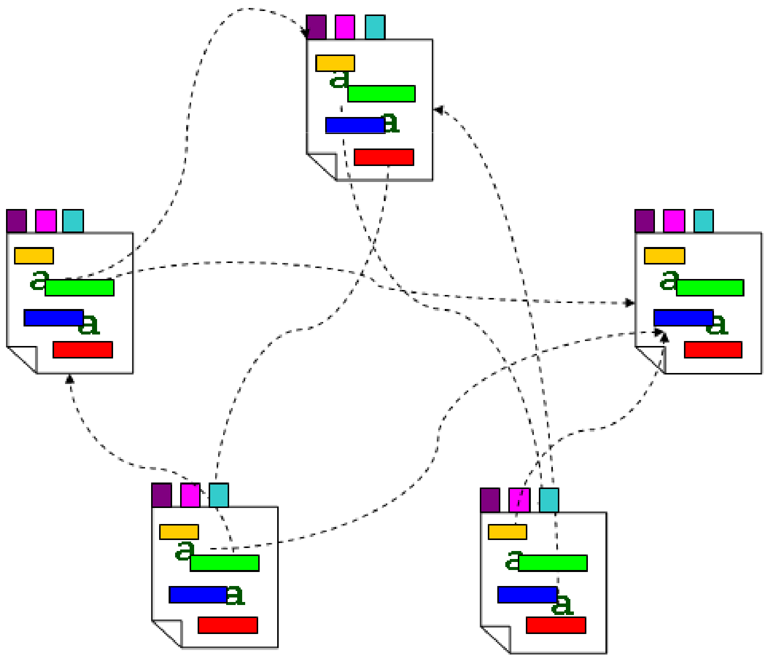
Semantic Web

- "The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." [Berners-Lee et al. 2001]
- The spirit:
  - Automatically processable metadata regarding:
    - the structure (syntax) and
    - the meaning (semantics)
    - of the content.
  - Presented in a standard form;
  - Dynamic interpretation for unforeseen purposes

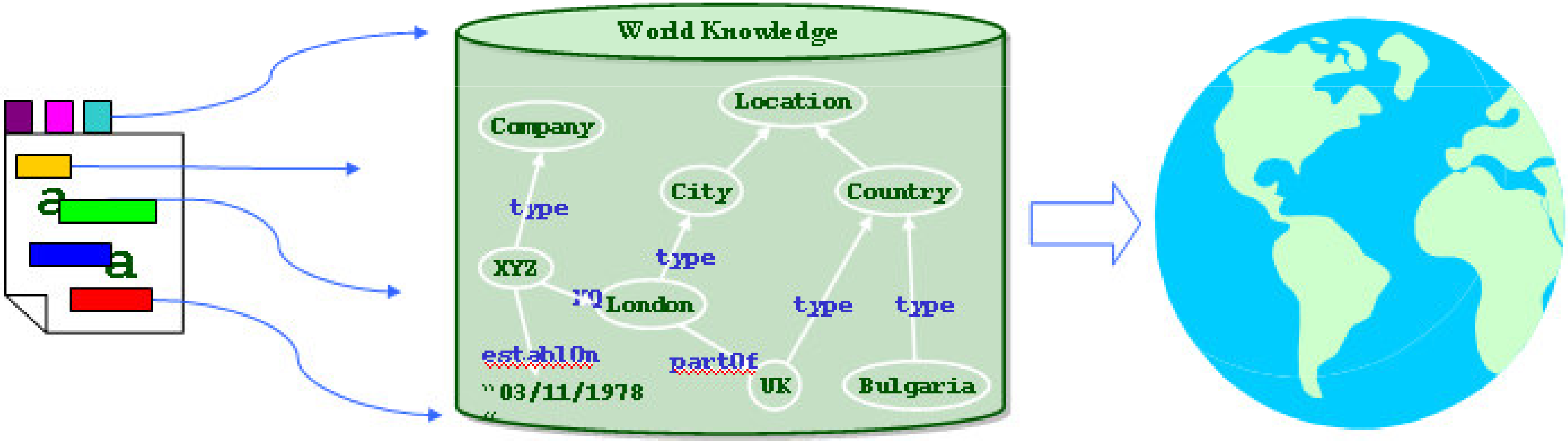


Semantic Web vs. WWW





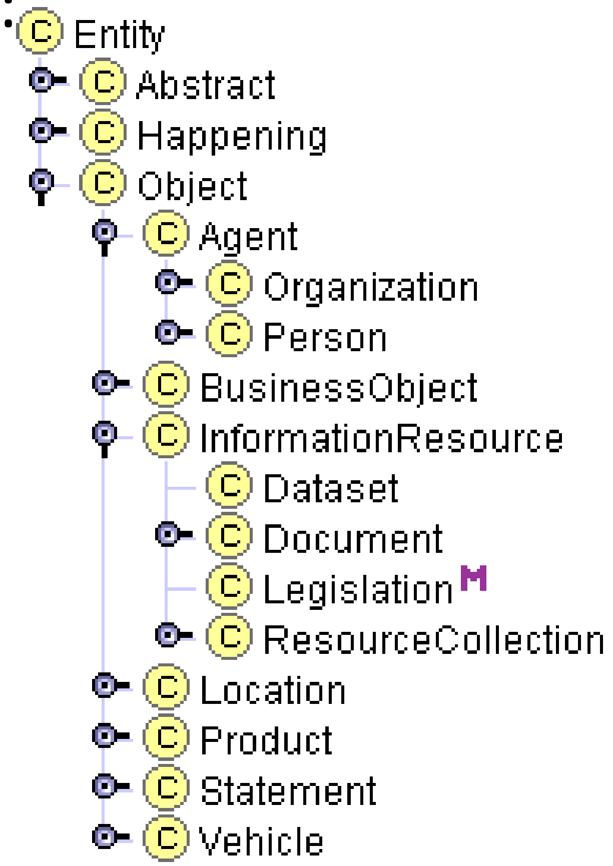
Semantic Web Is a Model of the World



Introduction to Ontologies

Despite the formal definitions, ontologies are:

- **Conceptual models** or schemata
  - Represented in a formalism which allows
  - Unambiguous “semantic” interpretation
  - Inference
- Can be considered a combination of:
  - DB schema
  - XML Schema
  - OO-diagram (e.g. UML)
  - Subject hierarchy/taxonomy (think of Yahoo)
- Ontologies enable agreement on the semantics across applications

