Следващо поколение бази данни и машини за търсене, базирани на технологии от Семантичния Уеб

Атанас Киряков, Л-ия Онтотекст, Сирма

Lunix-BG Конференция, София

Юли 2007

Presentation Outline

- **Sirma and Ontotext**
- Introduction to Semantic Web and Ontologies
- OWLIM: the “semantic database”
- KIM: the “semantic search engine”
  - CORE Search and Timelines Demo
- Applications

What is Sirma?

- Established in **1992** as a Bulgarian-Canadian AI Lab
- Currently it is a **group of diverse software businesses**
- **Offices in:**
  - Sofia, Kazanlak, Plovdiv, Varna – Bulgaria
  - Montreal, Ottawa – Canada
  - Sao Paolo – Brazil
  - Santa Rosa (the Bay area) – USA
- More than **10 companies and business units**
- **Top-3 software house** in Bulgaria, about **200 employees**
- **1999 EIST prize** winner
- **ISO 9001:2000** certified

THE EUROPEAN INFORMATION TECHNOLOGY PRIZE Winner

ontotext Semantic Technology Lab

Businesses and Joint Ventures of Sirma

- **Sirma Solutions:** e-Business, Banking, C3, IT consultancy
- **Ontotext Lab:** Semantic Technologies
- **EngView Systems:** CAD/CAM and measurement
- **Sirma Media:** E-publishing and edutainment
- **Sirma Business Consulting:** Banking, ERP
- **Pirina Technologies:** cutting plotters
- **WorkLogic:** groupware and e-Government
- **Eyebill:** VOIP billing and CRM systems (JV with Nexcom)
- **Innovantage:** recruitment intelligence in UK (joint venture)
- **SEP:** mobile payments operator (joint venture)

Ontotext Lab

- **R&D lab for Semantic (Web (Service)) Technologies**
- Active in several **research areas**, including:
  - Semantic Databases: Ontology Management, Reasoning;
  - Semantic Search: Information Extraction and Retrieval (IE, IR);
  - Semantic Web Services.
- **Core business:** research and core technology development
- **Applications in:** Semantic Web, Web Mining, KM, BI, Media Research, Life Sciences, Enterprise Application Intgeration, Business Process Management
- Aside from the scientific matters, most of the Ontotext fellows are just **professional software developers**

Leading Semantic Web Technology Provider

Ontotext is a leading Semantic Web **technology developer:**

- the developer of the **KIM** semantic annotation platform
- the developer of the **wsmo4j** semantic web services API and the **WSMO Studio** service development environment;
- the developer of **OWLIM** – the fastest OWL semantic repository;

Contributions to **open-source projects:**

- a major co-developer of **GATE** language engineering platform;
- a major co-developer of **Sesame** semantic repository;

Outstanding Research Projects

Ontotext is part of **outstanding European research projects**:

- On-To-Knowledge, SWWS, DIP, SEKT, PrestoSpace, SUPER, etc.;
- **100 MEuro** is the total budget of the European research projects of Ontotext
- Sirma was awarded as the most successful Bulgarian company in FP6

Academic & Technology Partners

- **NLP Group**, Sheffield University, UK;
- **Digital Enterprise Research Institute (DERI),** Innsbruck, Austria, and Ireland, Galway;
- **Linguistic Modelling Lab.** Bulgarian Academy of Sciences;
- **British Telecommunications Plc**, (BT), UK.
- **Institut AIFB** (FZI, Ontoprise), Karlsruhe, Germany;
- **DFKI Language Technology Lab**, Saarbrucken, Germany;
- **The Open University, Knowledge Media Institute**, UK;
- **Other partners:** SAP, IBM, HP Labs, Software AG, Capgemini, RAI, BBC, Telefonica

Ontotext Facts

- **Founded: year 2000**; part of Sirma Group
- Staff: **19 employees**
- http://www.ontotext.com: **~300 visits/day**
- **Google 1st place**: "semantic annotation", "semantic repository"
- Research **publications: 30**
- Research **projects**: **14** (8 running)
- **Products: 5**
- Partners: at least **20 partners** we directly cooperate with
- Number of servers **CPU cores: 33**; ~2 per engineer
- Average age: **29 years**

Products

- **OWLIM** – a semantic repository, http://www.ontotext.com/owlim
- The **KIM** Platform (the next slides), http://www.ontotext.com/kim.
- **WSMO Studio** (http://www.wsmostudio.org)
  - Semantic Web Service description development environment
  - **Top 1% of SourceForge**, Ohloh evaluation: ~ $730K
- **wsmo4j** (http://wsmo4j.sourceforge.net)
  - The WSMO API; Ohloh evaluation: ~ $500K
- **PROTON** (http://proton.semanticweb.org)
  - A light-weight upper-level ontology
- **ORDI** (http://www.ontotext.com/ordi)
  - An ontology-middleware and data integration framework

Projects and Joint Ventures

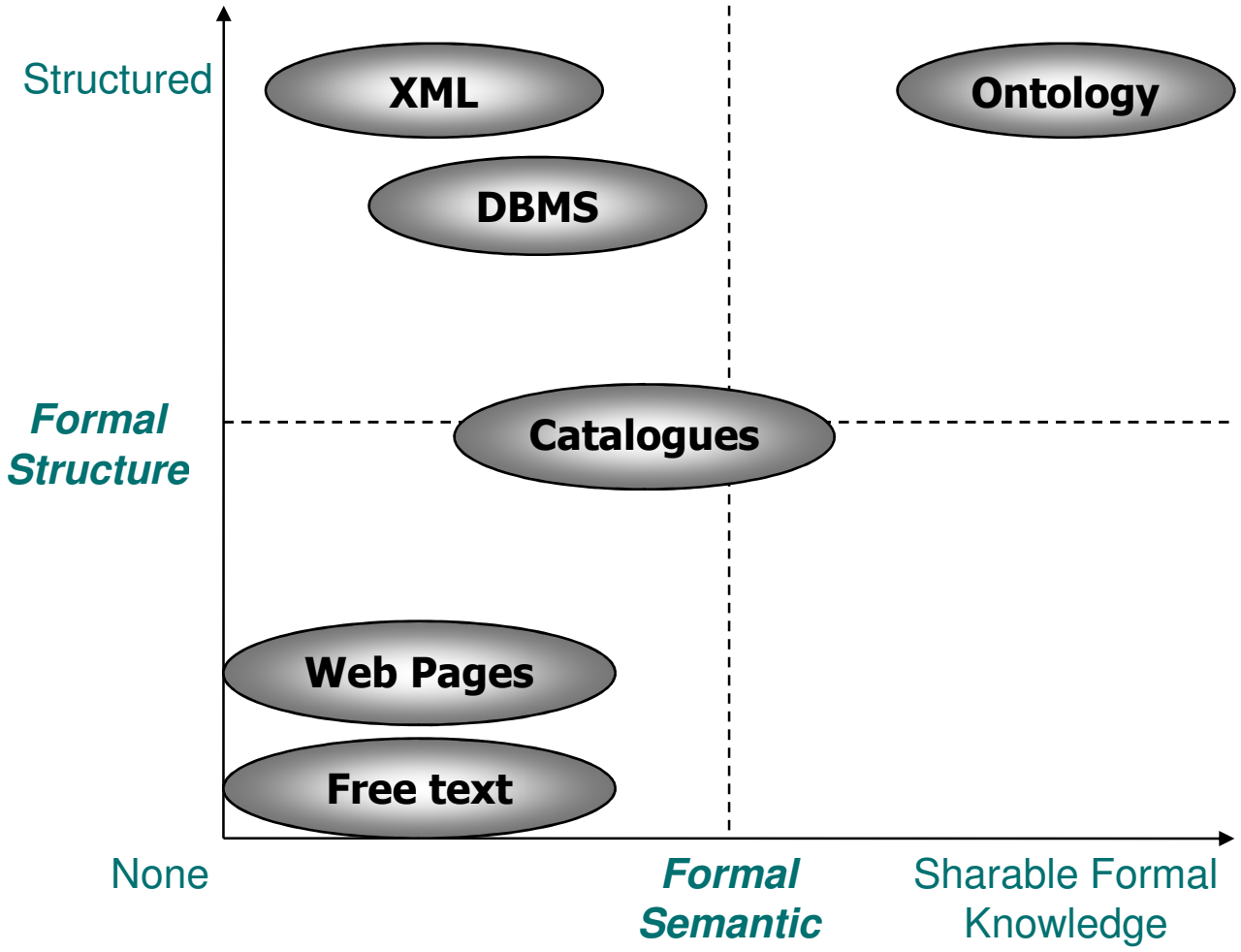
- **JOCI** – vertical search for recruitment in UK
- **LifeSKIM** – application of semantic technologies in Life Sciences
- **ETO** – a large scale BG portal, together with NetInfo

Life Science

Presentation Outline

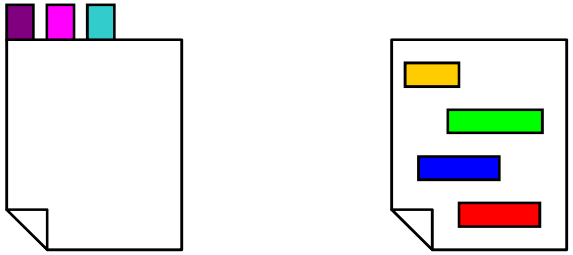
- Sirma and Ontotext
- **Introduction to Semantic Web and Ontologies**
- OWLIM: the “semantic database”
- KIM: the “semantic search engine”
  - CORE Search and Timelines Demo
- Applications

Sorts of Data



Annotations

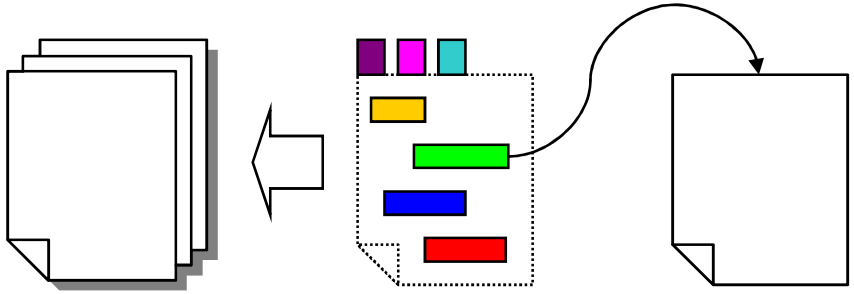
Embedded markup



Document-level

Character-level

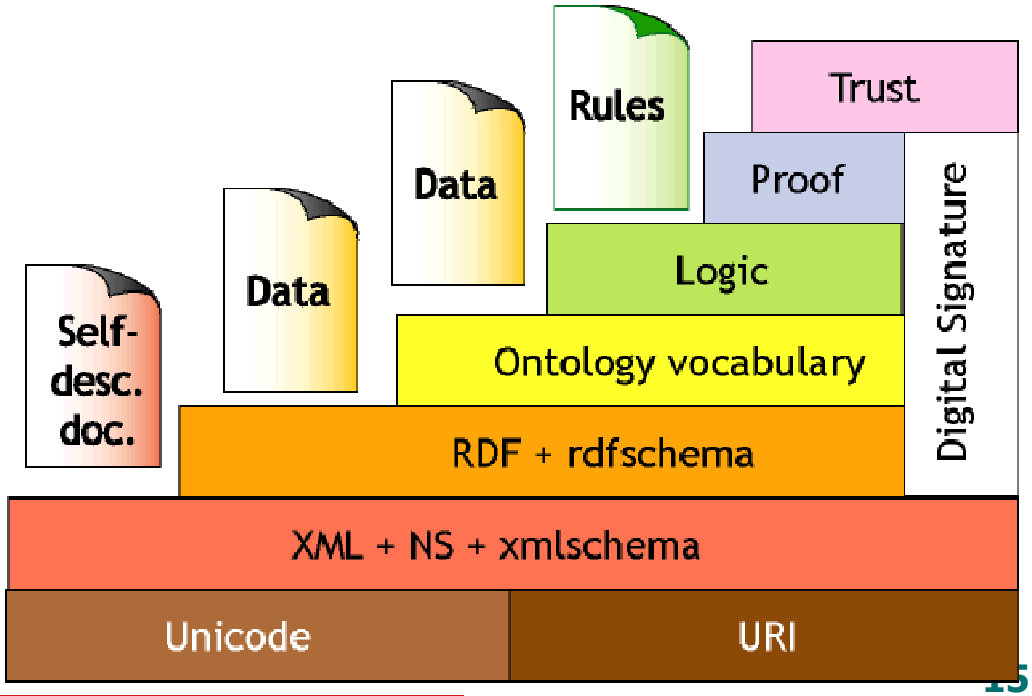
Standoff References



Hyperlink

Semantic Web

- The Semantic Web is the **abstract representation of data** on the WWW, based on the RDF and other standards
- SW is being **developed by the W3C**, in collaboration with a large number of researchers and industrial partners
- http://www.w3.org/2001/sw/ , http://www.SemanticWeb.org

