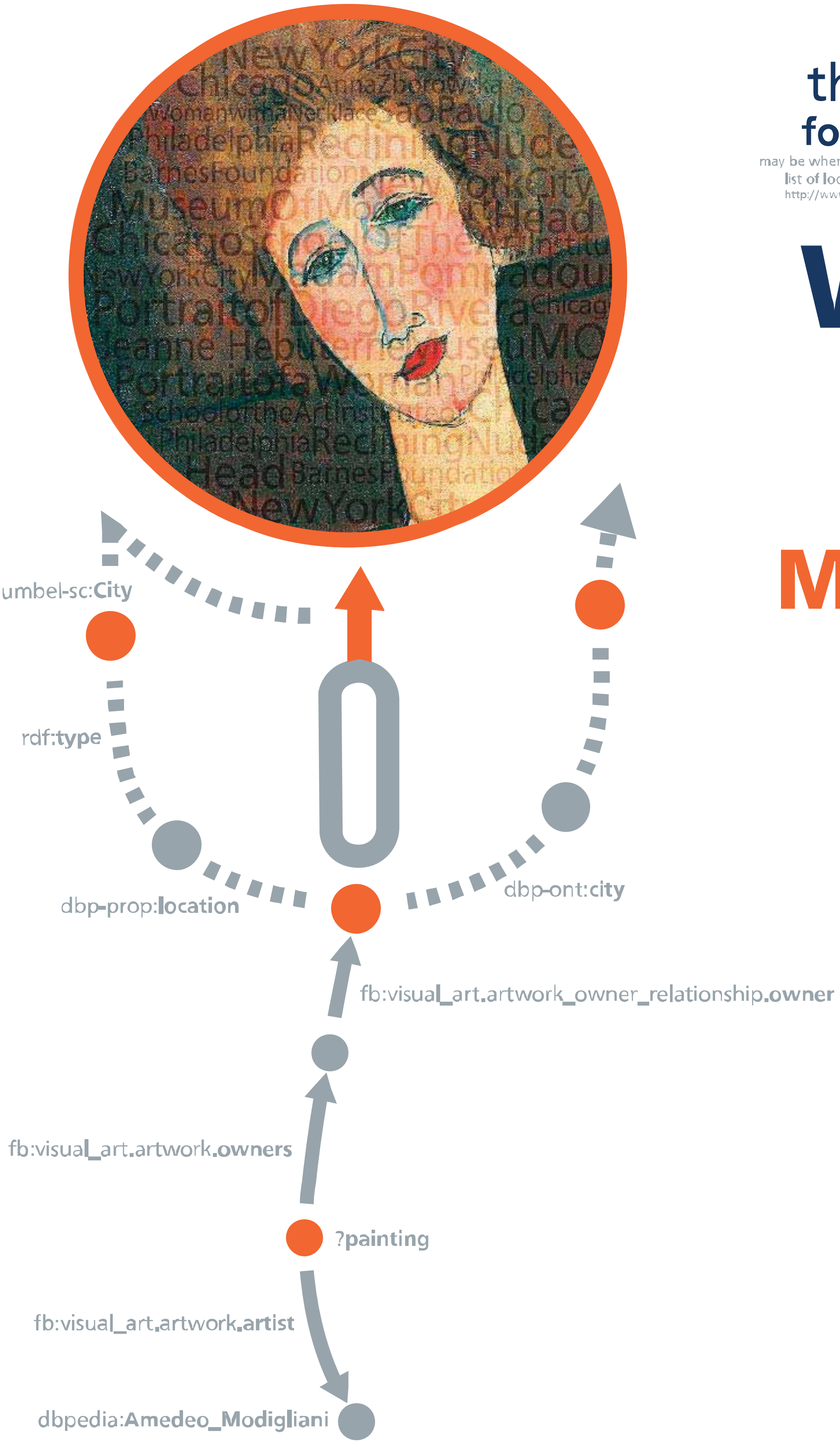
ReadWriteWeb: "...

the tipping point
for the **Semantic Web**

may be when one can ... deliver – using Linked Data –a comprehensive
list of locations of original Modigliani art works around the world"
http://www.readwriteweb.com/archives/the_modigliani_test_for_linked_data.php


umbel-sc:City
rdf:type
dbp-prop:location
dbp-ont:city
fb:visual_art.artwork_owner_relationship.owner
fb:visual_art.artwork.owners
?painting
fb:visual_art.artwork.artist
dbpedia:Amedeo_Modigliani

WHERE CAN one SEE Modigliani paintings?

See page 17 to learn how
Ontotext brings the Semantic
Web closer to its tipping point

ontotext

**We develop core semantic technology
and text mining solutions**

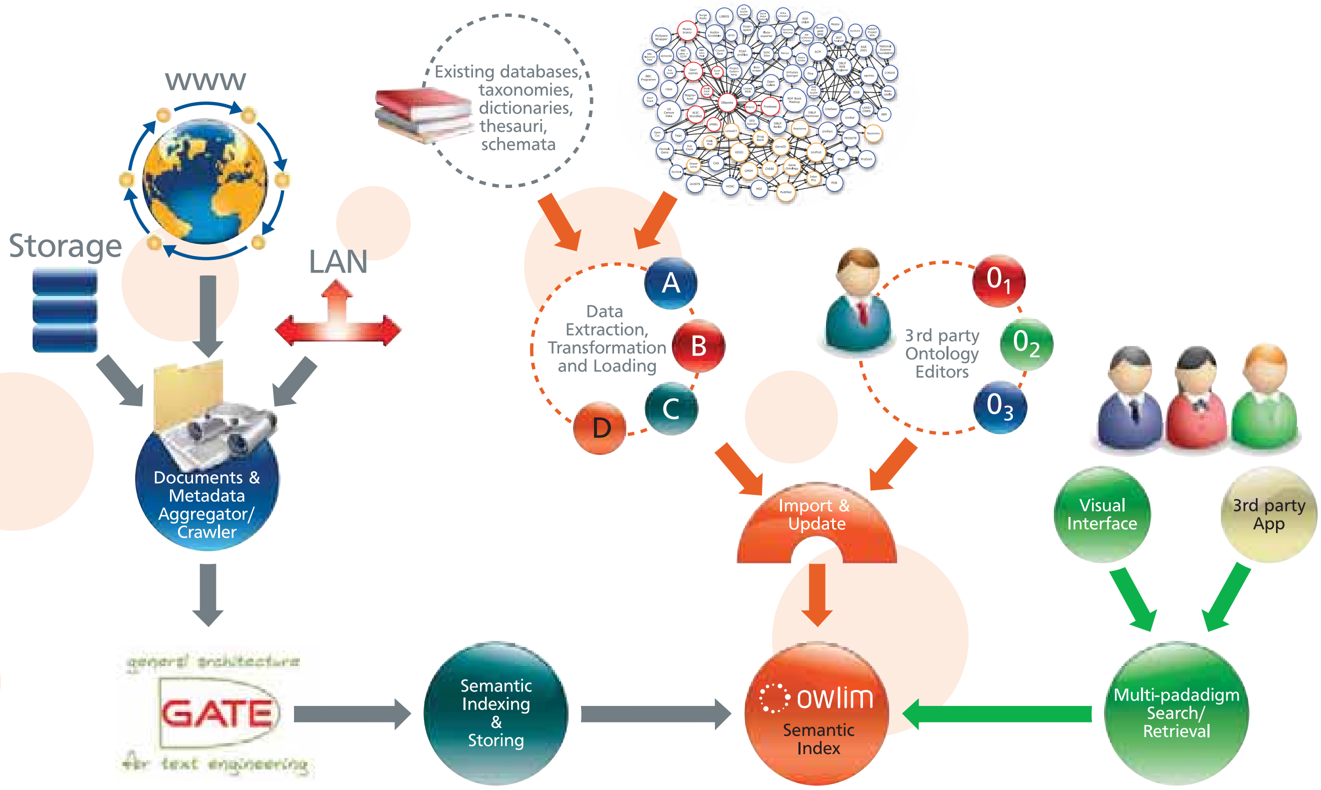
We link your data, your content and the Web

IN 10 WEEKS WE CAN BUILD A SOLUTION THAT:

- integrates **10 databases** with the linked data cloud (p.17)
- mines **10 million documents** and web pages (p.10,18)