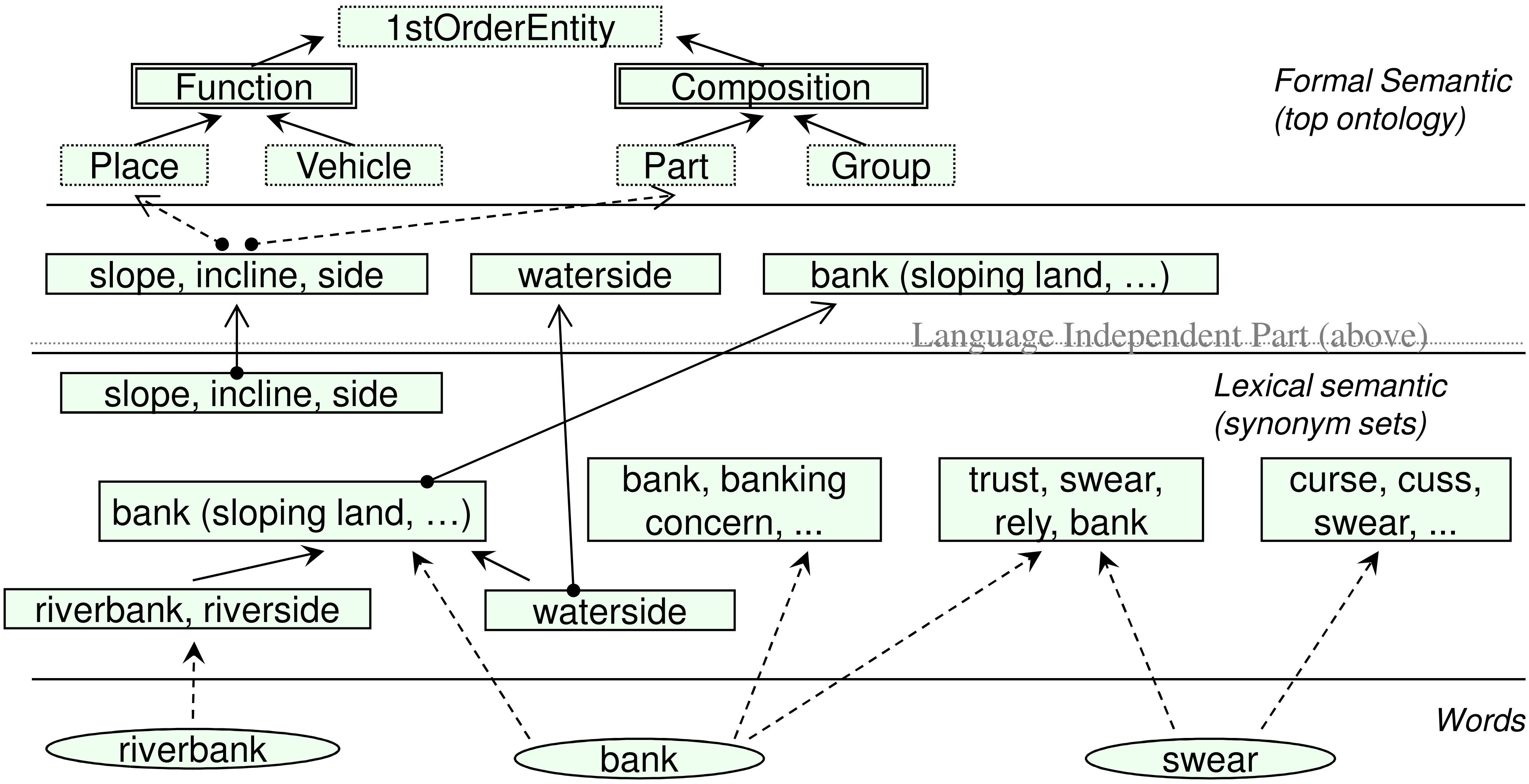


- Entity
  - Abstract
  - Happening
  - Object
    - Agent
      - Organization
      - Person
    - BusinessObject
    - InformationResource
      - Dataset
      - Document
      - Legislation
      - ResourceCollection
    - Location
    - Product
    - Statement
    - Vehicle

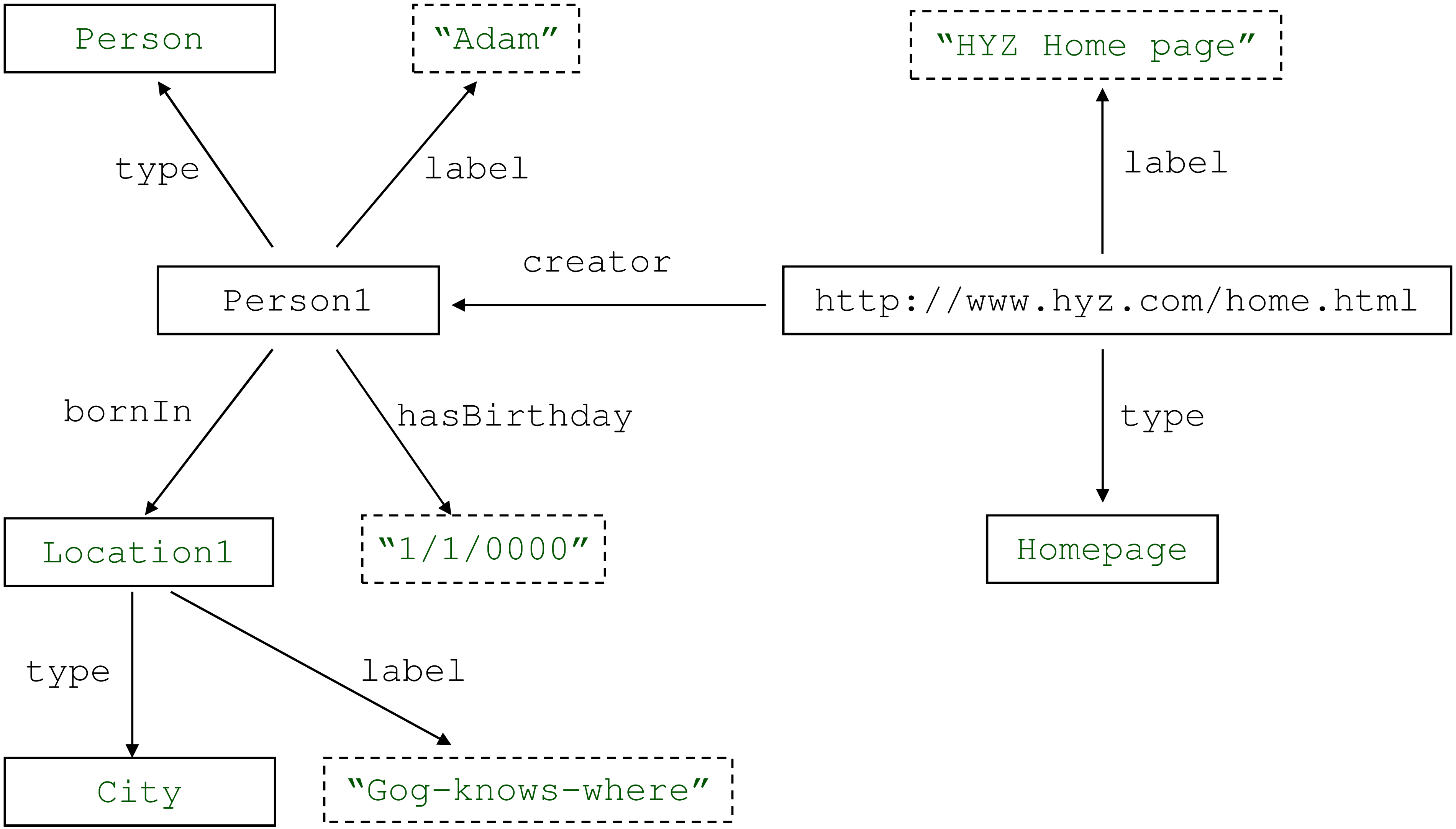
Types of Ontologies

- By Complexity of the representation language:
  - Light-weight vs. Heavy-weight
- By level of generality/reusability
  - Upper-level
  - Domain
  - Application and System
- By type of semantics being modelled
  - Schema-ontologies
  - Topic-ontologies
  - Lexical ontologies