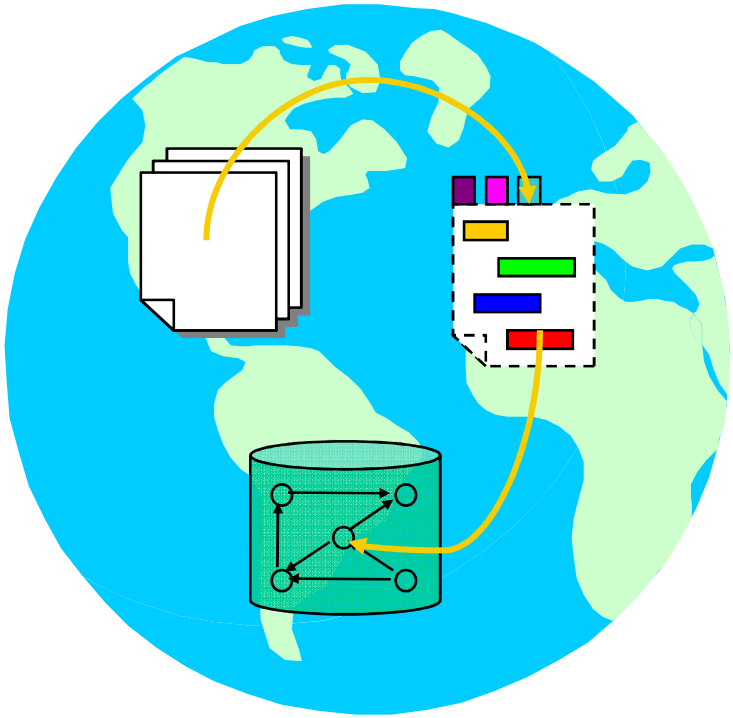


Semantic Web (II)

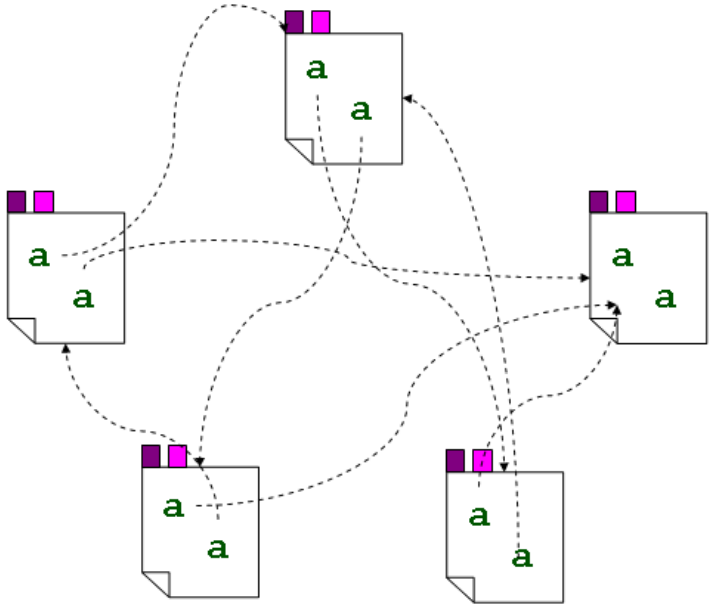
- "The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.“ [Berners-Lee et al. 2001]

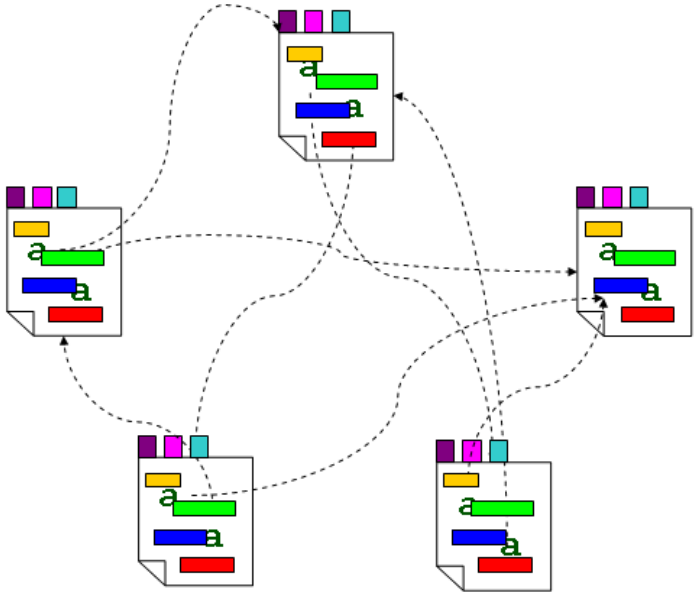
The spirit:

- Automatically processable **metadata** regarding:
  - the structure (syntax) and
  - the meaning (semantics)
  - of the content.
- Presented in a **standard form;**
- **Dynamic interpretation for unforeseen purposes**

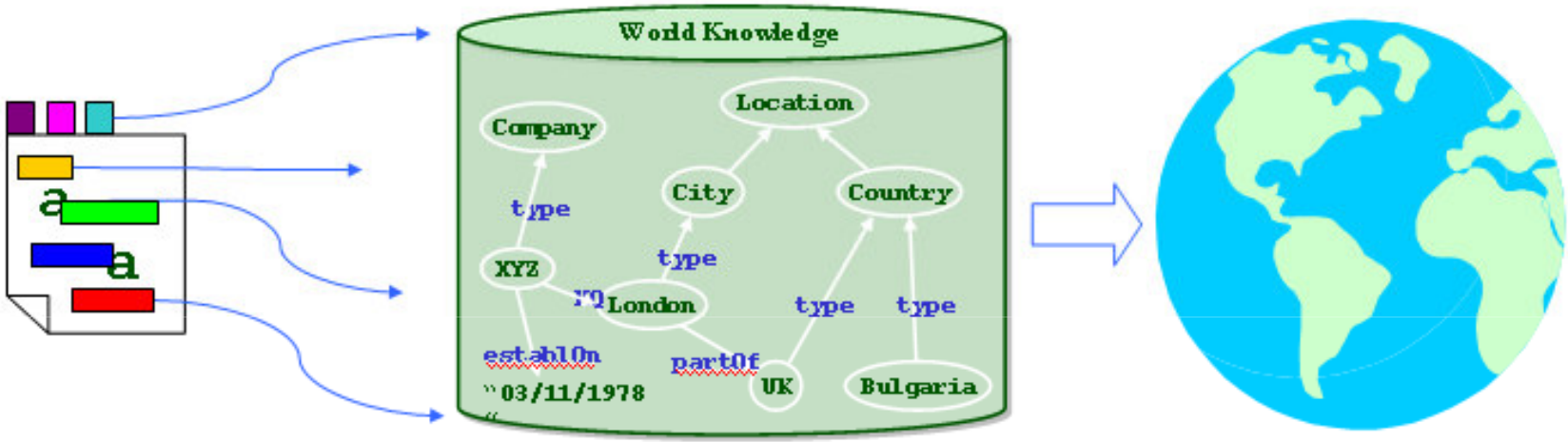


Semantic Web vs. WWW





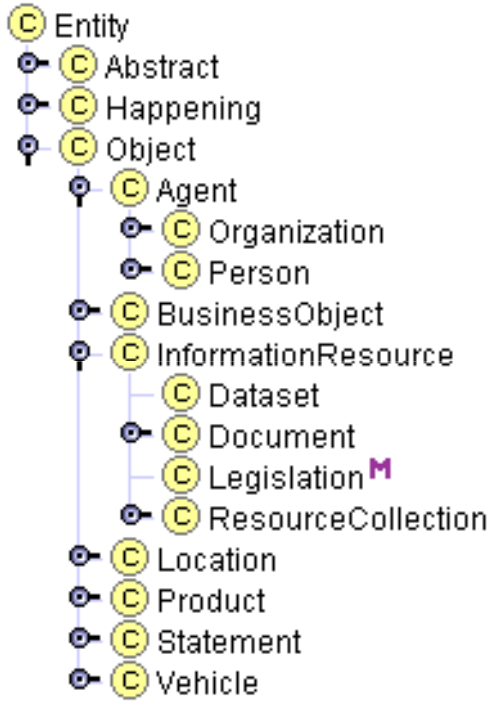
Semantic Web Is a Model of the World



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)

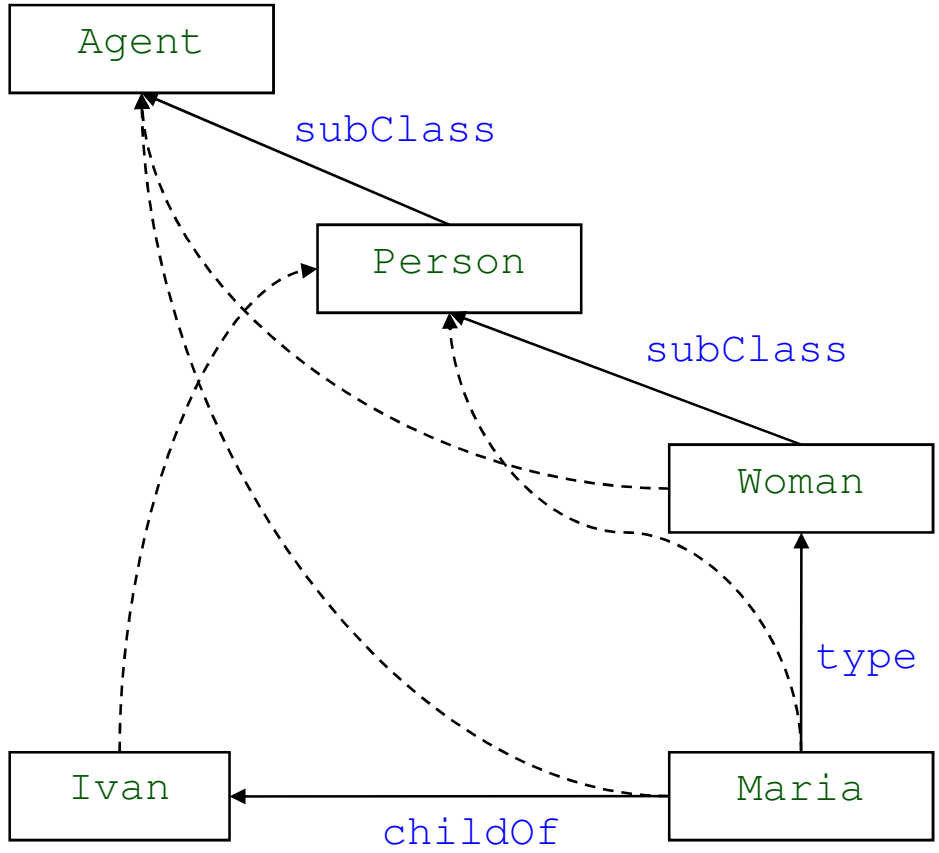


- Ontologies enable agreement on the semantics across applications

Inferred Closure

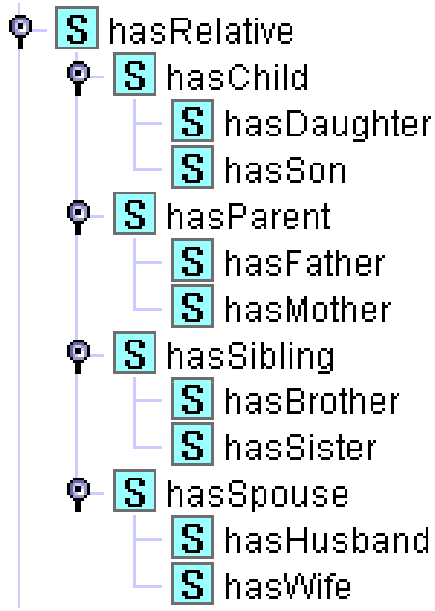
Sample rules:

```
     <C1, subclass, C2>
     <C2, subClass, C3>
=>   <C1, subClass, C3>

     <I, type, C1>
     <C1, subclass, C2>
=>   <I, type, C2>

     <I1, P1, I2>
     <P1, range, C2>
=>   <I2, type, C2>
```



Introduction to Ontologies (II)

- Imagine a DB storing“John is a son of Mary”.
- It will be able to "answer" just:
  - Which are the sons of Mary? Which son is John?
- An ontology with a definition of the family relationships. It could infer that:
  - John is a child of Mary (more general);
  - Mary is a woman;
  - Mary is the mother of John (inverse);
  - Mary is a relative of John (generalized inverse).

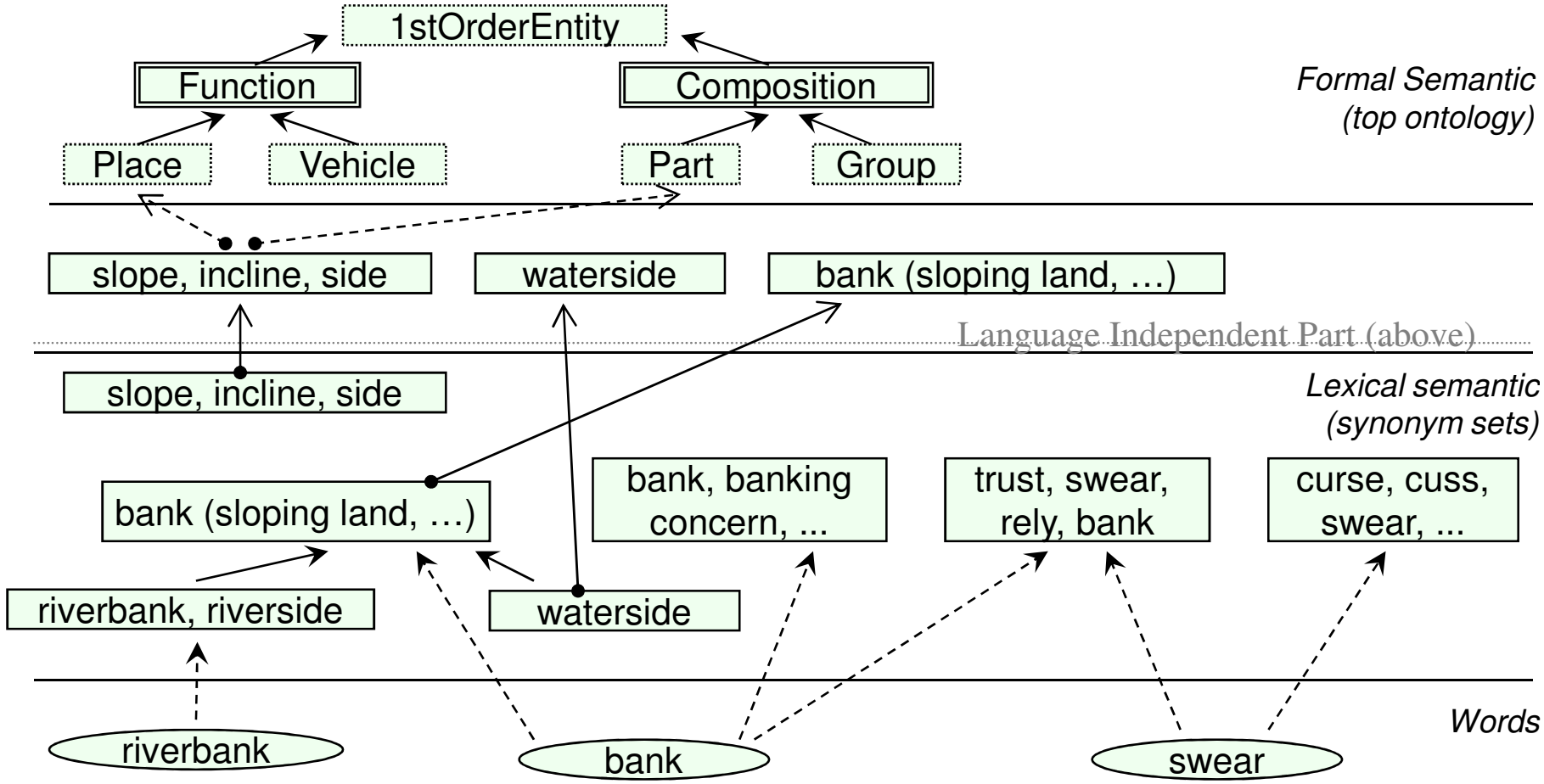


- The above facts, would remain "invisible" to a typical DB, which model of the world is limited to data-structures of strings and numbers.