AND LETS YOU SEARCH AND NAVIGATE OVER THIS INFORMATION

- in **10 different ways** (p.10)
- from a **$10K server**



WWW
Storage
LAN
Documents & Metadata Aggregator/ Crawler
general architecture
GATE
for text engineering
Semantic Indexing & Storing
Existing databases, taxonomies, dictionaries, thesauri, schemata
A
B
C
D
Data Extraction, Transformation and Loading
3rd party Ontology Editors
O_1
O_2
O_3
Import & Update
owlim
Semantic Index
Visual Interface
3rd party App
Multi-padadigm Search/ Retrieval

Our value proposition

THE BUSINESS CASE

- Deploy **flexible management** of data and content
- Allow **query variation,** for instance different vocabulary, syntax and level of generality
- Manage **heterogeneity** across multiple sources
- Discover implicit facts via **data semantics interpretation**
- Ensure **time and cost-effective** technology adaptation

OUR OFFER

We provide products and methodology that enable you to:

- **Integrate structured data** from a variety of sources
- **Collect information** from proprietary sources and the Web
- **Interlink text and data** to tag and index the text
- Uncover and **extract implicit facts,** interpreting the semantics of text and data
- Provide **multi-modal search** and navigation
- Process data with **evolving structure** as well as **sparse data**

OUR PROMISE

- We **save you time and money** when it comes to management of text and data from **multiple sources**
- We deliver **more answers with less effort**
- We find **hidden links**, matching facts scattered across huge volumes of diverse information
- We help **you access your information** and the Web

Our Vision and Products

DATA MANAGEMENT VISION

DATA MANAGEMENT NEEDS A CHANGE

- RDBMS and XML are inappropriate or too expensive for many tasks

RDF OFFERS A GOOD ALTERNATIVE

- RDF is suited for dynamic schemata and sparse data
- Semantic Web standards are designed for diversity

LIGHT-WEIGHT ONTOLOGIES AND REASONING ALLOW EFFICIENT DATA INTEGRATION

- Queries do not need to match the assertion syntax and vocabulary; thus, one can query unfamiliar data
- Semantics strengthens the links in data integration
- To avoid the AI pitfalls, ontologies should be as easy to manage as database schemata

TEXT MINING VISION

TEXT MINING INCREASES THE VALUE OF UNSTRUCTURED CONTENT

- Information extraction allows automated recognition of objects and extraction of facts from text at a reasonable accuracy and cost

INTERLINKING TEXT AND DATA ENABLES MORE EFFICIENT SEARCH AND NAVIGATION

THERE IS NO UNIVERSAL TEXT MINING TECHNOLOGY

- We offer practical solutions, designed to match specific requirements
- And a process for capturing the requirements, samples and feedback

PRODUCTS

- **OWLIM** Semantic Database (p.6)
- **KIM** Semantic Search Platform (p.10)
- **Web Mining Framework** (p.18)
- **LifeSKIM** Platform for Biomedical KM (p.14)

FORM FACTOR

LICENCES FOR SOFTWARE

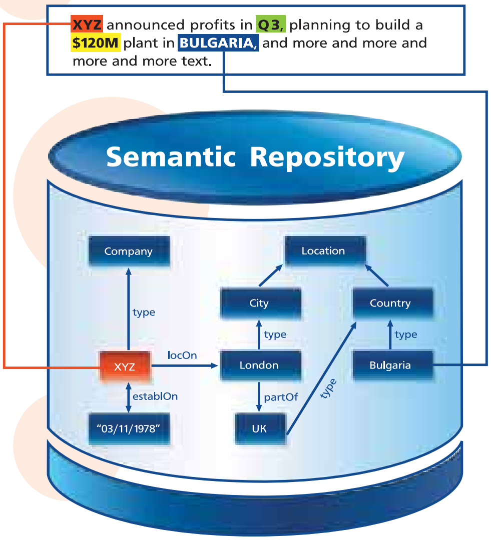
- Licences based on server capacity
- OEM and re-seller partnerships
- Free copies for research and evaluation purposes

PROFESSIONAL SERVICES

- Research, Analysis and Evaluation
- Development and Customization

SOFTWARE AS A SERVICE (SAAS)

- Semantic annotation services
- Virtual semantic repositories
- Virtual linked data "views" (p.17)


XYZ announced profits in Q3, planning to build a $120M plant in BULGARIA, and more and more and more and more text.
Semantic Repository
Company
Location
City
Country
type
type
type
XYZ
locOn
London
Bulgaria
establOn
partOf
type
"03/11/1978"
UK

Semantic Annotation in KIM (p.10)

Why choose our team

WE ARE BEST POSITIONED TO DEVELOP BESPOKE TEXT MINING SOLUTIONS

- We have established a solid reputation as a solution and technology leader, which is continuously reflected in our customer base, scientific standing and the developer community around our products (http://gate.ac.uk)
- To deliver the best expertise in consultancy, we engage the brightest semantics and text analysis gurus, e.g. the GATE developers with whom we cooperate closely

HOW DO WE SUPPORT OUR CUSTOMERS?

- Customers with their own development resources or preference for open technologies are provided with a strong open source and free licence track and we support the community around it
- Our solution cycle is completed by customization, maintenance and training services

PARTNERS AND CUSTOMERS

Austria: IRF, Seekda, SemSphere
Bulgaria: NetInfo
France: Mondeca, WHO
Finland: Profium
Germany: SAP, Volz Innovation, Wikimedia Deutschland
Italy: CEFRIEL, BPEng
Slovenia: Cycorp
South Korea: Saltlux
Spain: Telefonica, iSOCO, Atos Origin
Sweden: AstraZeneca
Switzerland: Basel Institute on Governance, ii4sm, Google Labs, Health On the Net
UK: BBC, BT, Press Association, Innovantage, The National Archive, System Simulation
USA: Panaton, Ronsmap, TopQuadrant

SOME OF OUR CLIENTS

REFERENCE CLIENTS AND IMPLEMENTATIONS

Applications ranging from clinical study analysis, based on multiple data sources, to target identification for drug discovery (p.14)

FINANCIAL INTELLIGENCE UNITS

Asset recovery and anti-corruption intelligence based on data from the Web, proprietary databases and data feeds, e.g. Dow Jones (p.10)

VERTICAL SEARCH ENGINE FOR CARS

Focus on gathering all vehicle offers from the US; Market analysis on all offer details: regions, vehicle types and many other features (p.18)

CONTENT DELIVERY FOR PUBLISHERS

Matching incoming news with customer profiles to deliver relevant articles and photos in the right format (p.7)

A TOP-5 TELEVISION PRODUCTION COMPANY

Cluster configuration of BigOWLIM handles millions of requests per day, serving as a back-end for the website of one of the largest TV companies in the world (p.9)

OWLIM

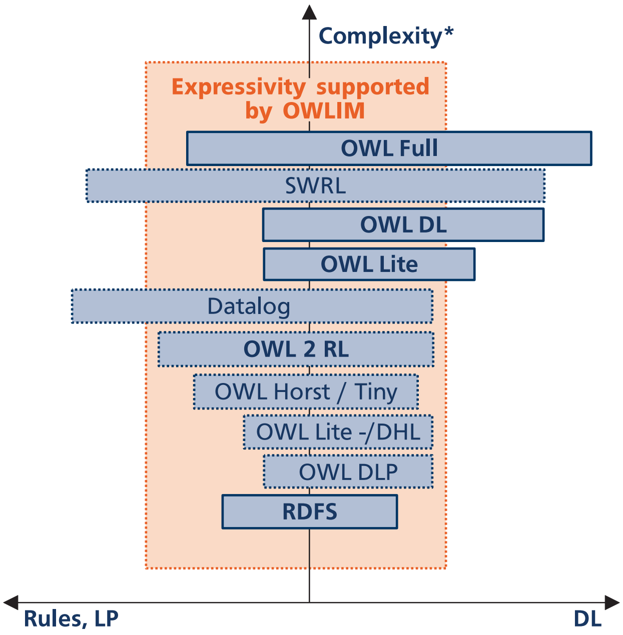
SEMANTIC REPOSITORIES

Semantic repositories or semantic databases are DBMS – their main function is to store and query structured data. The essential difference from relational DBMS is that semantic repositories can infer non-explicit facts using:

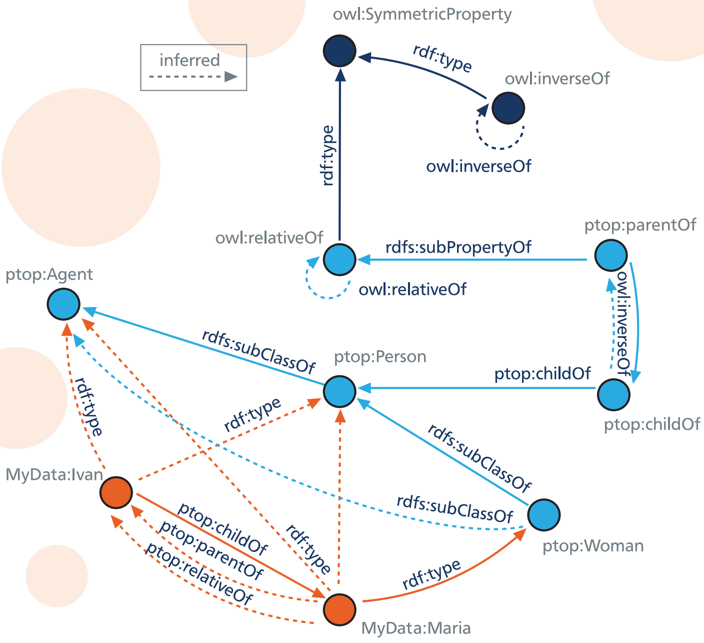
- More expressive schema definitions (ontologies), encoding some of the semantics of the data
- Inference mechanisms to interpret stored data

MORE INTELLIGENT QUERY ANSWERING

Semantic repositories offer greater analytical power. A query can match criteria and return results based on data that differs from the query patterns, but bears relevant meaning. For instance, a query pattern "Maria, relative-of, ? x" can return Ivan as a match based only on the assertion "Ivan,child-of,Maria" (see the graph below).



Naive OWL Fragments Map



OWLIM IS A HIGH-PERFORMANCE RDF DATABASE

OWLIM is a mature, native RDF semantic repository. Its performance, efficiency and robustness allow it to replace legacy database management systems in a very wide range of applications.

OWLIM IS PARTICULARLY SUITABLE FOR:

- Analytical tasks and Business Intelligence (OLAP)
- Integration of heterogeneous and sparse data

OWLIM IS A SCALABLE INFERENCE ENGINE

IT USES RULE-BASED REASONING TO SUPPORT:

- RDFS, OWL Horst and OWL 2 RL
- Custom semantics via rules and axiomatic triples

Robust Semantic Repository

SEMANTIC DATA INTEGRATION

Semantic repositories provide an ideal platform for data-integration because RDF is designed for the management of data created without centralized control:

- New data sources can be adopted with little effort
- Schema changes are easy to accommodate

RDF REPRESENTS A GENERIC DATA MODEL :

- The logical structure of data is not fixed in its physical representation
- Structure and semantics are interpreted, based on RDFS schemata and OWL ontologies

The diagram below illustrates the differences between data representation in a sample relational database model (on the right) and the corresponding RDF model (on the left).

OWLIM IN USE

BIGOWLIM IN LIFE SCIENCES:

Integration of large-scale KB in the LifeSKIM platform, consolidating biomedical databases (p.14)

BUNDLED IN GATE AS AN ONTOLOGY SERVICE

GATE is the most popular text mining platform

INTEGRATED IN PROFIUM METADATA SERVER

BigOWLIM is integrated in Profium Metadata Server, which is heavily used for content delivery in the publishing industry

IN KIM PLATFORM AS A SEMANTIC REPOSITORY

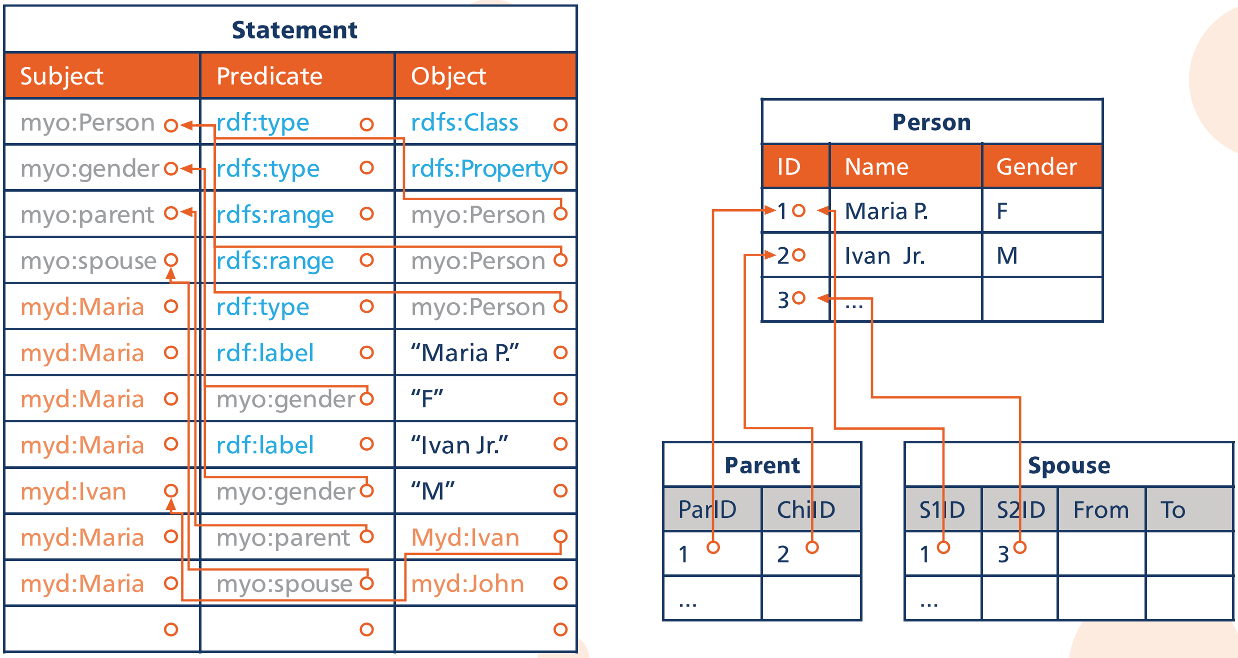
KIM is a semantic annotation and search platform (p.10)

BUNDLED IN TOPBRAID COMPOSER AS A REASONER

TBC is one of the most advanced RDF(S)/OWL editors

THE HEART OF THE DATA LAYER OF THE LARKC PROJECT

LarKC is probably the most ambitiuous large-scale reasoning project - http://www.larkc.eu/



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

Person

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

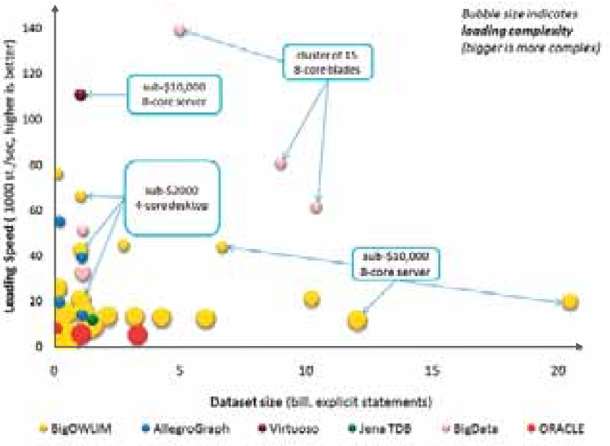
Parent

ParID	ChilD
1	2
...	

Spouse

S1ID	S2ID	From	To
1	3		
...			

OWLIM



Scalable Inference Map

SCALABLE INFERENCE

Benchmarking semantic repositories is a challenging task due to the complex set of relevant criteria and factors.

The map on the left presents the loading speed of some of the most scalable repositories in relation to the size of the dataset and the complexity of the inference involved:

- The best published evaluation results for each engine are presented on the diagram on the left
- Loading in OWLIM and ORACLE includes materialization

OUTSTANDING REASONING PERFORMANCE

SWIFTOWLIM IS THE FASTEST OWL ENGINE ON EARTH!

- It scales up to 10 mill. statements on a 32-bit PC
- It loads LUBM(50) in 42 sec. at 161 KSt./sec.

BIGOWLIM IS THE MOST SCALABLE OWL ENGINE!

- It can load 20 bill. statements on a $9000 server
- Loads LUBM(8000), 1 bill. statements, in 14 hours and answers the queries in one hour on a $2000 workstation

Loading includes parsing, inference, and indexing.