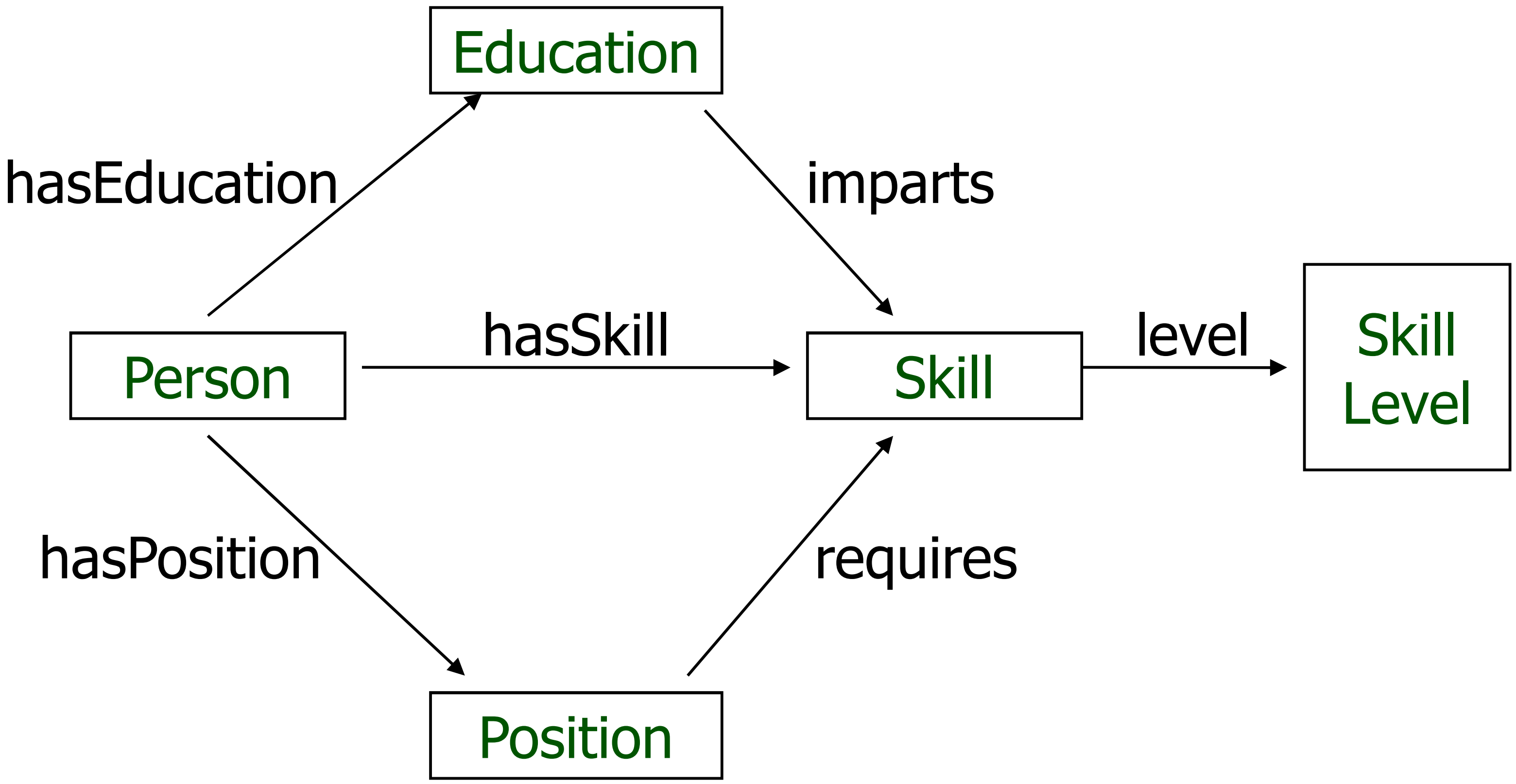
Lexical Semantics: EuroWordnet



Piece of RDF Graph

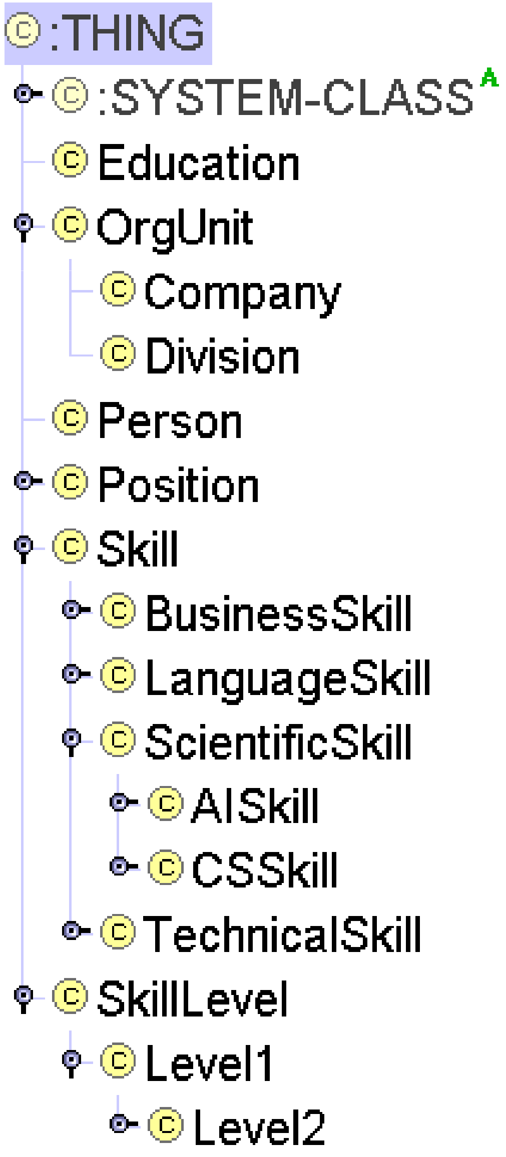


RDF: The schema as ER-graph



Simplified conceptual schema of a Skills Management Application

RDF: Sample Class Hierarchy



```
:THING
  :SYSTEM-CLASS
  Education
  OrgUnit
    Company
    Division
  Person
  Position
  Skill
    BusinessSkill
    LanguageSkill
    ScientificSkill
      AISkill
      CSSkill
    TechnicalSkill
  SkillLevel
    Level1
      Level2
```