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: Wordnet



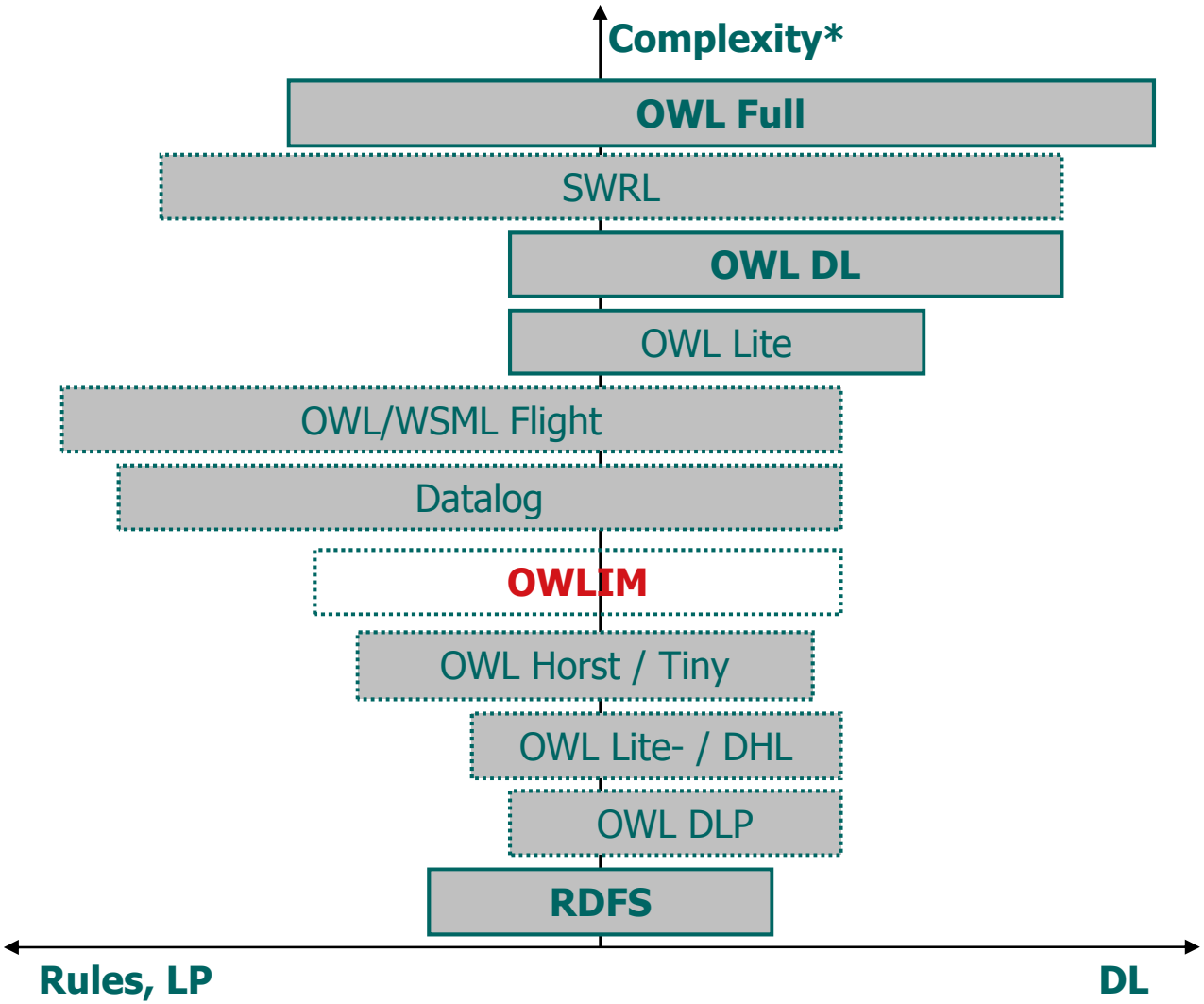
Presentation Outline

- Sirma and Ontotext
- Introduction to Semantic Web and Ontologies
- **OWLIM: the "semantic database"**
- KIM: the "semantic search engine"
  - CORE Search and Timelines Demo
- Applications

OWLIM

- 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
  - Suitable for **analytical tasks** and **Business Intelligence** (OLAP)
  - Inappropriate for highly dynamic transaction-oriented environments
- OWLIM is RDF repository with reasoning support:
  - **Full RDFS and limited OWL Lite,** Entailment **rules**
  - **Custom semantics** defined in terms of rules and axioms

Naïve OWL Fragments Map



Versions, Features, and Benchmarks

- OWLIM is implemented as:
  - Storage and Inference Layer, **SAIL, for Sesame** 1.2.x
  - It uses the TRREE engine for reasoning
- There are two versions: **SwiftOWLIM** and **BigOWLIM**
  - Both using TRREE, but different versions
  - The same inference and semantics (rule-compiler, etc)
- SwiftOWLIM is the **fastest OWL engine**!
  - It scales to 10 million statements on a desktop PC
  - It loaded LUBM(50,0) in 5 minutes, at average speed 25 KSt./sec.
- BigOWLIM is the **most scaleable OWL engine**!
  - It can process 1 billion statements on a $5000-worth server
  - It loaded LUBM(8000,0) and answered the queries in 69 hours

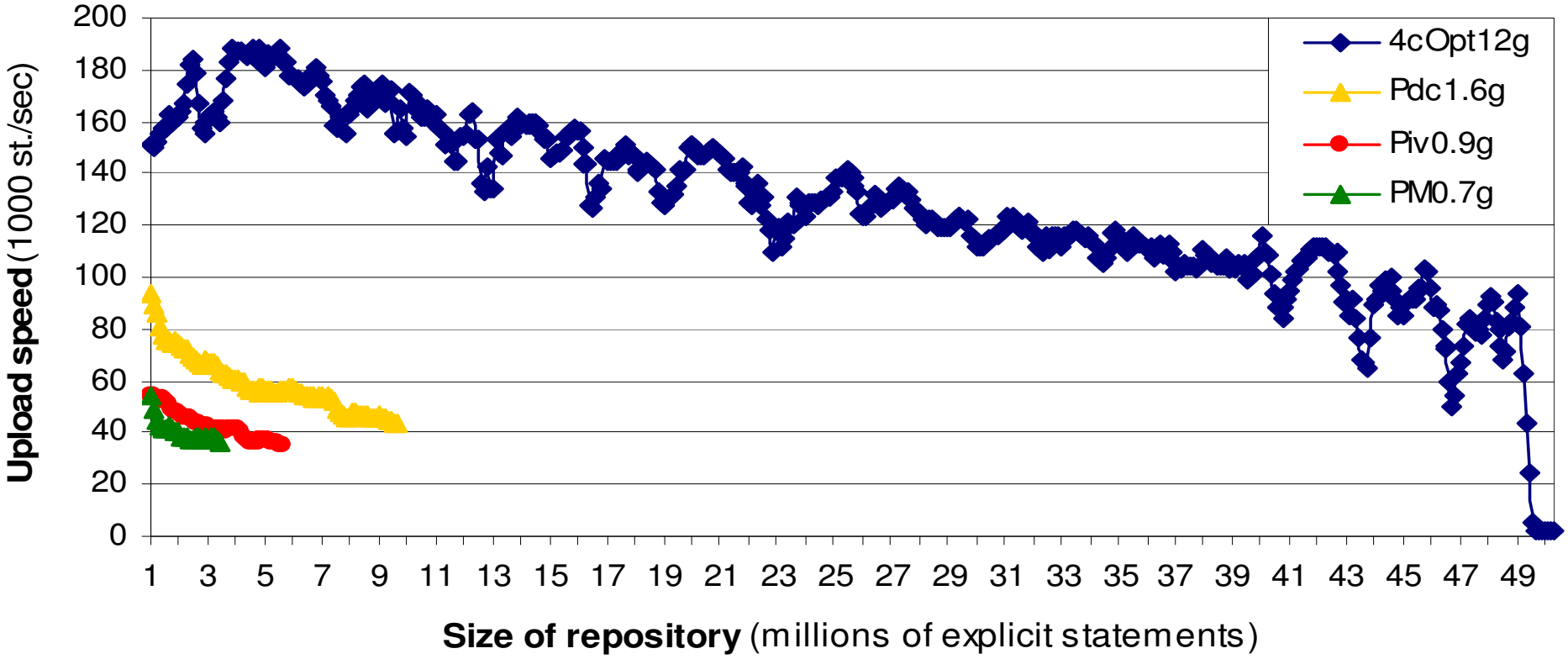
Versions and Features

	SwiftOWLIM	BigOWLIM
Scale (Mill. of explicit statem.)	10 MSt, using 1.6 GB RAM **100 MSt**, using 16 GB RAM	130 MSt, using 1.6GB **1068 MSt**, using 12GB
Processing speed (load+infer+store)	30 KSt/s on notebook **200 KSt/s** on server	4 KSt/s on notebook **20 KSt/s** on server
Query optimization	No	Yes
Persistence	Back-up in N-Triples	Binary files, allowing instant initialization
Efficient owl:sameAs	No	Yes
Licence and Availability	Open-source under LGPL; Uses SwiftTRREE that is free, but not open-source	Commercial Evaluation copies provided on request

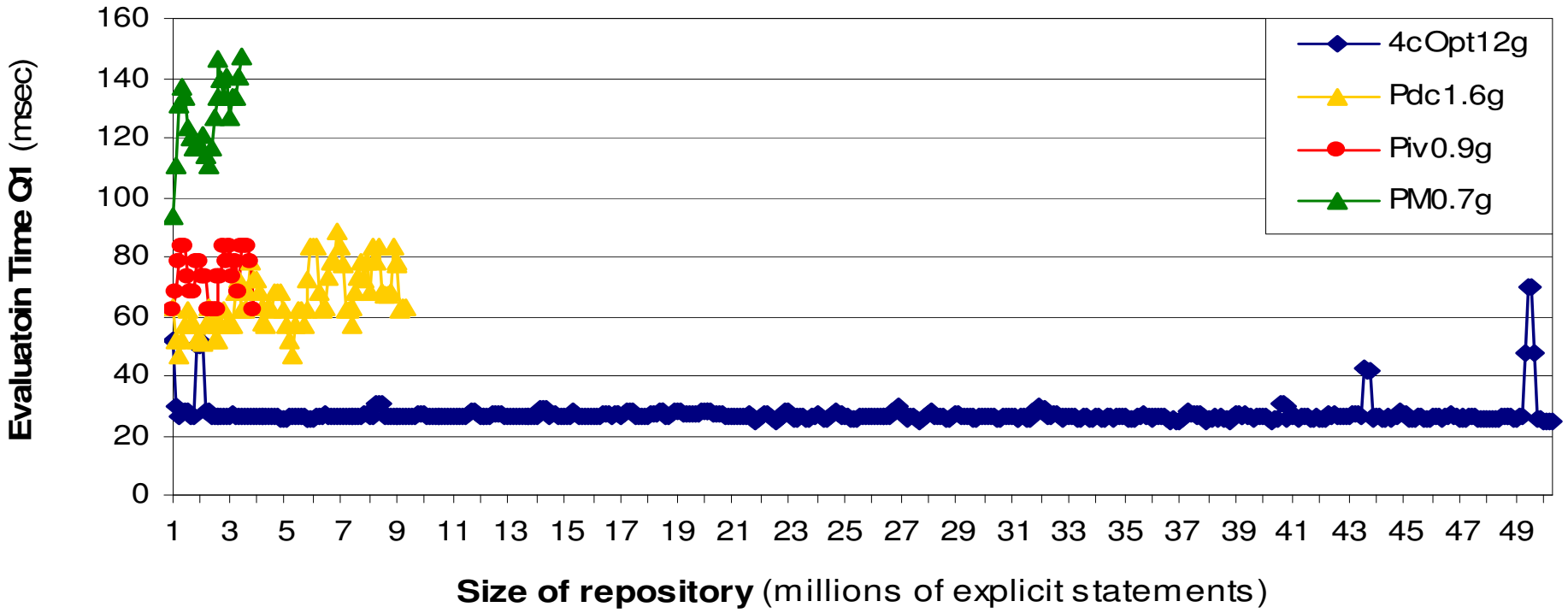
Performance Evaluation Configurations

Name	Configuration	RAM (Xmx)	JDK	Comment
4cOpt12g	2xOpteron 270 (2.0GHz), dual-core Suse Linux v.10, 64-bit	12GB, DDR400	JDK 1.6 64-bit	A database/application server; 4 SATA2 drives on RAID10; assembly cost ~4000 EURO
Pdc1.6g	Pentium D 920 (2.8GHz), Win XP	1.6GB, DDR400	JDK 1.6	Workstation
Piv0.9g	Pentium IV 630 (3.0GHz), Win XP	900MB, DDR2 533	JDK 1.6	Office desktop
Pm0.7g	Pentium Mobile 1.6GHz, Win XP	700MB, DDR266	JDK 1.6	Notebook (Q2'03)