For more details: http://www.ontotext.com/owlim/

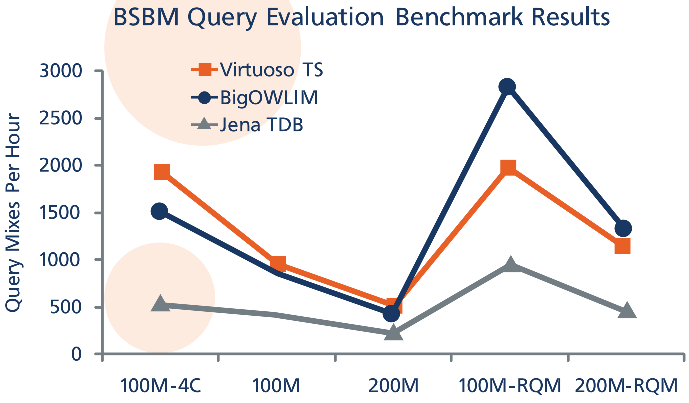
"OWLIM performed very well, while still being able to process OWL DLP, and hence should be the choice for ABox reasoning with lightweight ontologies."
Bock, Haase, Volz: Benchmarking OWL Reasoners. In ARea2008 - Workshop on Advancing Reasoning on the Web

"The reduced query mix (RQM) consists of the same query sequence as the complete mix but without queries 5 and 6. The two queries were excluded as they alone consumed a large portion of the overall query execution time for bigger dataset sizes."
Bizer, Ch., Schultz, A.: BSBM Results for Virtuoso, Jena TDB, BigOWLIM. 30/11/2009.

EXCELLENT HANDLING OF QUERY LOADS

The Berlin SPARQL Benchmark evaluates the performance of query engines in an e-commerce use case: searching products and navigating through related information. Randomized "query mixes" (of 25 queries each) are evaluated continuously against datasets of different sizes. Multiple-clients query loads are simulated as well.

The diagram below presents the performance of BigOWLIM in comparison with some of the most popular engines, as measured by the BSBM team in Nov. 2009.



High Performance RDF Database

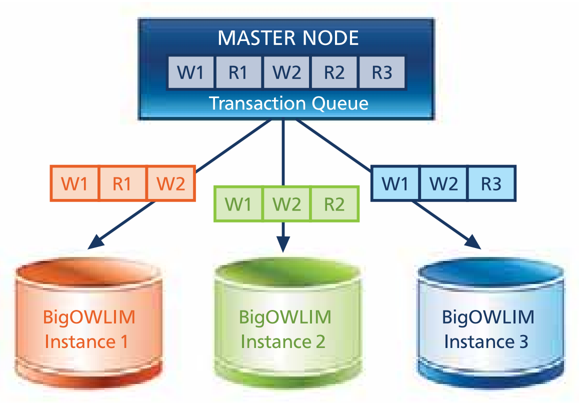
USING OWLIM

END USERS CAN ACCESS OWLIM THROUGH:

- The Web UI routines of Sesame (www.openrdf.org)
- Ontology editors integrated with Sesame
- Forest - the front-end library used in FactForge (p.17)

APPLICATIONS CAN:

- Embed it through the Sesame API
- Access it as a standalone server
- Access it in a cluster setup, as presented on the right



Sample OWLIM Replication Cluster Configuration

SWIFTOWLIM AND BIGOWLIM

There are two major OWLIM species, identical in terms of RDF storage and inference functionality

SWIFTOWLIM: SUITED FOR MEDIUM-SIZED DATASETS AND PROTOTYPING

- Extremely fast loading of data

BIGOWLIM: HUGE VOLUMES OF DATA AND INTENSIVE QUERYING

- Query optimizations ensure faster query evaluation
- Scales better, having lower memory requirements

EASY APPLICATION SETUP

THE DISTRIBUTION OF OWLIM HAS BEEN OPTIMIZED TO MINIMIZE THE EFFORTS AND TIME FOR SETTING UP AN APPLICATION:

- **Getting-started:** a sample application setup, allowing easy bootstrapping of applications that use OWLIM
- **Wordnet:** a sample application loading the RDF/OWL representation of Wordnet is provided. It illustrates how one can query its own dataset by simple modifications of Getting Started setup
- **Quick Start Guide**

	SwiftOWLIM 2.9	SwiftOWLIM 3.X	BigOWLIM 3.X
Scale (Millions of explicit statements)	10 MSt, on 1.3GB RAM **100 MSt**, on 16GB RAM	10 MSt, on 2GB RAM **100 MSt**, on 20GB RAM	130 MSt, using 1.6GB **20 BSt**, using 64GB
Processing speed (load+infer+store)	40 KSt/s on a notebook **250 KSt/s** on a server	25 KSt/s on a notebook **200 KSt/s** on a server	10 KSt/s on a notebook **80 KSt/s** on a server
Query optimization	No	No	Yes
owl:sameAs optimization	No	No	Yes
Licence and Availability	Open-source, LGPL; SwiftTRREE is free, but not open-source	Open-source, LGPL; SwiftTRREE is free, but not open-source	Commercial. Research and evaluation copies provided for free
Comment	Fastest OWL database; multi-threaded hybrid inference	Fastest RDF engine with Named Graph and SPARQL support	Ultimate scalability and SPARQL evaluation; full-text search; clustering

KIM Platform

kim

WHAT KIM DOES:

CREATES SEMANTIC LINKS between your documents, pages and structured data

HELPS YOU USE ONTOLOGIES, describing your data and domain

Allows you **to DEFINE TEXT MINING ALGORITHMS** to identify entities and relationships

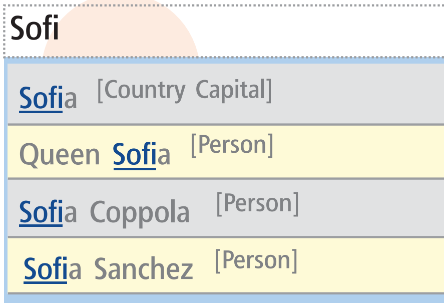
CREATES A MULTI-PARADIGM INDEX over all modalities of your data to meet your search and navigation needs

SHOWCASE:

News contains valuable (and often reliable) information about the world around us. Still, it comes from many sources and is often redundant or simply overwhelming in volume. To find your way in the news you can identify key entities and concepts mentioned in articles, and link them with background knowledge about our world. We have created a simple service that monitors newswires and creates a semantic index of their contents. You can use it to search and navigate the news in various ways, analyze emerging trends or provide press clipping services for custom fields of interest.

INCREMENTAL AUTO-SUGGEST

Look for entities through their names and get suggestions:



Sofi

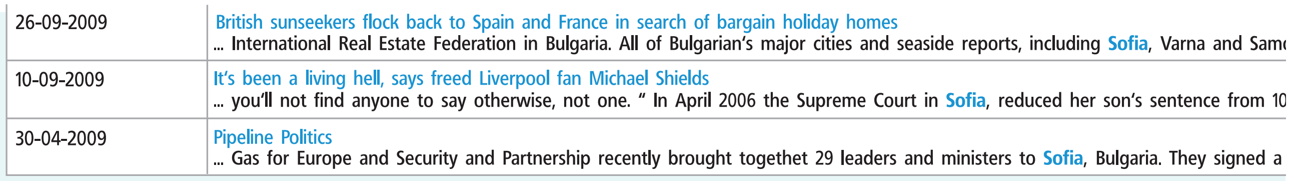
Sofia [Country Capital]

Queen Sofia [Person]

Sofia Coppola [Person]

Sofia Sanchez [Person]

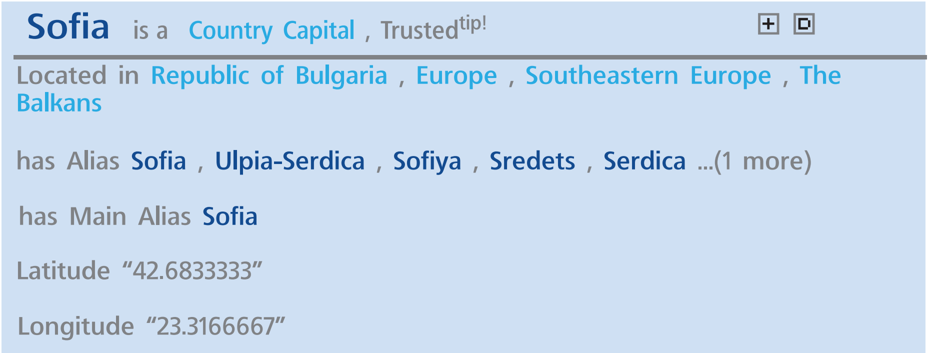
Select entities and get documents mentioning them:



26-09-2009	British sunseekers flock back to Spain and France in search of bargain holiday homes ... International Real Estate Federation in Bulgaria. All of Bulgarian's major cities and seaside reports, including Sofia, Varna and Samo
10-09-2009	It's been a living hell, says freed Liverpool fan Michael Shields ... you'll not find anyone to say otherwise, not one. " In April 2006 the Supreme Court in Sofia, reduced her son's sentence from 10
30-04-2009	Pipeline Politics ... Gas for Europe and Security and Partnership recently brought togethet 29 leaders and ministers to Sofia, Bulgaria. They signed a

Semantic Annotation and Search

Explore entity descriptions and find related entities:



Sofia is a Country Capital , Trusted tip!