ontotext
Semantic Technology Lab

Presentation Outline

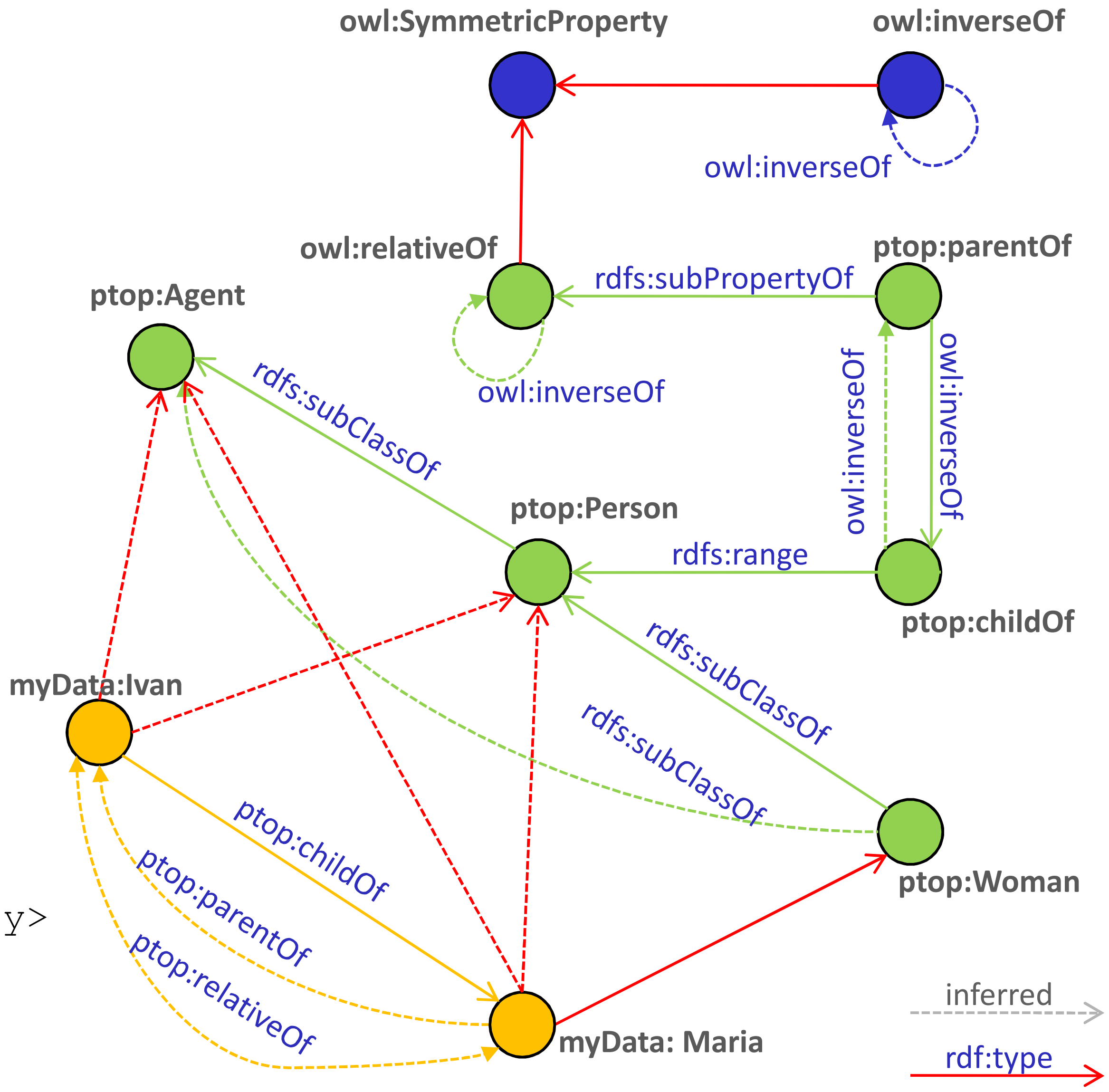
- Introduction
- **Semantic Databases**
- Linked Data
- Reason-able views
- Priming RDF Graphs

Semantic Repository for RDFS and OWL

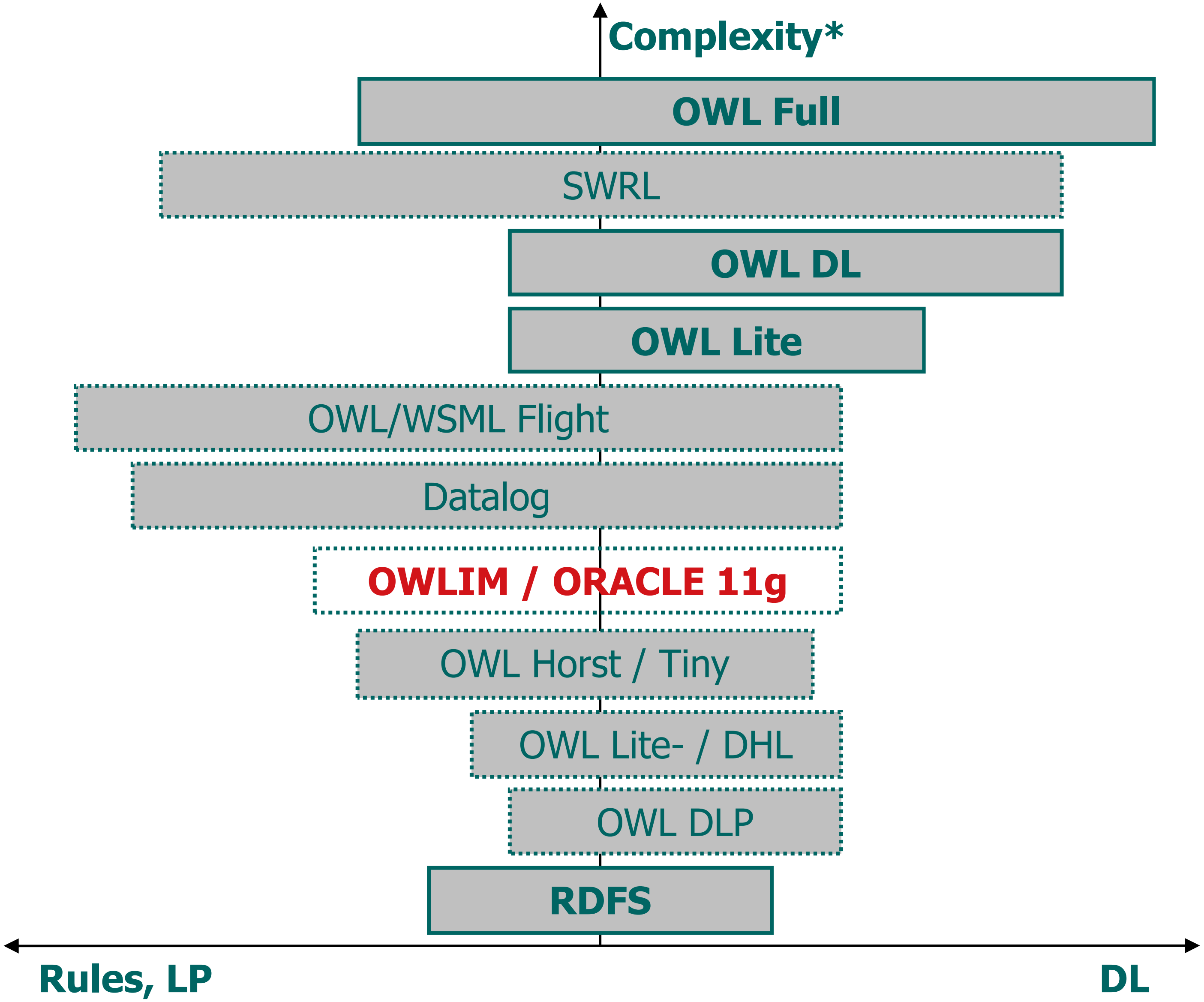
- OWLIM is a **scalable semantic repository** which allows
  - Management, integration, and analysis of heterogeneous data
  - Combined with light-weight reasoning capabilities
- Its performance allows it to replace RDBMS in many applications
  - More suitable for OLAP than for OLTP
- OWLIM is RDF database with **high-performance reasoning**:
  - The inference is based on logical **rule-entailment**
  - Full RDFS and limited **OWL Lite** and Horst are supported
  - **Custom semantics** defined via rules and axiomatic triples