OWLIM Performance: Upload and Inference

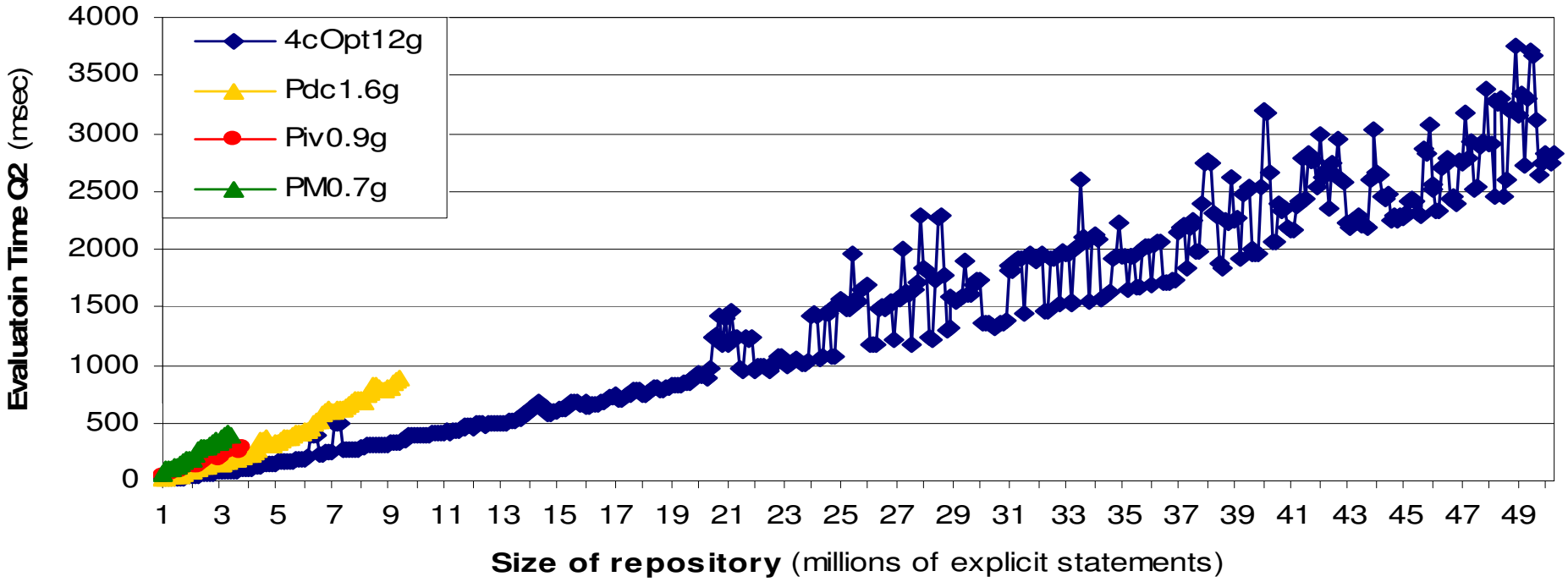


OWLIM Performance: Query Answering



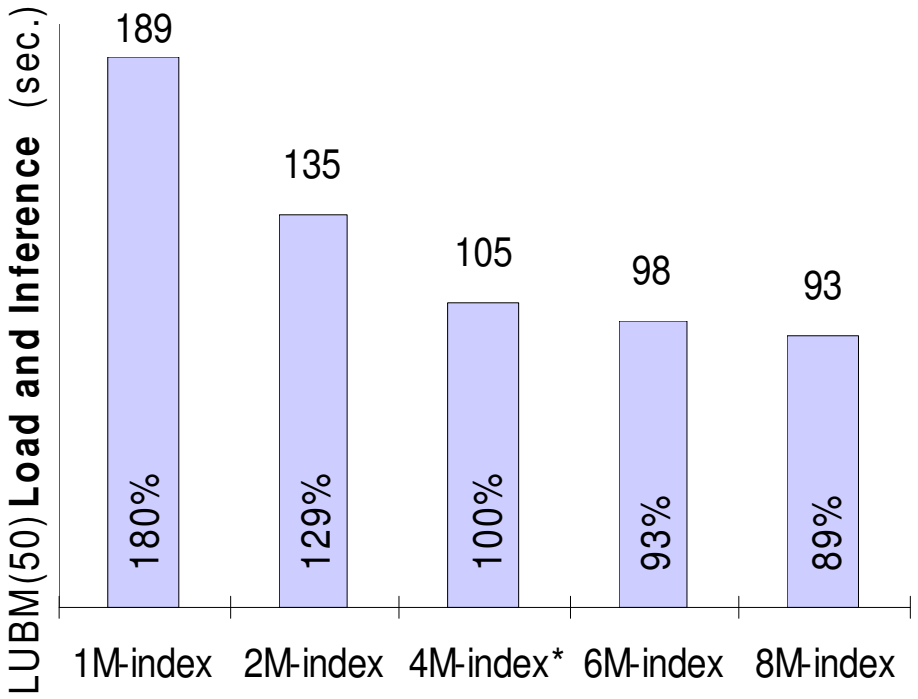
- Q1: Pattern of 11 statement-joins
- Fixed small resultset – retrieval time close to 0
- The query evaluation time is almost constant

OWLIM Performance: Query Answering (II)



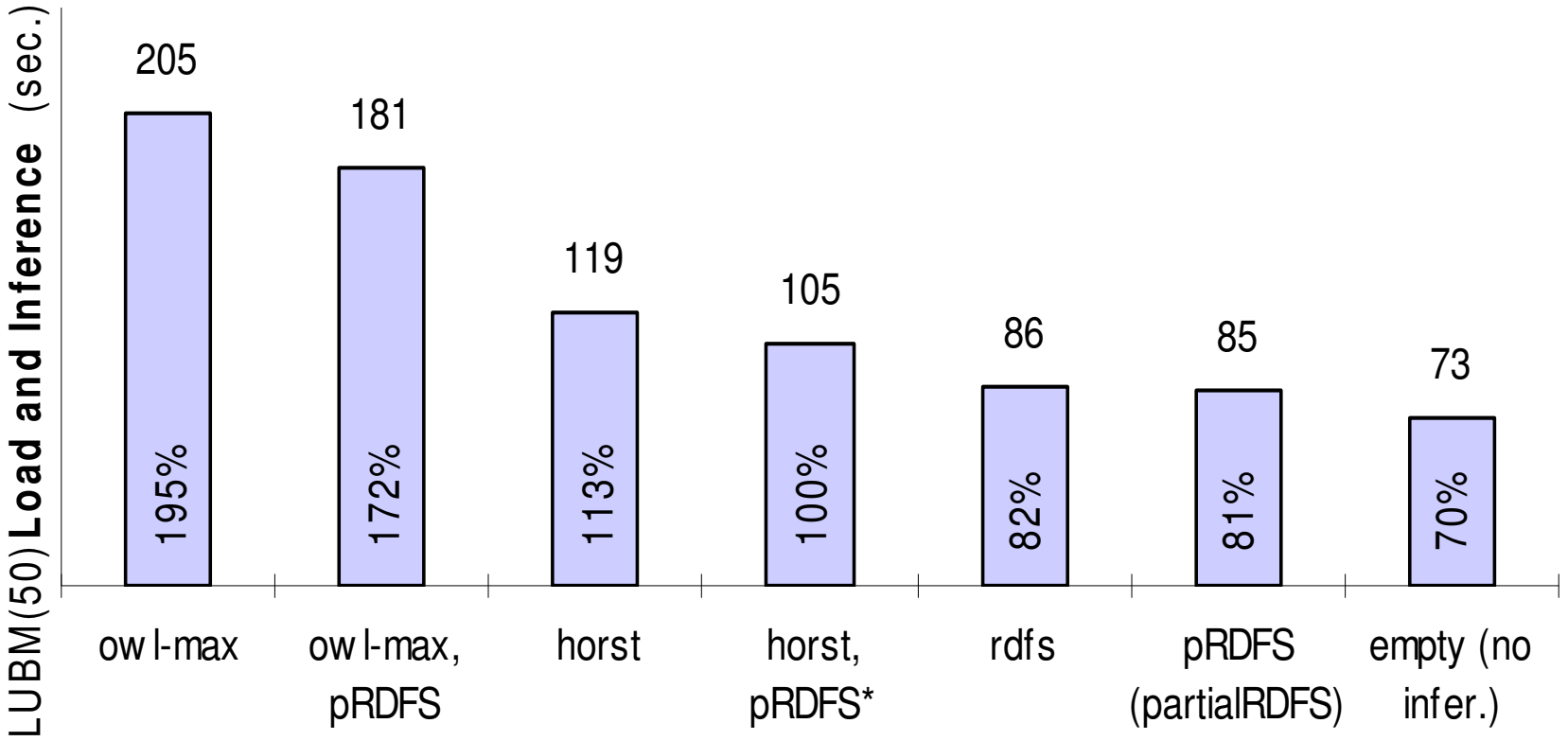
- Q2: Pattern of 12 statement-joins and LIKE "*xyz*" literal constraint
- Large result set which grows linearly with the repository
- The query evaluation and retrieval time also grows linearly

LUBM(50,0): The Optimal Index Size Analysis

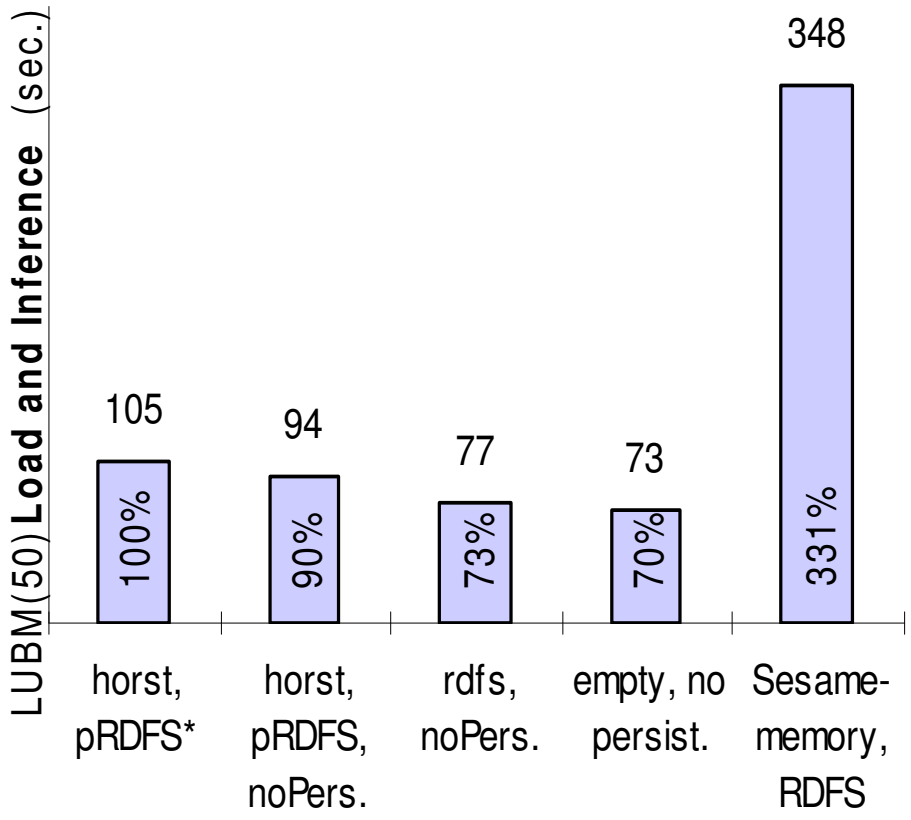


- As expected, larger index sizes lead to better performance.
- Critical for the performance on LUBM(50,0) is the border line between 1 and 2 millions of index entries.
- Index sizes larger than the default setting (4 million entries, 64MB of RAM) seem to deliver very little improvement.