Located in Republic of Bulgaria , Europe , Southeastern Europe , The Balkans

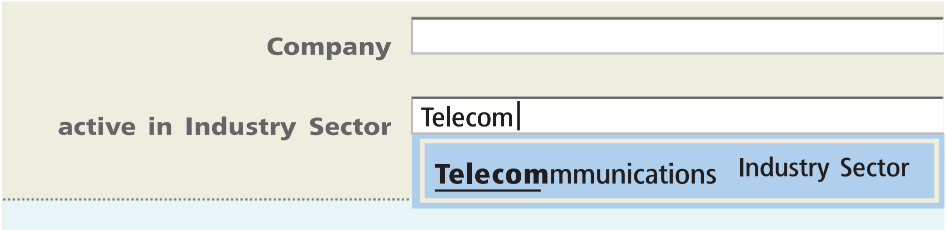
has Alias Sofia , Ulpia-Serdica , Sofiya , Sredets , Serdica ...(1 more)

has Main Alias Sofia

Latitude “42.6833333”

Longitude “23.3166667”

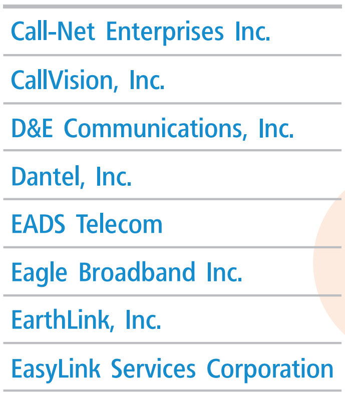
Employ relations in search, e.g. “companies from the telecommunications sector” and get a list of companies or documents about them:



Company

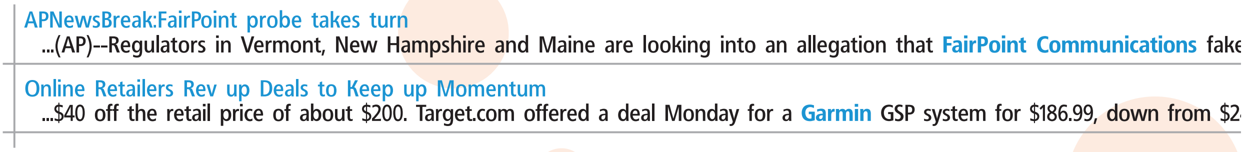
active in Industry Sector: Telecom

Telecommmunications Industry Sector



- Call-Net Enterprises Inc.
- CallVision, Inc.
- D&E Communications, Inc.
- Dantel, Inc.
- EADS Telecom
- Eagle Broadband Inc.
- EarthLink, Inc.
- EasyLink Services Corporation

EMPLOY RELATIONS IN SEARCH



APNewsBreak:FairPoint probe takes turn
...(AP)--Regulators in Vermont, New Hampshire and Maine are looking into an allegation that FairPoint Communications fake

Online Retailers Rev up Deals to Keep up Momentum
...$40 off the retail price of about $200. Target.com offered a deal Monday for a Garmin GSP system for $186.99, down from $2

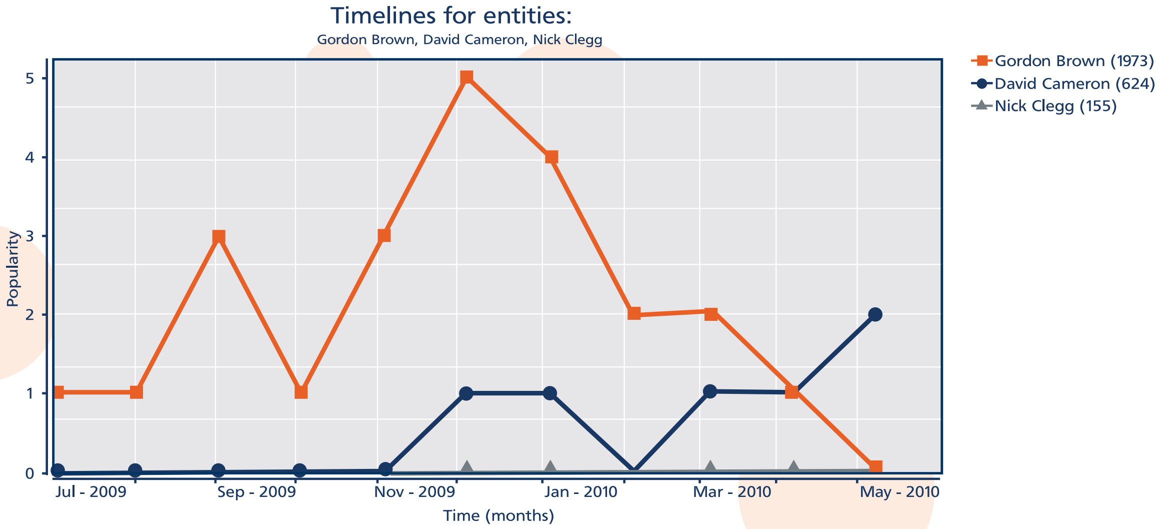
KIM Platform

Search for entities appearing together with a selected one, e.g. “people appearing in the news along with Gordon Brown”, and look into their descriptions or documents mentioning them:

- David Cameron
- Alistair Darling
- Barack Obama
- George Osborne
- Nicolas Sarkozy
- Tony Blair
- Angela Merkel
- Nick Clegg
- Vince Cable
- Hamid Karzai
- David Milibard

SEARCH FOR ENTITIES

Analyze popularity trends in a data subset, e.g. “relative popularity of Brown, Cameron and Clegg”:



WHAT WE CAN DO FOR YOU:

WE PROVIDE KIM free for evaluation and development

IT COMES WITH A PRE-DEFINED BASIC ontology and extraction of the most popular entity types (people, organizations, locations, dates ...)

WE CAN SUPPORT YOUR CUSTOMIZATION and development or fully tailor KIM to suit your needs

Semantic Annotation and Search

SOME KIM APPLICATION AREAS:

ENTERPRISE SEARCH AND SEMANTIC CONTENT MANAGEMENT: linking multiple sources of structured data and texts to provide an integrated view over your data

HEALTHCARE, PHARMACEUTICAL AND CHEMICAL INDUSTRIES: annotation, search and semantic data integration in clinical studies, patient records, scientific articles and patents

SALES-RELATED BUSINESS INTELLIGENCE: British Telecom surprised us with an internal, sales analysis system based on KIM

NEWS ENRICHMENT: currently working with BBC to power their metadata enriched coverage of the news