Rule-Based Inference

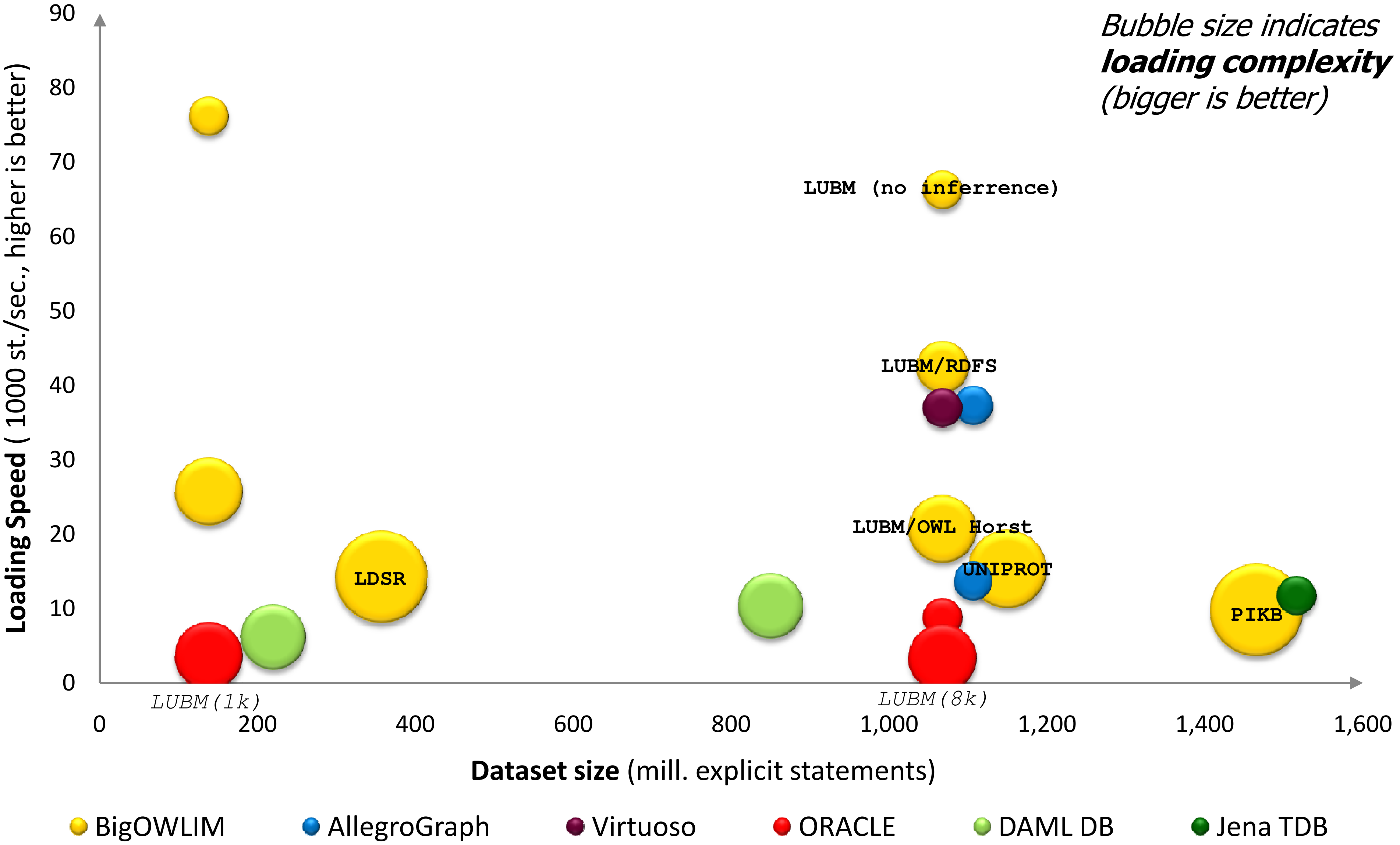
```
<C1,rdfs:subClassOf,C2>
<C2,rdfs:subClassOf,C3>
⇒ <C1,rdfs:subClassOf,C3>

<I,rdf:type,C1>
<C1,rdfs:subClassOf,C2>
⇒ <I,rdf:type,C2>

<I1,P1,I2>
<P1,rdfs:range,C2>
⇒ <I2,rdf:type,C2>

<P1,owl:inverseOf,P2>
<I1,P1,I2>
⇒  <I2,P2,I1>

<P1,rdf:type,owl:SymmetricProperty>
⇒  <P1,owl:inverseOf,P1>
```