LUBM(50,0): Rule-set and Inference Mode

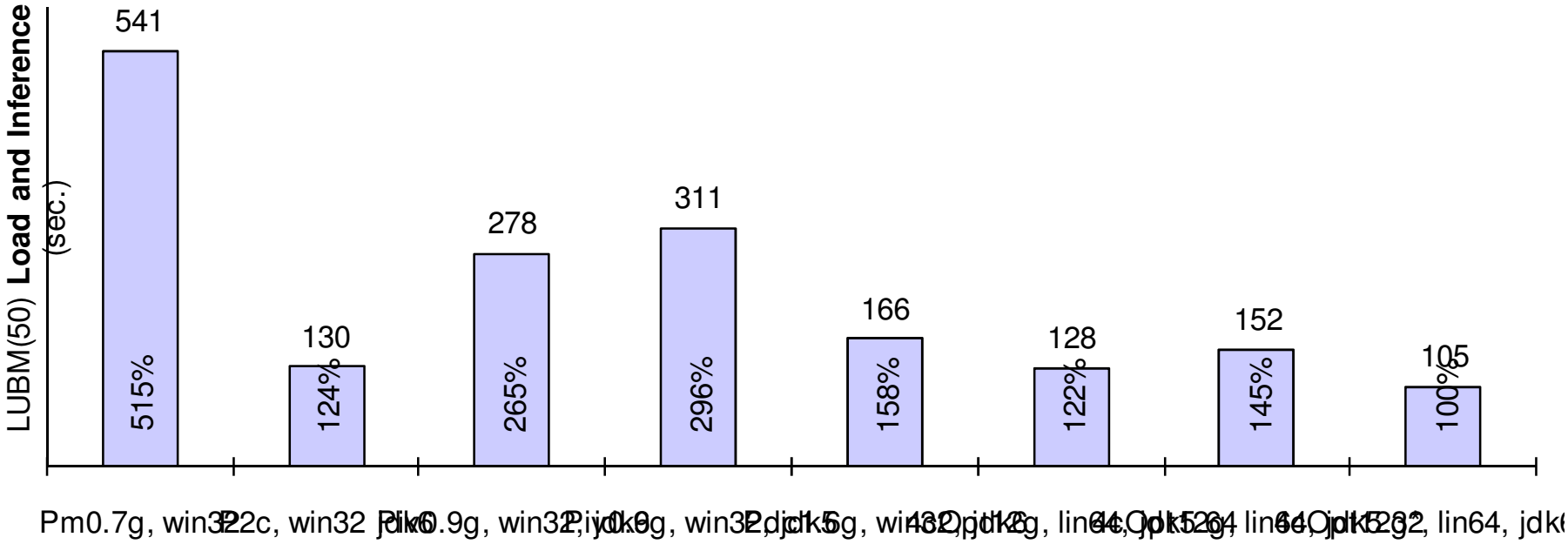


LUBM(50,0): The Impact of the Persistence



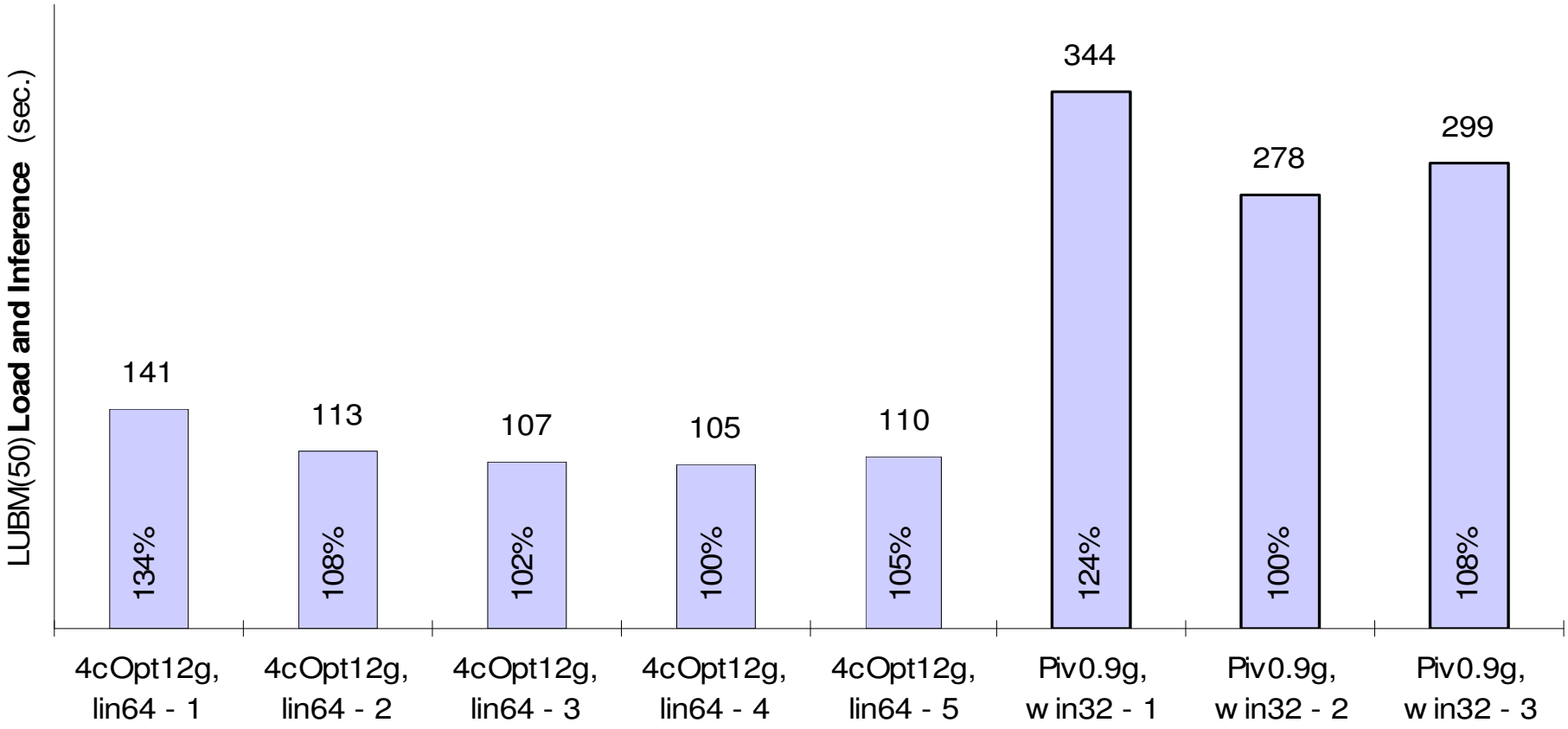
- The parsing and building of in-memory representation takes about half of the load time (see "empty, noPers.")
- The persistence takes 10% on top of the time for loading (see "horst, pRDFS, noPers.");
  - Or flat 10 sec., 10% on top of the time for parsing;
- "Sesame-memory" is four times slower than the "rdfs, noPers." setup of OWLIM.
  - It performes faster on 32-bit JDK 1.5 (305 s.); the time on 64-bit JDK was even higher.

LUBM(50,0): Different Hardware, OS, JDK



Refer to OWLIM's system documentation for analysis and comments.

LUBM(50,0): Multi-threaded Inference



Refer to OWLIM's system documentation for analysis and comments.

Presentation Outline

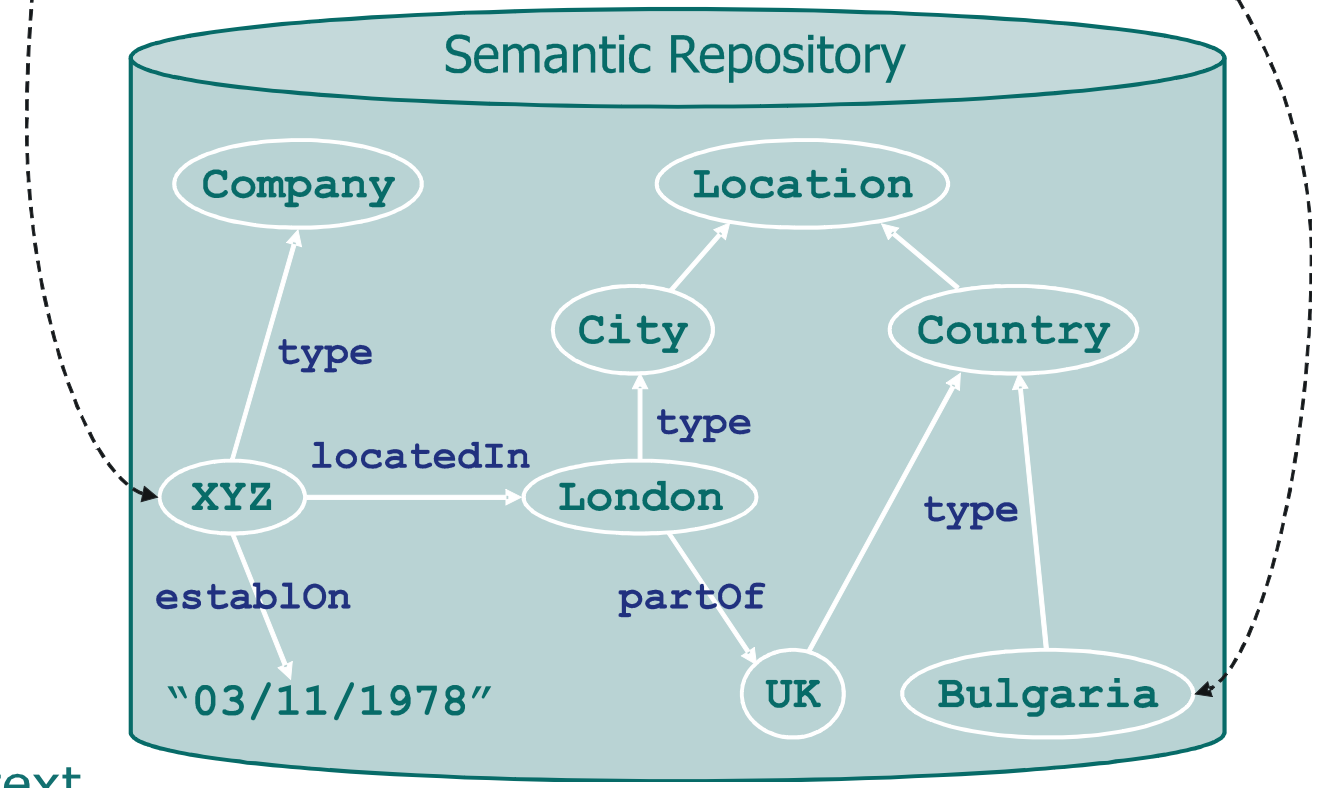
- Sirma and Ontotext
- Introduction to Semantic Web and Ontologies
- OWLIM: the “semantic database”
- **KIM: the “semantic search engine”**
  - CORE Search and Timelines Demo
- Applications

The KIM Platform

- A platform offering services and infrastructure for:
  - (semi-) automatic **semantic annotation** and
  - **ontology population**
  - **semantic indexing** and **retrieval** of content
  - **query** and **navigation** over the **formal knowledge**
- Based on an **Information Extraction** technology
- **Aim:** to arm Semantic Web applications
  - by providing a **metadata generation** technology
  - in a **standard, consistent,** and **scalable framework**