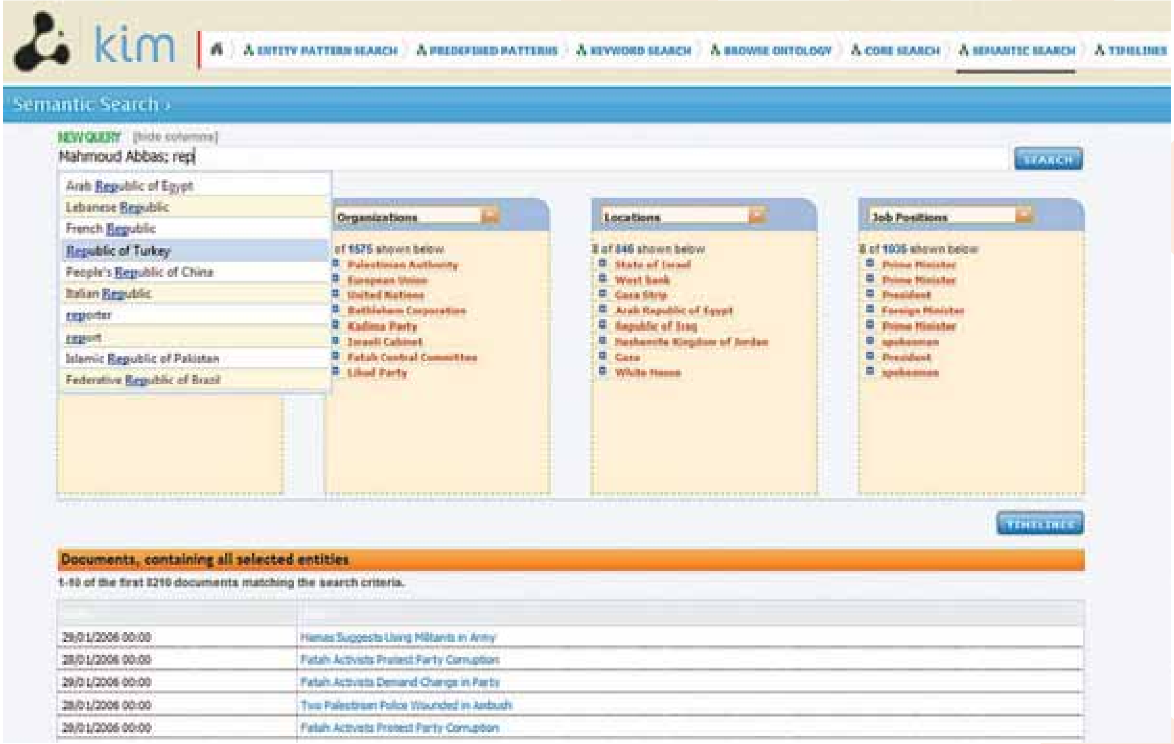
LINKED DATA BASED ANNOTATION AND SEARCH: for public and corporate documents

ENTITY PROFILING: based on profile feeds like World Check and Dow Jones and extending profiles automatically with what appears about people, organisations and their relations on the Web, to be applied for asset recovery with several governments' financial intelligence units and with over a hundred new deployments a month

ARCHITECTURE

Designed as a modular platform with well-defined APIs, KIM is based on:

- GATE as information extraction (text mining) platform
- OWLIM and Sesame as semantic repositories



Mahmoud Abbas; rep

- Arab Republic of Egypt
- Lebanese Republic
- French Republic
- Republic of Turkey
- People's Republic of China
- Italian Republic
- reporter
- report
- Islamic Republic of Pakistan
- Federative Republic of Brazil

Organizations

of 1575 shown below

- Palestinian Authority
- European Union
- United Nations
- Bethlehem Corporation
- Kadima Party
- Israeli Cabinet
- Fatah Central Committee
- Likud Party

Locations

8 of 846 shown below

- State of Israel
- West bank
- Gaza Strip
- Arab Republic of Egypt
- Republic of Iraq
- Hashemite Kingdom of Jordan
- Gaza
- White House

Job Positions

8 of 1036 shown below

- Prime Minister
- Prime Minister
- President
- Foreign Minister
- Prime Minister
- spokesman
- President
- spokesman

Documents, containing all selected entities

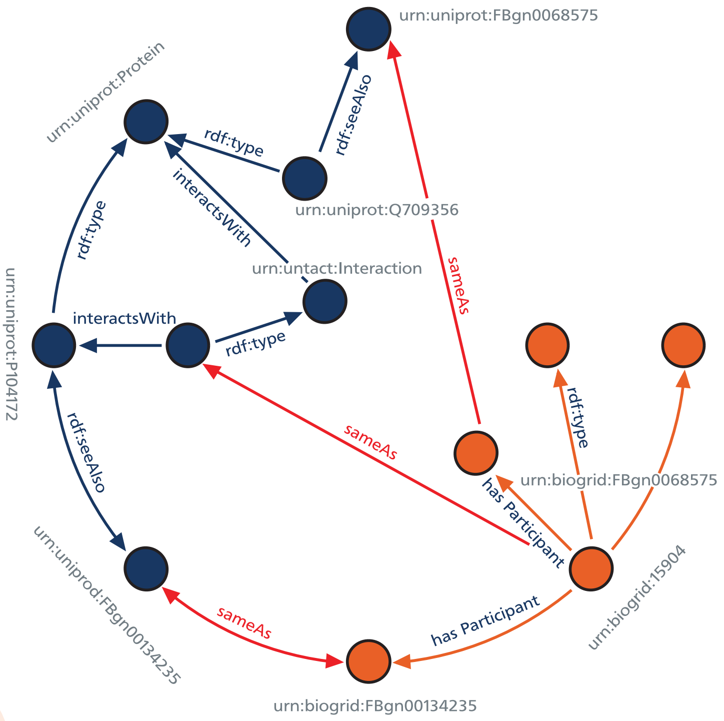
1-10 of the first 8210 documents matching the search criteria.

29/01/2006 00:00	Hamas Suggests Using Militants in Army
28/01/2006 00:00	Fatah Activists Protest Party Corruption
29/01/2006 00:00	Fatah Activists Demand Change in Party
28/01/2006 00:00	Two Palestinian Police Wounded in Ambush
29/01/2006 00:00	Fatah Activists Protest Party Corruption

Life SKIM

SERVICES AND SOLUTIONS

Ontotext offers professional services and consultancy for the development of applications and solutions in the life science and healthcare domain.

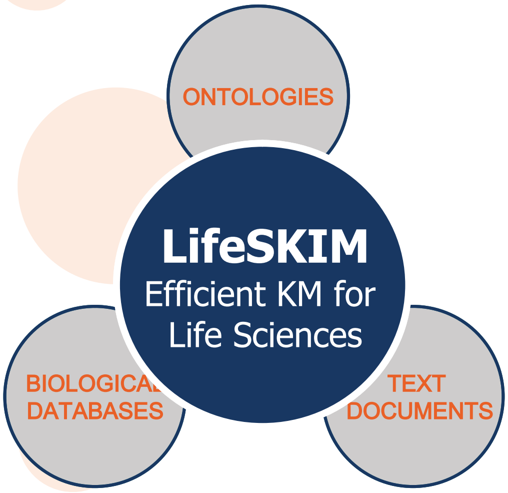


INFORMATION EXTRACTION

Ontotext develops text processing components that can process biomedical documents and extract structured knowledge and meta-data. Ontotext provides professional services such as developing named-entity recognition modules and integrating third-party text mining algorithms and applications. Based on the open-source General Architecture for Text Engineering (GATE), our solutions are vendor independent, customizable and allow extensions driven by customer needs.

SEMANTIC WAREHOUSES

Data integration is the most expensive and pressing problem faced by biotech and pharmaceutical industries. Hundreds of narrow, domain-specific databases are available, but there is no single integrated view to this heterogeneous knowledge. Ontotext implements customer-centric RDF warehouse solutions and automates the data feeding process by developing extract, transform and load scripts. We capture the semantics of your structured and unstructured information and provide you with ontologies, thesauri or customized reasoning that understands both the data and your problems.


ONTOLOGIES

LifeSKIM
Efficient KM for
Life Sciences

BIOLOGICAL
DATABASES

TEXT
DOCUMENTS

PRODUCTS

LINKED LIFE DATA (p.16)

LIFESKIM

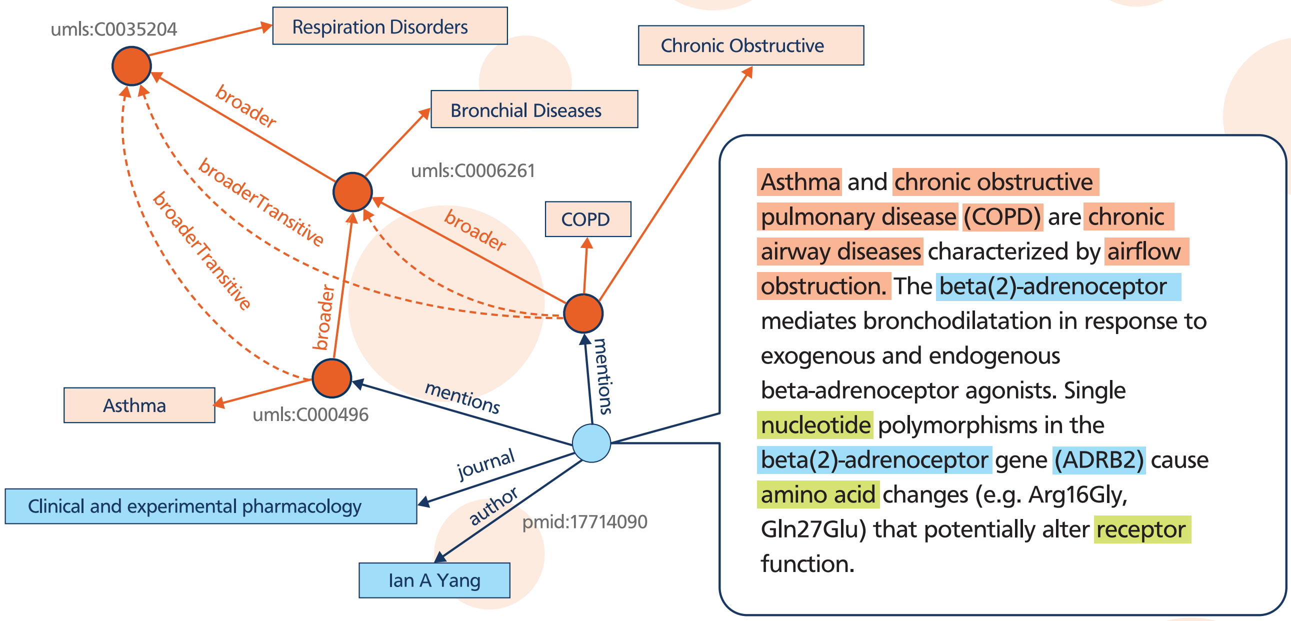
LifeSKIM is an end-user application that offers a wide range of functionality – from standard information retrieval to advanced semantic search. By combining semi-structured data and information extraction technologies, LifeSKIM delivers unmatched query precision in the life science and healthcare vertical sector.