Naïve OWL Fragments Map

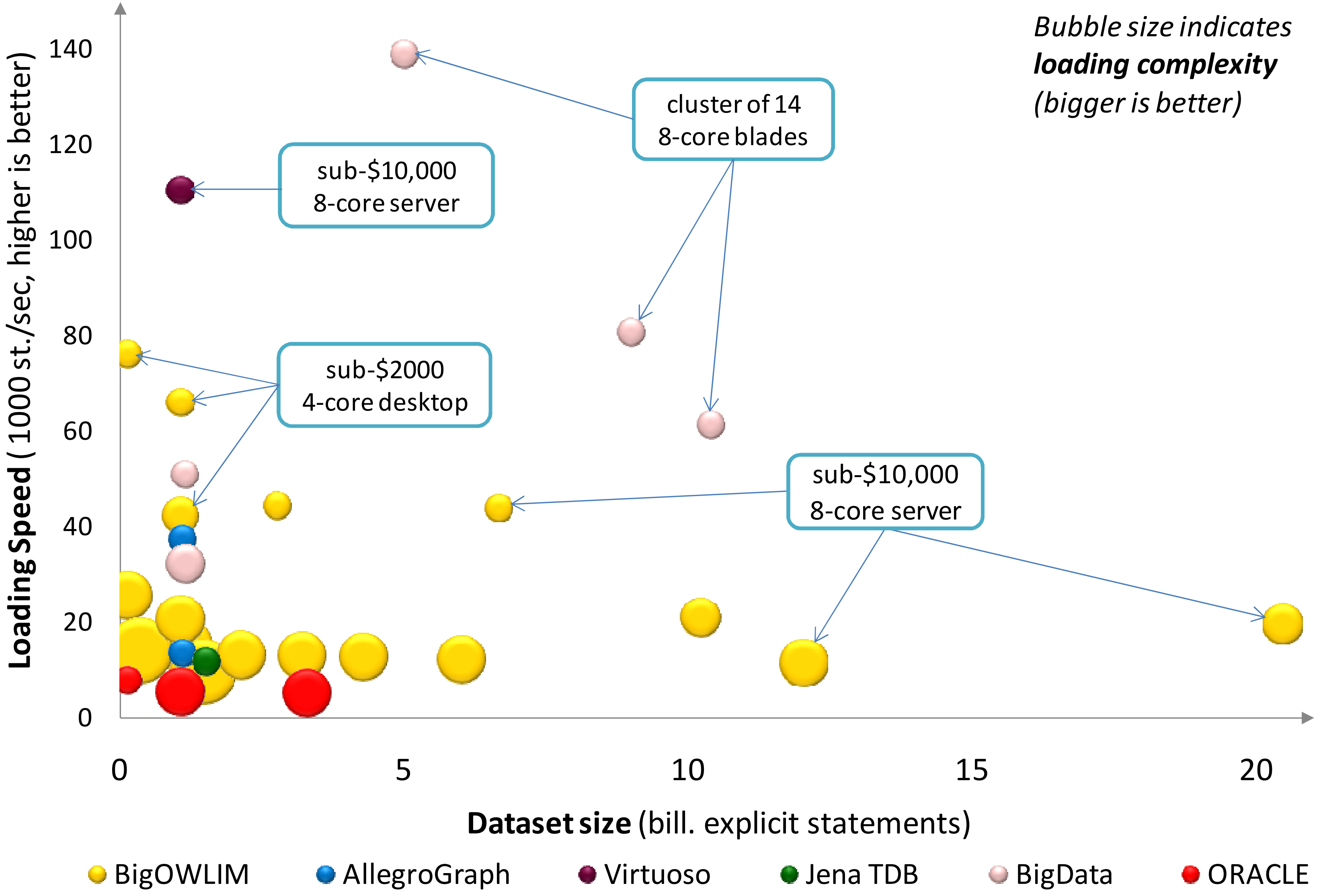


Scalable Reasoning Map (up to 1.5B)



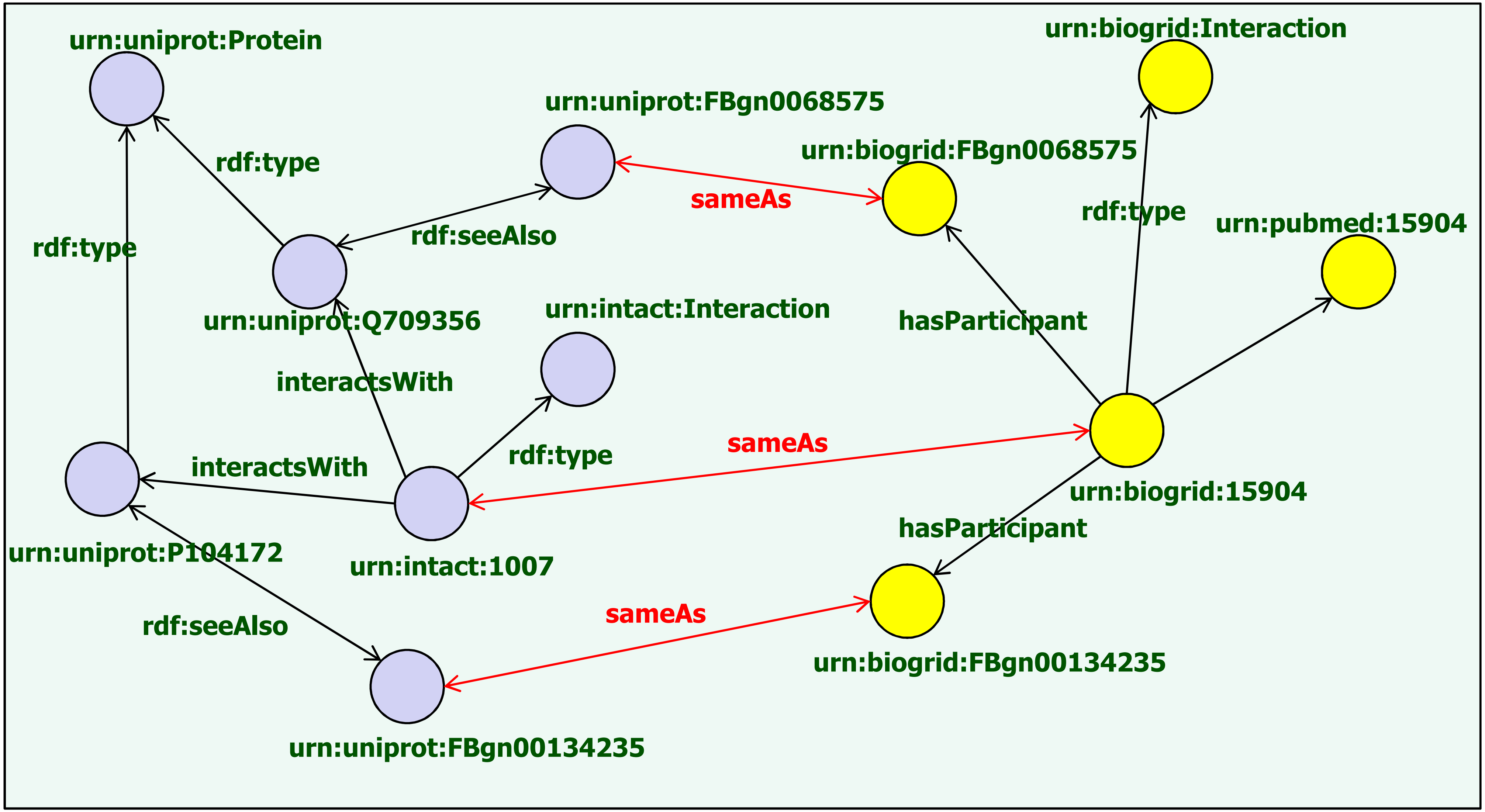
Bubble size indicates ***loading complexity*** *(bigger is better)*