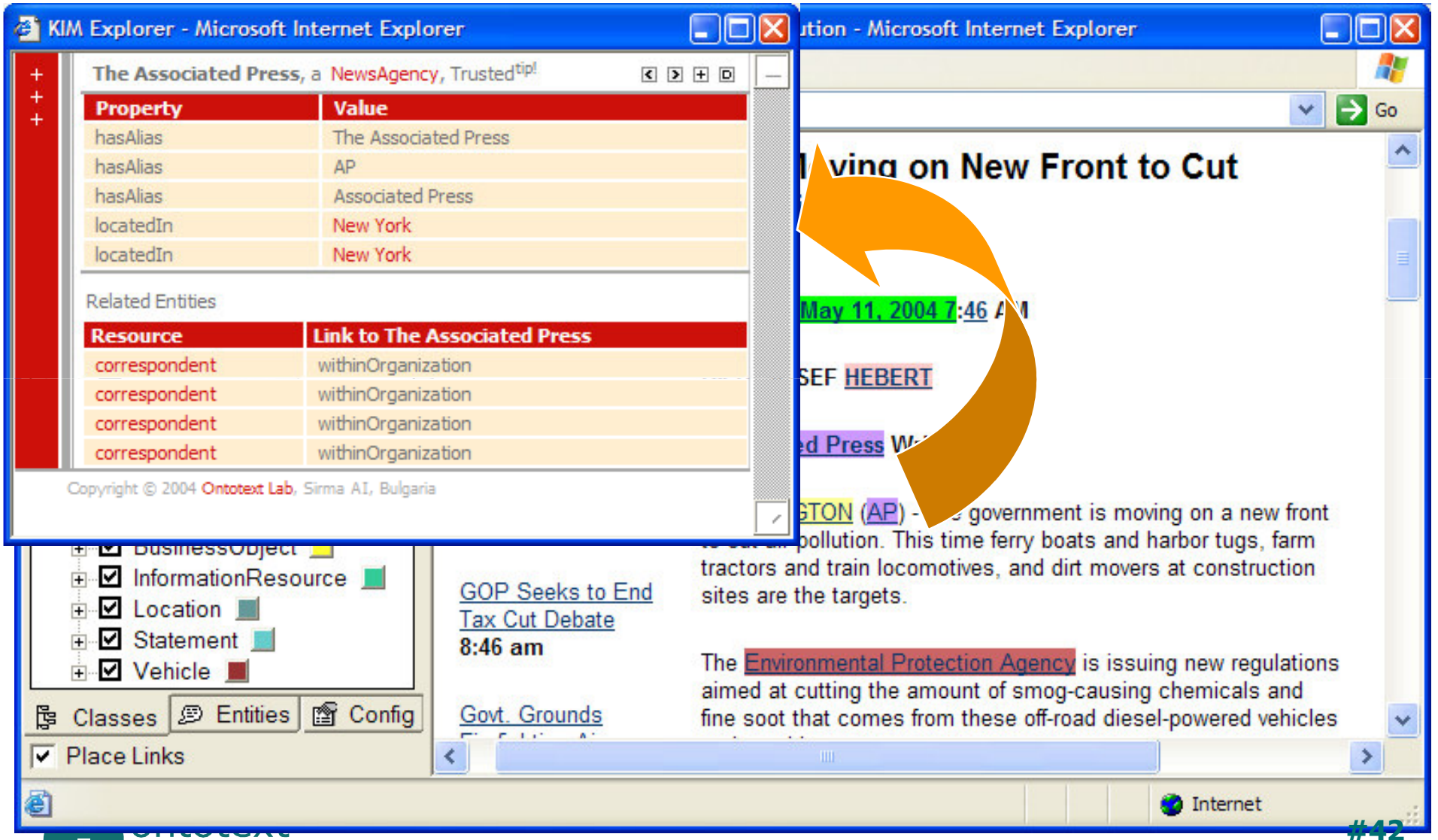
Semantic Annotation: 2001

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

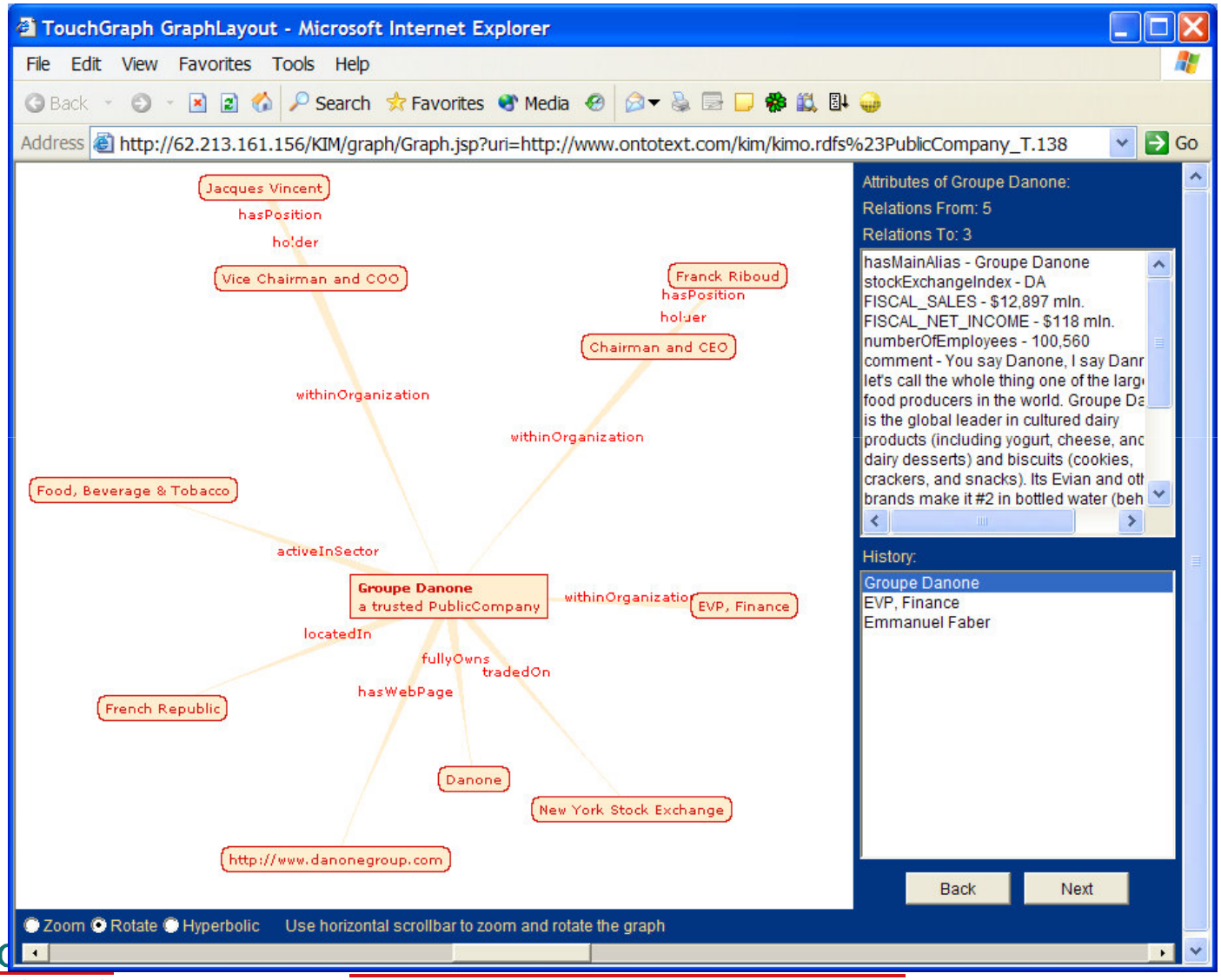


Simple Usage: Highlight, Hyperlink, and …

Simple Usage: … Explore and Navigate



Simple Usage: … Enjoy a Hyperbolic Tree View



How KIM Searches Better

KIM can match a **Query:**

Documents about a telecom company in Europe, John Smith, and a date in the first half of 2002.

With a document containing**:**

At its meeting on the 10th of May, the board of Vodafone appointed John G. Smith as CTO

The classical IR could not match:

- Vodafone with a "telecom in Europe“, because:
  - Vodafone is a mobile operator, which is a sort of a telecom;
  - Vodafone is in the UK, which is a part of Europe.
- 5th of May with a "date in first half of 2002“;
- “John G. Smith” with “John Smith”.

CORE: Co-occurrence and Ranking of Entities

Be able to efficiently query for:

- **Number of appearances** and **popularity** of entities

  Q1: How often has a company appeared in the international business news during a given period ?

- **Co-occurrence** of entities

  Q2: Give me the people that co-appear with telecom companies

- Combination of the above with **semantic queries** and **Full-Text Search,** time-constraints, etc.

  Q3: Q2 + where the documents from 2004 contain "fraud" and the company is located in South-east Europe

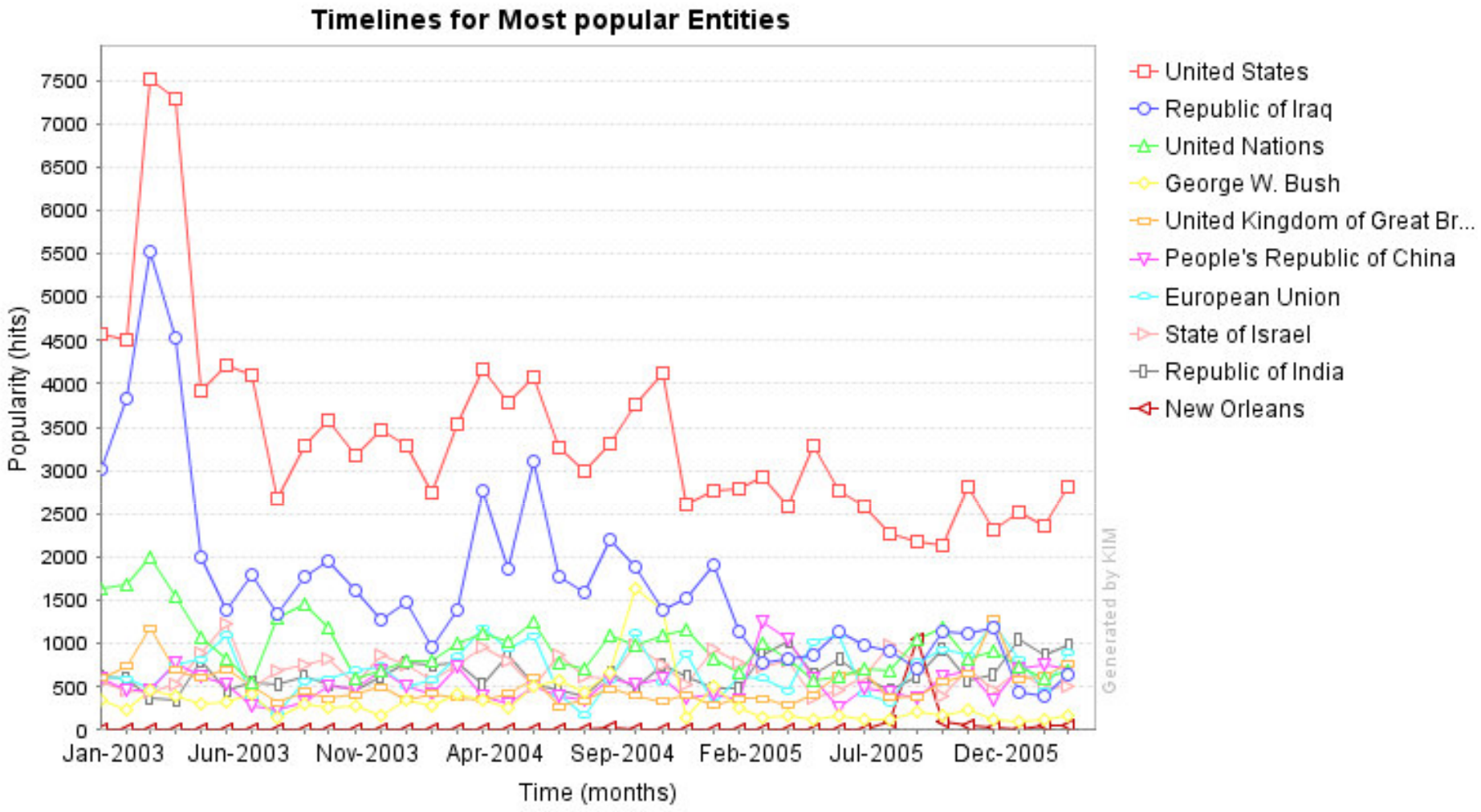
- **Popularity ranking**

  Q4: the 5 most popular persons for each month in 2005, based on news for South Africa, showing a timeline of their ranking

CORE: Scale and Applications

- Allow such queries in *efficient* manner over data with cardinality:
  - 10^6 entities/terms in 10^7 documents (tens of millions)
  - 10^2 entities occurring in an average document
  - managing and querying efficiently 10^9 entity occurrences!
- **Detection of "associative" links** between entities
  - based on co-occurrence in context;
  - an alternative to extraction of "strong links" by parsing local context
- **Media monitoring**: the ranking is as good/relevant/representative as the set of documents is
- **Computing timelines** for entity ranking or co-occurrence
  - "How did our popularity in the IT press changed during June" (i.e. "What is the effect of this 1.5MEuro media campaign ?!?")
  - "How does the strength of association between organization X and RDF changes over Q1 ?"

Timelines Result

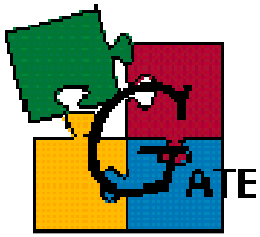


Document Filter: **ALL** docs, containing **Keywords :** (none) **Entities :** (none)
Time Period: **01/01/2003** to **31/03/2006** Granularity: **Month**
Options: display **2** topmost entities of type **Entity** for **each time unit**

KIM is Based On…

KIM is based on the following open-source platforms:

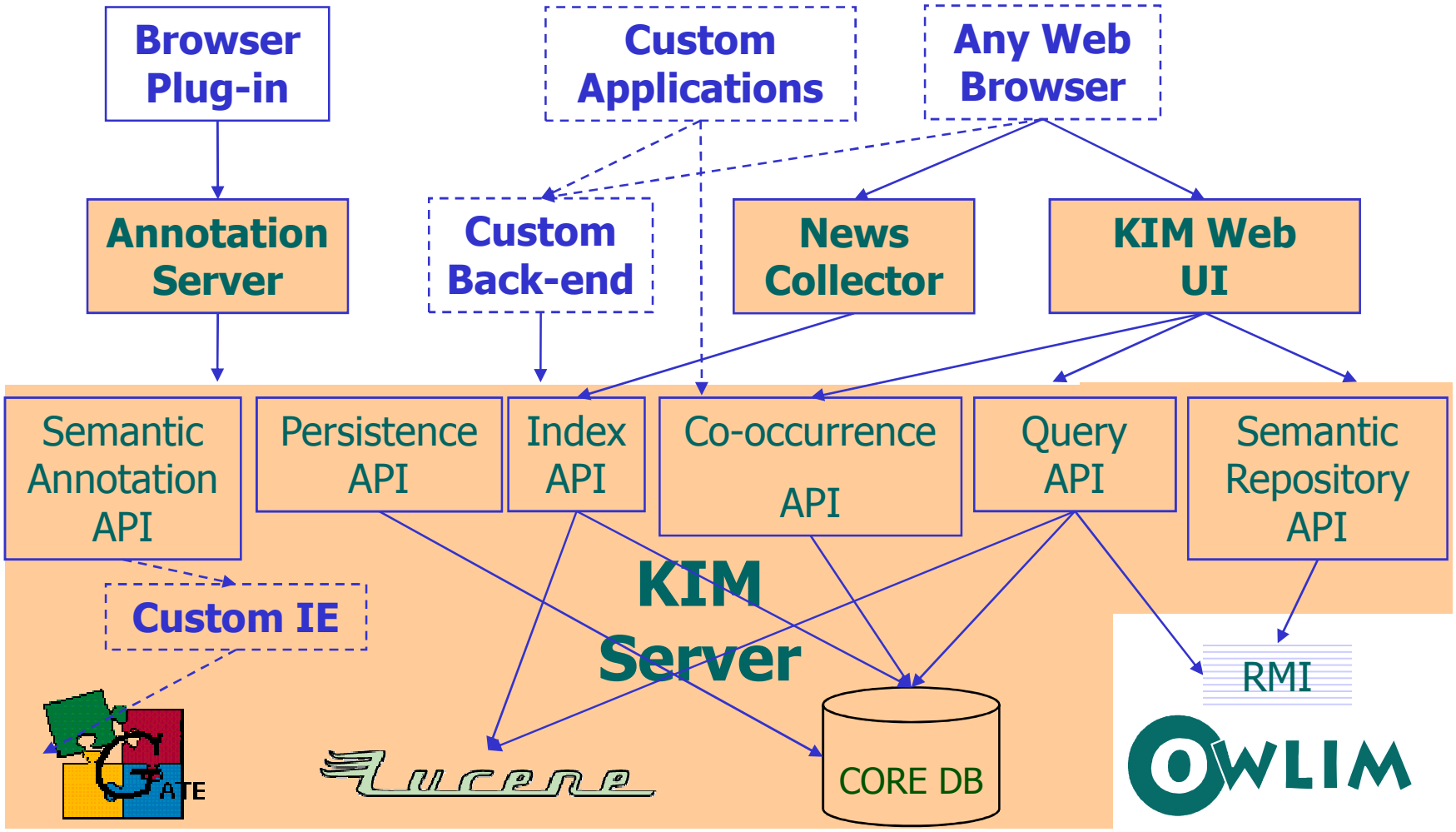
- **GATE** – the most popular NLP and IE platform in the world, developed at the University of Sheffield. Ontotext is its biggest co-developer. www.gate.ac.uk and www.ontotext.com/gate