The application automates document meta-data management, document mining, scalable named entity recognition and has the following main features:

- Semantic indexing and retrieval of documents using knowledge bases
- Flexible querying and navigation of knowledge that has been generated from both structured and unstructured text

Search Types	Example Queries
Semantic Search	Documents that contain a human gene and a respiratory disease
Text Search and Analysis	PROC" (gene) AND "lung cancer"
Simple Text Search	"PROC gene" AND "lung cancer"

- Automatic ontology learning and population from text
- Efficient reasoning against extracted and structured information
- Co-occurrence and ranking of entities in documents
- Classical information retrieval queries
- Persistence and annotation of arbitrary data types



Asthma and chronic obstructive pulmonary disease (COPD) are chronic airway diseases characterized by airflow obstruction. The beta(2)-adrenoceptor mediates bronchodilatation in response to exogenous and endogenous beta-adrenoceptor agonists. Single nucleotide polymorphisms in the beta(2)-adrenoceptor gene (ADRB2) cause amino acid changes (e.g. Arg16Gly, Gln27Glu) that potentially alter receptor function.

Linked Life Data

Linked Life Data (LLD) is a scalable RDF warehouse solution that supports inference using the OWLIM engine (p.6). By supporting heterogeneous data sources, simple information updates, easy incremental extensions and integration of datasets, it helps you to develop a large scale KB and efficiently answer queries. Company information can be mined and reused through standard interfaces!

Many organizations have started to publish their data in a reusable way. Thus, they mitigate the problems of traditional search engines, which fail to directly address the specific needs of users. LLD makes this integrated information accessible, complies with existing standards and adds semantic multidimensional search as a powerful tool for exploring and querying knowledge bases.

Linked Life Data helps to solve the semantic data integration challenge by:

ENABLING THE SEMANTIC INTEGRATION of multiple disconnected sources

SIGNIFICANTLY REDUCING THE COST of combining internal company knowledge with the information available on the Web.

Linked Life Data

linked life data

Semantic data integration platform for biomedical domain

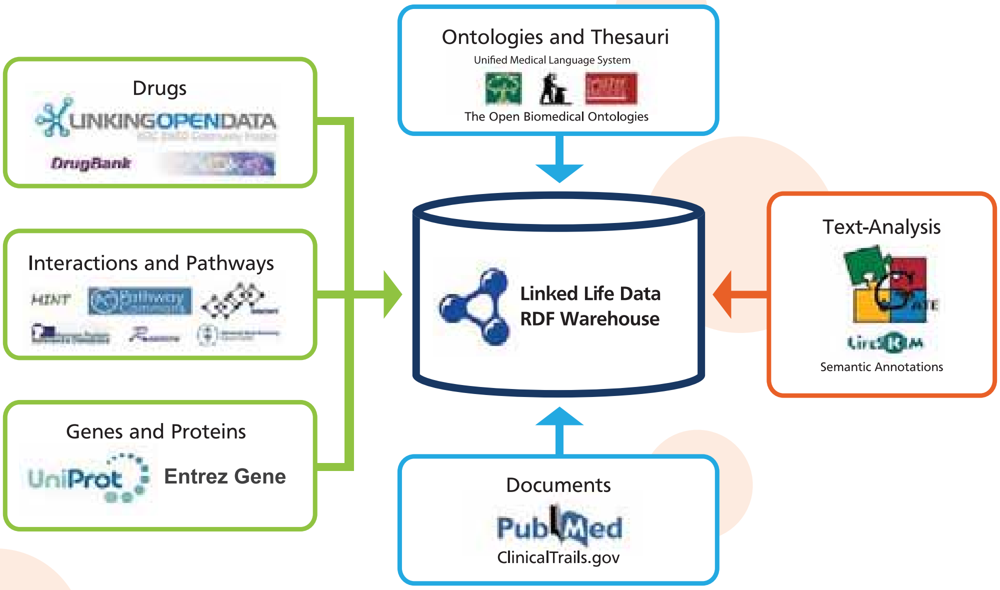
chronic obj

DEMONSTRATING EXCELLENT SCALABILITY:

- Offering 20+ popular biomedical data sources
- Supporting standard data integration patterns in the internal system reasoner
- Linking document text to your company entities with state-of-the-art information extraction algorithms that are compliant with community standards

LLD offers superior search performance over highly heterogeneous datasets. The service is publicly available at: http://linkedlifedata.com

The LLD datasets are indicated in yellow on the LOD map on the right. It contains over 4 bill. explicit statements



FactForge - Manageable Linked Data

REASON-ABLE VIEW TO LOD

FactForge represents a reason-able view to the Web of data. It is meant to serve as an index that allows users to find resources and facts based on the semantics of the data. http://www.factforge.net

In FactForge we selected several of the central datasets of the LOD project and loaded them in the OWLIM (p.6) semantic repository. Reasoning was performed to "materialize" the facts that could be inferred from these data.

DATASETS, ONTOLOGIES AND STATISTICS

DATASETS INCLUDED: DBPedia, Geonames, UMBEL, Wordnet, CIA World Factbook, Lingvoj, MusicBrainz. These are indicated in red on the map of LOD below.

ONTOLOGIES: several schemata referred by the datasets were loaded in FactForge: DC, SKOS, FOAF, RSS.

SIZE: 1.2B explicit and 881M inferred statements were indexed; the total number of retrievable triples is 10 billion.

INFERENCE

Materialization is performed with respect to semantics very similar to **OWL 2 RL**.

FactForge benefits from the optimized handling of **owl:sameAs** in BigOWLIM, which allows considerable reduction of the indices, without loss of semantics or performance.

SAMPLE QUERIES

AN EXTENSIVE SET OF SAMPLE QUERIES AVAILABLE AT FACTFORGE AIMS TO:

- **Guarantee data consistency**, in the same way in which unit tests guarantee software quality
- **Lower the cost of entry**, demonstrating how data from multiple datasets can be joined in a useful manner

One of them answers the question "In which cities can one see Modigliani paintings?". FactForge (previously known as LDSR) was the first system to pass The Modigliani Test, defined as a criterion for the tipping point of the Semantic Web at ReadWriteWeb,

http://www.readwriteweb.com/archives/the_modigliani_test_for_linked_data.php

LINKING OPEN DATA (LOD)

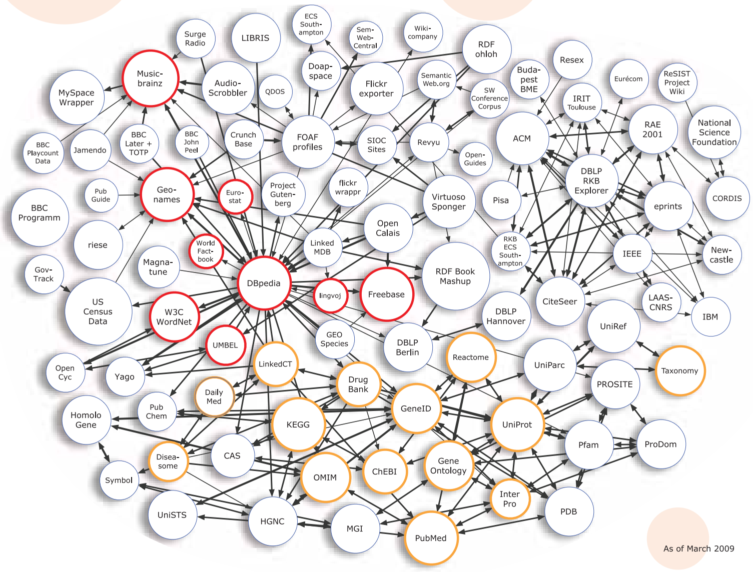
Linking Open Data is a W3C SWEO Community Project that aims to facilitate the emergence of a Web of linked data by means of publishing and interlinking open data.
http://esw.w3.org/topic/SweoIG/TaskForces/...
CommunityProjects/LinkingOpenData