Scalable Reasoning Map (the big picture)



Bubble size indicates ***loading complexity*** *(bigger is better)*

RDF Warehousing of Protein Interactions

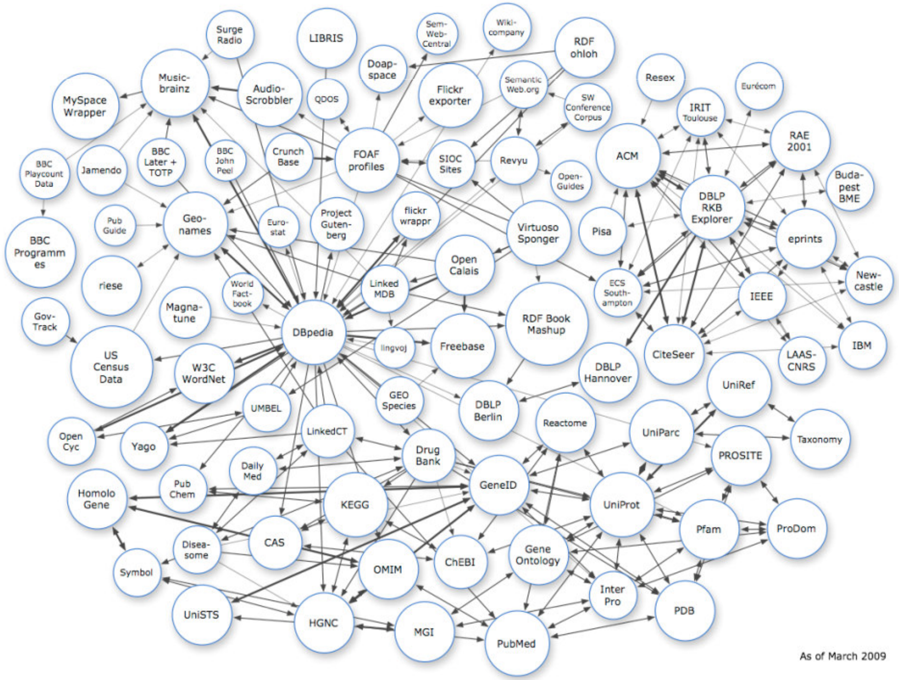


Presentation Outline

- Introduction
- Semantic Databases
- **Linked Data**
- Reason-able views
- Priming RDF Graphs

Linking Open Data

- Linking Open Data (LOD) W3C SWEO Community project
  http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData



As of March 2009

- Initiative for publishing “linked data” – a set of principles, which allows browsing of RDF data, spread across different servers, in the way HTML is browsed

Reason-able views to the LOD

- Classical sound and complete reasoning is unfeasible to a web of linked data.
- The major obstacles:
  - Most of the popular reasoning setups count on “closed world assumption” (which is irrelevant in web context)
  - The complexity of reasoning even the with the simplest DL (say OWL Lite) is prohibitatively high
  - Some of the datasets of LOD (or some parts of them) are not suitable for reasoning. It seems that some data publishers use the OWL and RDFS vocabulary without account for its formal semantics
  - Although reasoning with data distributed across different web servers is possible it is much slower than reasoning with local data. The fundamental reason is related to the so called "remote join" problem known from the distributed DBMS

Reason-able views to the LOD (2)

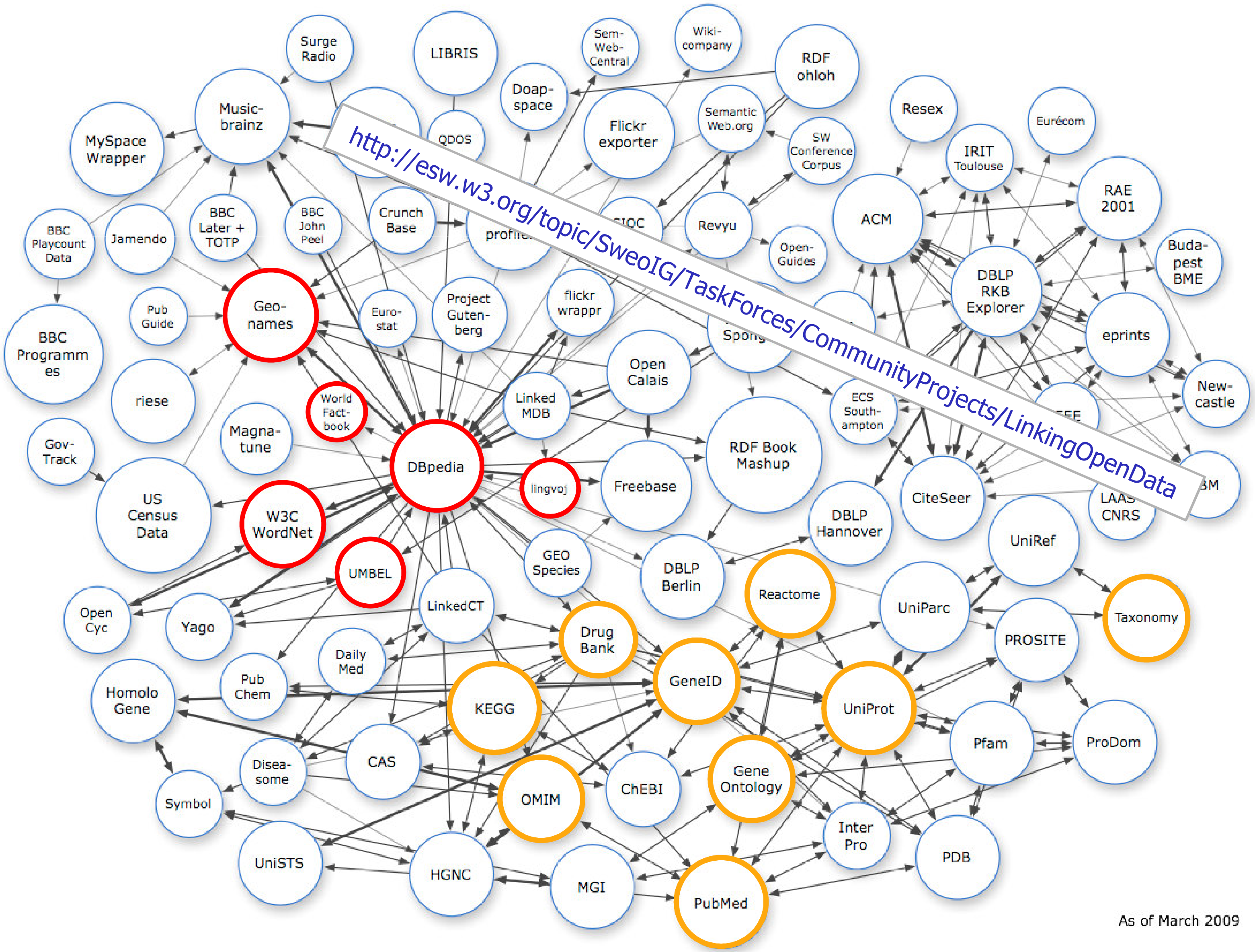
- *Reason-able views* represent an approach for reasoning with the web of linked data
- Key ideas:
  - Inference with respect to **tractable OWL** dialects
  - group **selected datasets and ontologies** in a reason-able view
  - load all ontologies and data in a **single semantic repository**
- Selection Criteria:
  - the dataset (or a part of it that is easy to define and isolate) allows inference, which delivers meaningful results under the semantics determined for the view;
  - the dataset is more or less static, i.e. not a wrapper for a database or service