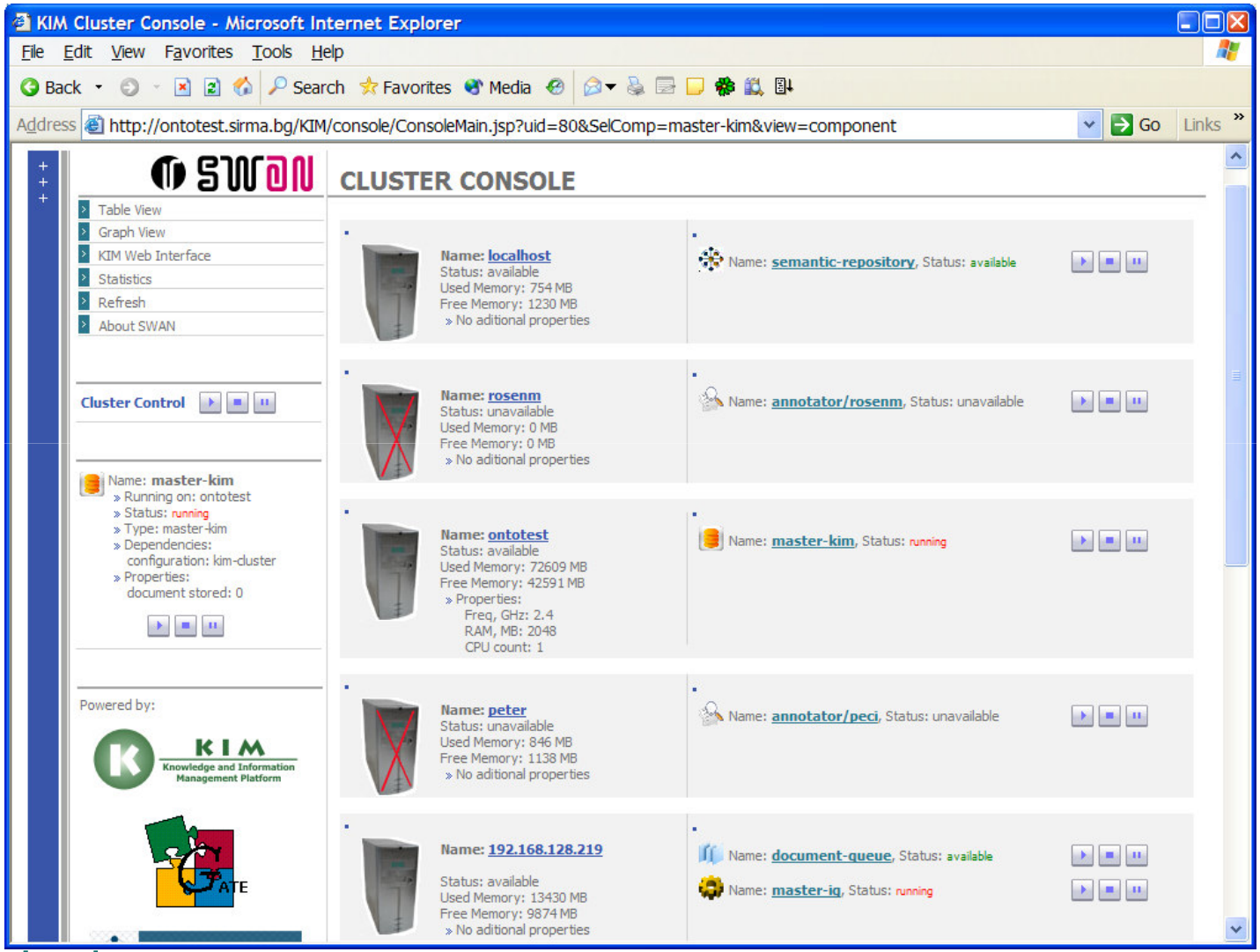
- **Sesame** – RDF(S) repository by Aduna B.V. Ontotext is its biggest co-developer. www.openrdf.org

- **Lucene** – an open-source IR engine by Apache. jakarta.apache.org/lucene/

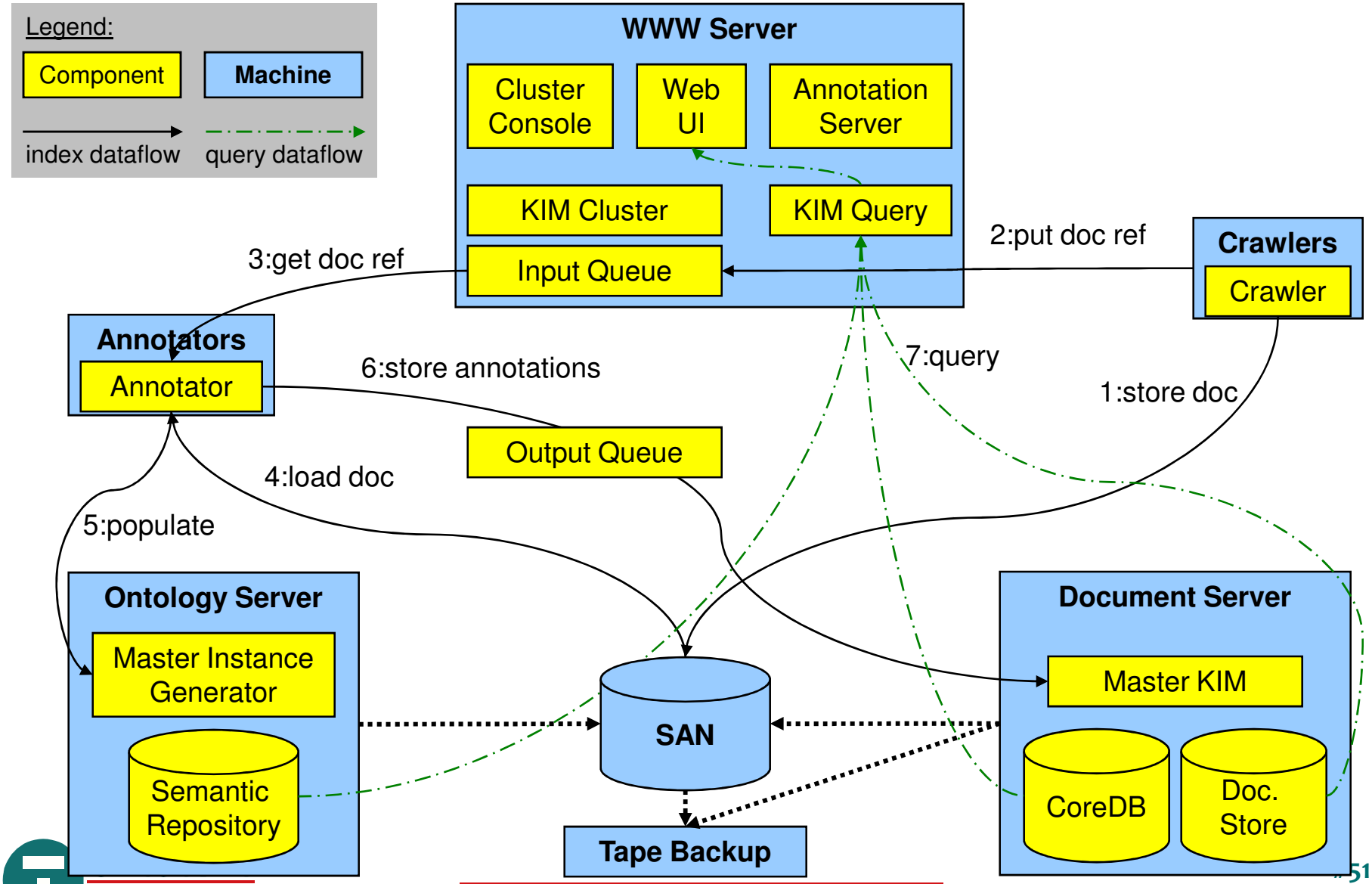
KIM Architecture



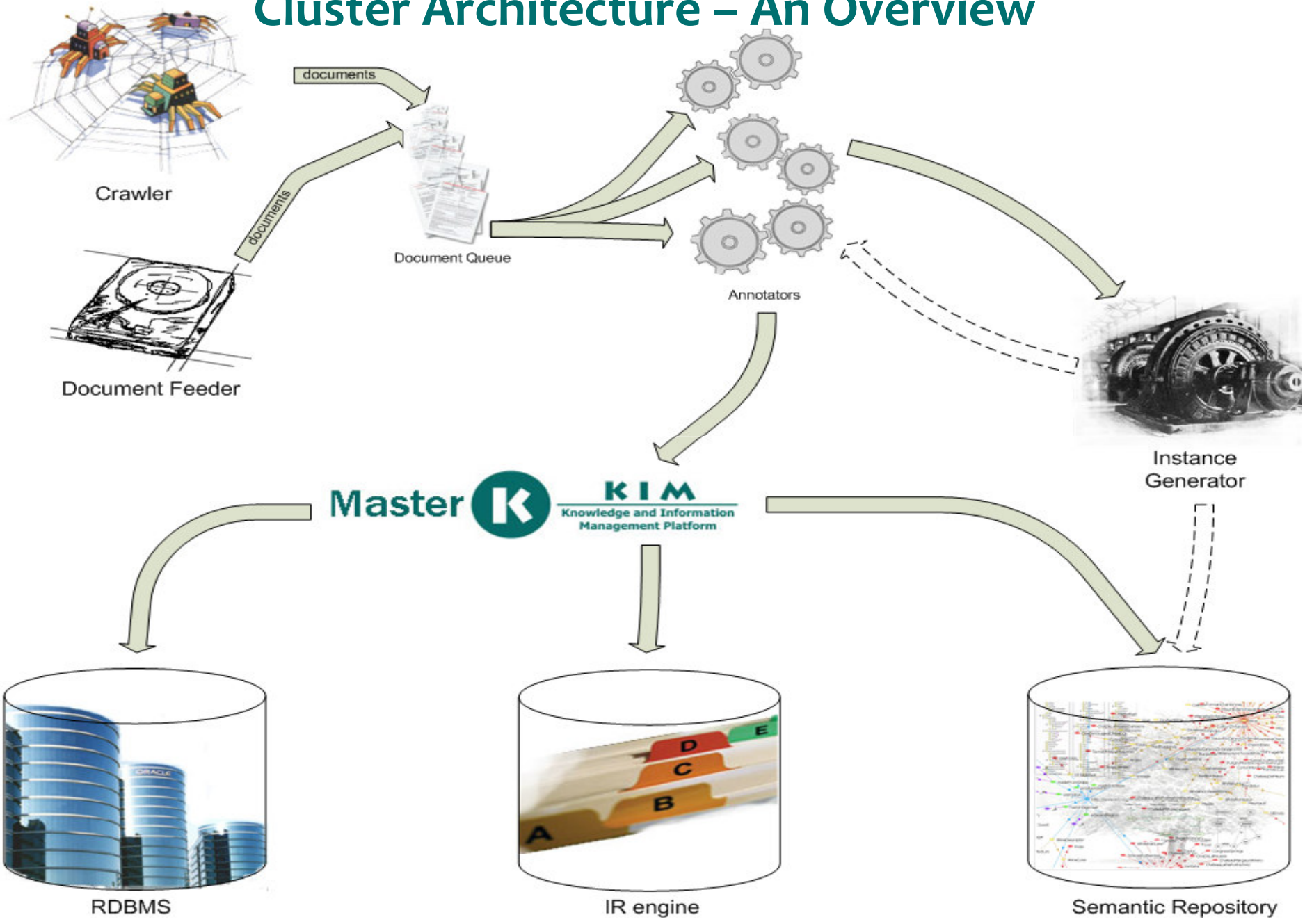
KIM Cluster Architecture



Sample Cluster Configuration



Cluster Architecture – An Overview



Demo with 1 Million Documents

- CORE Demonstration:


Hang Glider Crash Lands
Was Attempting Flight To LaBelle

» **1 million documents**
» International **News** Articles (2002-2006)
» Approx. 1000 articles per business day

- Statistics

» More than **1 million entities** (50K pre-populated)
» Described in about **10 million RDFS/OWL triples**
» On average, 30 entities occurring per document
» Number of occurrences: **27 M**