EXPLORATION AND QUERYING

FACTFORGE IS A FREE PUBLIC SERVICE DELIVERING FAST AND RELIABLE SINGLE POINT OF ACCESS TO THE CENTRAL LOD DATASETS

The data can be accessed in several ways:

- Incremental **URI auto-suggest**
- **RDF Search,** returning a ranked list of RDF snippets
- **Exploration** - traversing the data, one resource at a time
- **Evaluation of queries** in SPARQL
- **SPARQL end-point**



As of March 2009

Web Mining Framework

Ontotext's Web Mining Framework (WMF) integrates screen scraping and focused extraction and efficiently handles a wide range of web intelligence and web search applications.

It provides infrastructure and implements components for:

- Focused web crawling: collecting only web pages with specific information
- Screen scraping: acquiring large volumes of data from the "deep web" with high precision (e.g. job boards)
- Information extraction: acquiring structured data from plain text
- Identity resolution: merging data from different sources
- Semantic annotation and search: using KIM and OWLIM

The framework provides a technological basis for a couple of Ontotext's joint ventures (Innovantage and Namerimi) as well as for a number of other applications.

KEY ADVANTAGES

Unlike some other open source and commercial solutions our framework:

- Is optimized for big volumes of data and provides web mining for search engines
- Applies shallow NLP techniques based on GATE's proven and comprehensive technology
- Balances genre/domain specificity, task complexity and usable accuracy
- Covers the whole life cycle of building, executing, monitoring and maintenance of web mining components
- Allows real time continuous data collection and ensures 24/7 run
- Provides basis for accurate data analysis and statistics

INDUSTRIAL APPLICATIONS

We have deployed our solution in several large scale industry applications, e.g. recruitment intelligence and vehicle trading.

INNOVANTAGE, http://innovantage.co.uk

Provider of online recruitment intelligence in UK:

- Maintains an up-to-date database of about 500 000 vacancies and about 2 million company profiles
- Harvests vacancies from 300 000 websites and more than 20 job boards (see the diagram on the right)

RONSMAP, http://ronsmap.com

One-stop online shop for comparing vehicle offers:

- Targets the gathering of all vehicle offers from the US
- Monitors US car dealers' activities and inventory dynamics
- Stores all details from the offers allowing for market analysis by regions, vehicle types and many other features

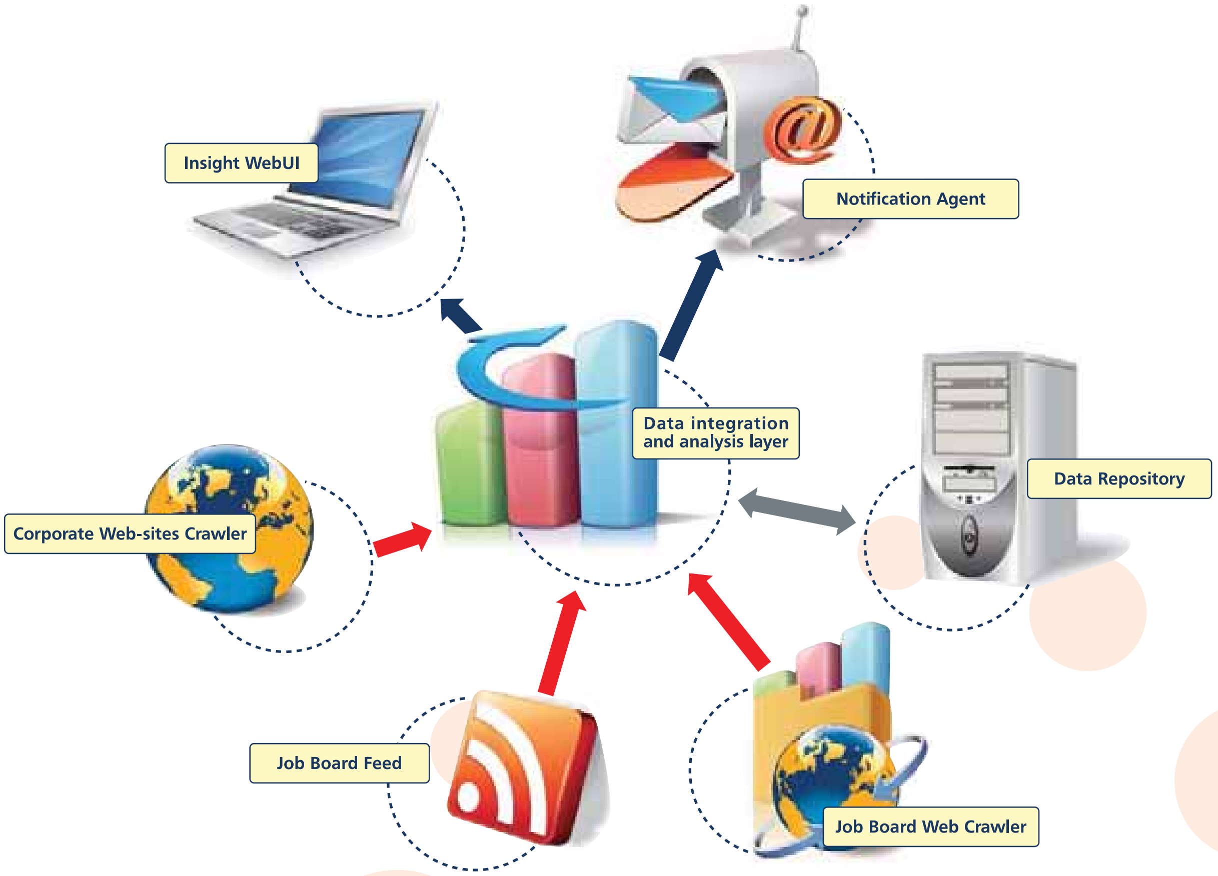
NAMERIMI, http://eto.bg

Namerimi is a search engine offering Bulgarian web content, estimated to be about 100 million pages. The engine provides highly-interactive user interface featuring user tagging and feedback collection. Its back-end implements semantic suggestions and ranking, and entity-based search.

Namerimi is based on the following technologies:

- NUTCH – distributed crawling and indexing
- The most comprehensive Bulgarian morphology engine
- Named entity recognition and semantic annotation
- IR (indexing) tuned for national/Bulgarian language

The Dataflow at Innovantage



ontotext

RESEARCH PROJECTS

ONTOTEXT PARTICIPATES IN THE FOLLOWING PROJECTS:

SOA4ALL - a platform based on Web 2.0, Semantics and Context Management to enable a Web of bill. of services

LARKC - a platform for large-scale integrated reasoning and Web-search

NoTube - semantic technologies for personalised creation, distribution and consumption of TV content

INSEMTIVES - incentive models and framework for semantic metadata and content authoring

MOLTO - developing tools for translating texts between multiple languages in real time with high quality

KHRESMOI - develops a multi-lingual, multi-modal search and access system for biomedical information

CUBIST - combines features of semantic techniques with those of standard business intelligence

RENDER - tackles the enormous diversity of sources, culture and information on the Web

JOINT VENTURES

Innovantage: provider of recruitment intelligence data and technology in the United Kingdom

Harvesting vacancies from 300 000 websites
Integrating about 500 000 vacancies from job boards

Namerimi: developer of focused Web mining and search technology based on semantics

Used in a Bulgarian national search engine (~100M pages)
Joint venture with NetInfo

MEMBERSHIPS

W3C®

DSLL

STI · INTERNATIONAL

Ontotext AD
a Sirma Group Company
135 Tzarigradsko Chaussee, Floor 3
1784 Sofia, Bulgaria

Tel: (+359 2) 974 61 60 info@ontotext.com
Fax: (+359 2) 975 32 26 www.ontotext.com