Two reason-able views to the web of linked data

Ontotext persents:

- **Linked Data Semantic Repository**
  - Some of the central LOD datasets
  - General-purpose information
  - 358M explicit and 512M inferred triples
  - **http://www.ontotext.com/ldsr/**
- **Linked Life Data - PIKB** (in yellow)
  - Several popular life-science datasets
  - Complemented by gluing ontologies
  - 1.47B explicit and 842M inferred triples
  - **http://www.linkedlifedata.com**

Linking Open Data Datasets



http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

As of March 2009

Linked Data Semantic Repository

- **Datasets**: DBPedia, Geonames, UMBEL, Wordnet, CIA World Factbook, Lingvoj
- **Ontologies:** Dublin Core, SKOS, RSS
- **Inference:** materialization with respect to owl-max
  - One of the richest tractable fragments of OWL
  - Proper extension of LarKS’s OWL-Lepton-I profile
  - Seems to completely cover the semantics of the data
  - **owl:sameAs optimisation** in BigOWLIM, allows almost 10-fold reduction of the indices, without loss of semantics or performance
- Publicly available at http://ldsr.ontotext.com
  - **Query** and **explore** through openrdf’s Workbench (web UI)
  - **SPARQL end-point**

LDSR Statistics

Dataset	Explicit Triples	Inferred after import	RDF nodes after import
Umbel	3,167,205	56,833	1,230,550
DBpedia (sameAs)	145,120	278,139	1,414,157
Geonames	72,747,880	428,696,785	34,813,153
DBpedia 3.2 core	280,697,077	38,922,702	100,131,770
lingvoj	19,692	848,978	100,141,681
Wordnet	1,946,838	8,575,920	100,769,150
CIA Factbook	35,956	291,877	101,005,679
Total	**357,844,134**	**511,522,747**	**101,005,679**

- Total statements in the repository indices: **869M**
- Number of retrievable statements (considering owl:sameAs expansion): **above 1.1B**

Which size?

- Linked Data Semantic Repository loaded in 7 hours on “solid” desktop
- Numbers:
  - Number of imported statements (NIS): 357M
  - Number of new inferred statements: 512M
  - Number of stores statements (NSS): 869M
  - Number of retrievable statements (NRS): 1.14B
- owl:sameAs optimisation allowed reducing the indices by 280M statements

What to do with Wikipedia categories

- On the original data:
  - 420 000 categories in DBPedia 3.2:
  - 415 729 explicitly declared ones and 4357 mentioned as broader-to
  - 786K SKOS broader relations among those categories
  - Human-assigned and contained cycles and errors due to lexical ambiguity and other (human) factors.
- Cycles
  - 2023 simple cycles were detected
  - 1070 of which were trivial – a category broader to itself
  - Remaining 953 non-trivial cycles
  - Largest length: 188

ontotext
Semantic Technology Lab

What to do with Wikipedia categories (2)

- Experiments on refinement of the hierarchy:
  - Discard broader cycles automatically
  - Manually discard mis-assigned broader statements causing cycles
    - Changing 737 relations from broader to related, cuts all cycles
  - Remove System Categories
  - Remove multiple inheritance, leaving up to two parents
    - Various heuristics to chose which ones to leave
- Latest metrics
  - 1.96M explicit statements
  - 836K entities
  - 189M total statements (explicit + inferred)
  - About 44M inferred broader links
  - Loading took 4 hours

ontotext
Semantic Technology Lab

Presentation Outline

- Introduction
- Semantic Databases
- Linked Data
- Reason-able views
- Priming RDF Graphs

LDSR in LarKC

- LDSR was set up as a testbed for selection and ranking components and plug-ins in LarKC
- PageRankRDF performance on LDSR:
  - it takes only 10 seconds to perform one iteration of PageRank
  - **3 minutes to compute the ranks of the 100 million nodes** in LDSR
- DualRDF (an RDF priming component)
- More details are presented in D2.4.1

DualRDF: Spreading Actioncation over RDF Grahps

- The performance of the spreading activation tasks varies considerably depending on the parameters of the process
- As a reference point use the following result: it takes **7 seconds to activate about 7 thousand nodes** after spreading of activation from resource http://dbpedia.org/resource/Berlin with decay factor 0.25.
- Queries on the “primed” or “selected” part of a dataset run up to 20 times faster and return only focussed results

LDSR Priming Experiments

Single-node activation results

Activation seed	Decay factor	Firing Threshold	Fired Entities	Time (sec)
dbpedia:London	0.85	0.25	54,032	26
dbpedia:Paris	0.85	0.25	96,275	27
dbpedia:Berlin	0.85	0.25	25,721	19

Dependence on the Decay Factor

Activation seed	*Decay factor*	Firing Threshold	Fired Entities	Time (sec)
dbpedia:Berlin	*0.75*	0.25	25,720	10
dbpedia:Berlin	*0.50*	0.25	7212	9
dbpedia:Berlin	*0.25*	0.25	7193	7

Dependence on the Firing Threshold

Activation seed	Decay factor	*Firing Threshold*	Fired Entities	Time (sec)
dbpedia:London	0.50	*0.20*	54,030	8
dbpedia:London	0.50	*0.40*	10,959	7
dbpedia:London	0.50	*0.60*	9,700	6

Multi-node activation

Activation seed	*Decay factor*	Firing Threshold	Fired Entities	Time (sec)
London,Paris,Berlin	*0.75*	0.25	594,938	149
London,Paris,Berlin	*0.50*	0.25	475,043	32
London,Paris,Berlin	*0.25*	0.25	368,483	25

Query performance after selection through priming

	Query 1:	Query 2
# of active nodes after graph priming	368,381	368,413
# of results (time for eval.) **before priming**	2,174 (87ms)	163,438 (518ms)
# of results (time for eval.) **after priming**	23 (5ms)	530 (47ms)
Reduction ratio for the number of results	94.5	308.4

Thanks!

Artificial Intelligence is not death!

It is just searching for a batter label …