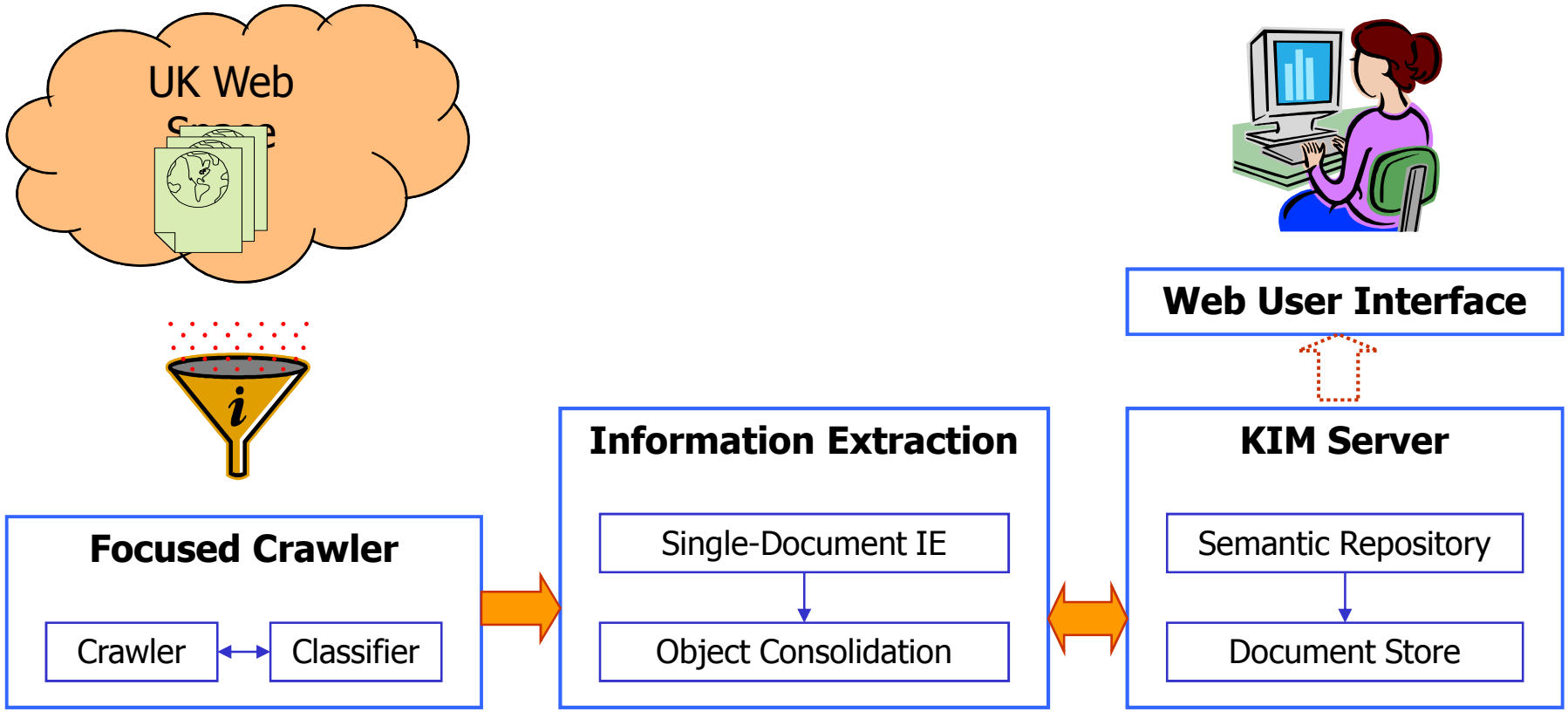
Presentation Outline

- Sirma and Ontotext
- Introduction to Semantic Web and Ontologies
- OWLIM: the “semantic database”
- KIM: the “semantic search engine”
  - CORE Search and Timelines Demo
- **Applications**

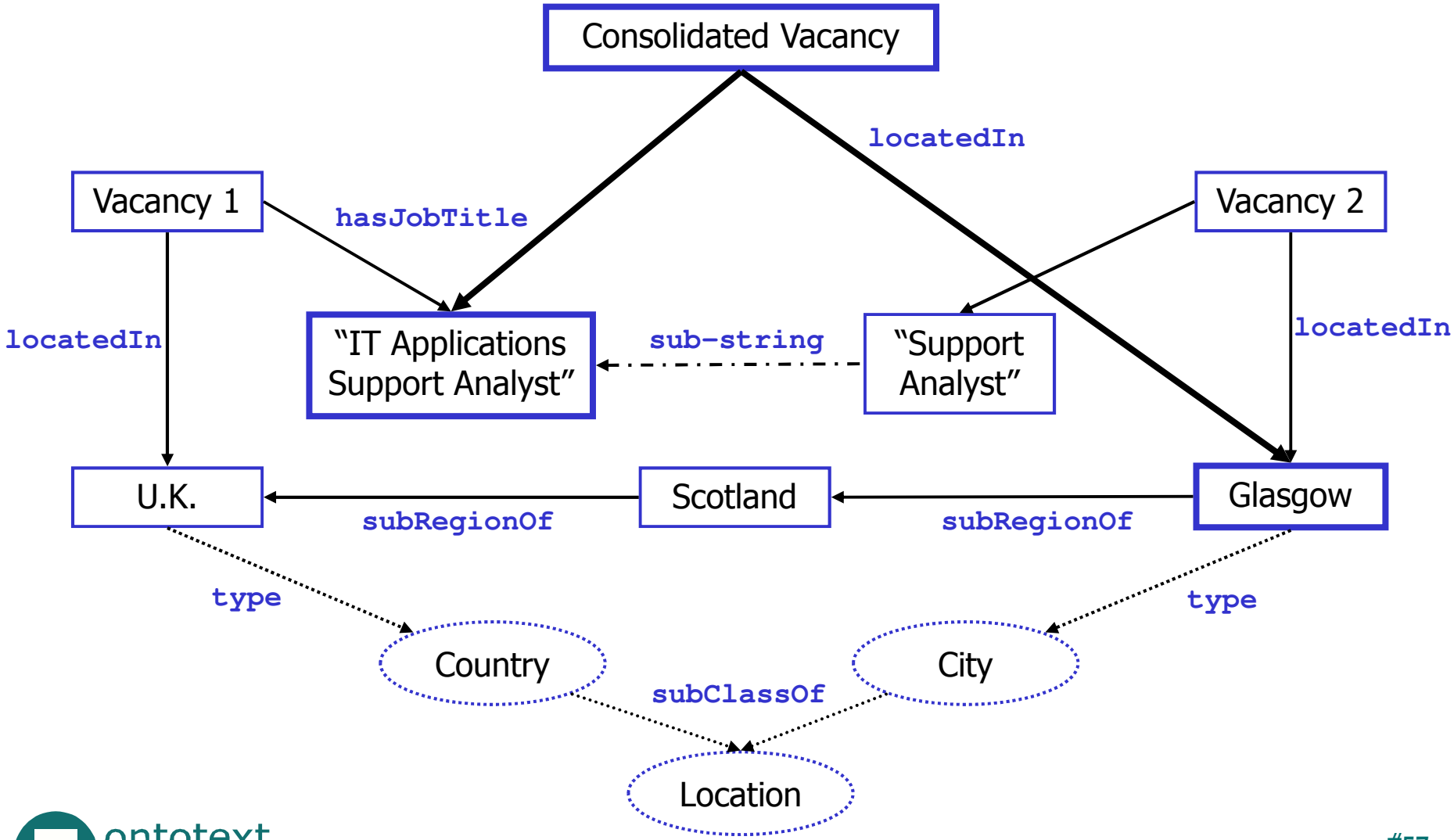
JOCI

- “Jobs & Contacts Intelligence”, Innovantage, Fairway Consultants
- Gathering recruitment-related information from web-sites of UK organizations
- Offering services on top of this data to recruitment agencies, job portals, and other.
- Based on KIM
- Includes focused crawling and many other techniques
- Launched April 2005
- Sirma is shareholder in Fairway Consultants
- First round of investment Jan 2007 (FWI, HSBC)

JOCI Dataflow



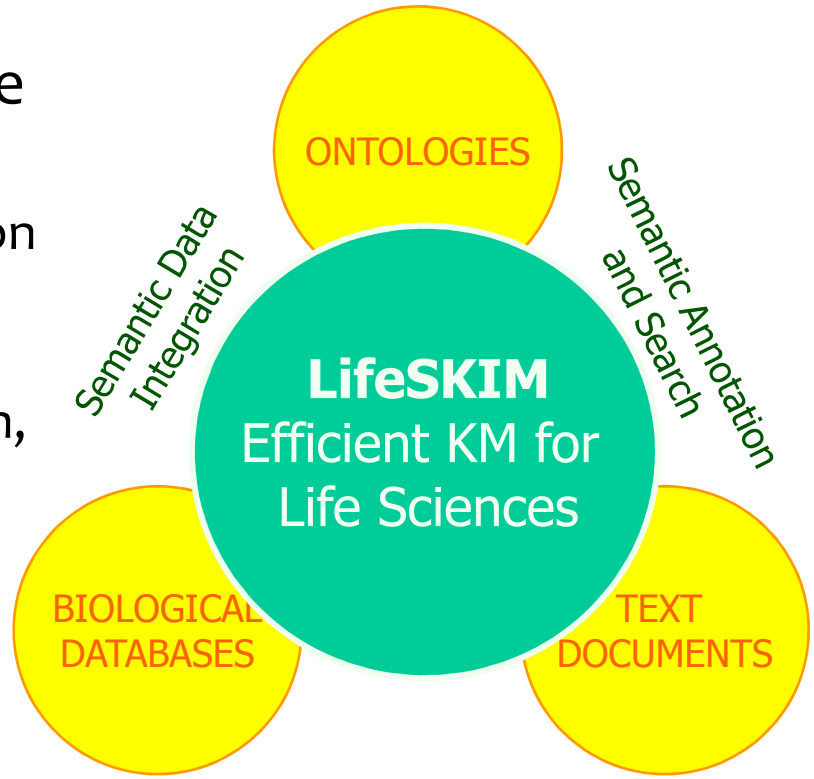
JOCI: Vacancy Consolidation/Matching



LifeSKIM

LifeSKIM group in Ontotext, provides KM technology for the Life Sciences:

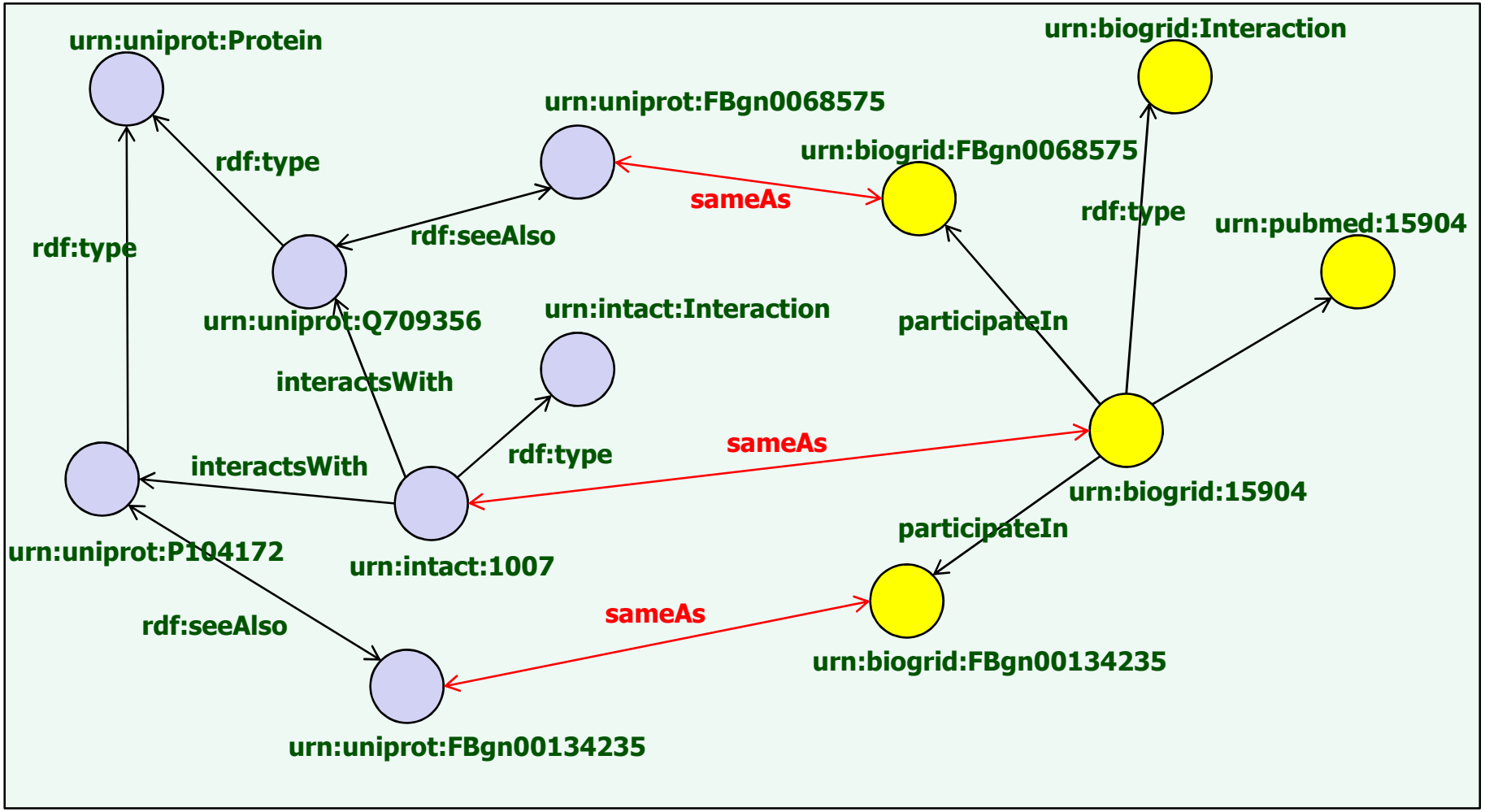
- **Semantic data integration:** integration of multiple structured and semi-structured biological data sources.
- **Text-mining** and semantic annotation, linking text to the structured data
- **Semantic search**, navigation, hyperlinking, visualization, etc.



Existing technology is adapted:

- OWLIM, ORDI (structured data)
- KIM, GATE (text-mining & search)

Semantic Data Integration Benefits



Thanks!

Ontotext Lab: core semantic technologies

- Employing most recent research results
- Outstanding performance and scale
- Based on open formats

?

http://www.ontotext.com/

По пътя ...

Търсим най-добрите

Semantic Search and Data Bases Developer

Semantic Web Services (Senior) Developer

Web User Interfaces Developer

Web Mining Systems Developer

Database & System Admin

за да продължим заедно!

http://www.ontotext.com/jobs.html