ontotext

Semantic Technology on the Field

June 2012 Brochure

OpenPolicy Improves Search for Government Regulations p.4

Live at BBC Sports and Olympics 2012 Sites p.4

Pharma Company Outsources Linked Data Management p.5

Museums Publish Data with OWLIM p.5

App to Table: Semantic Technology Helping People Eat Better!

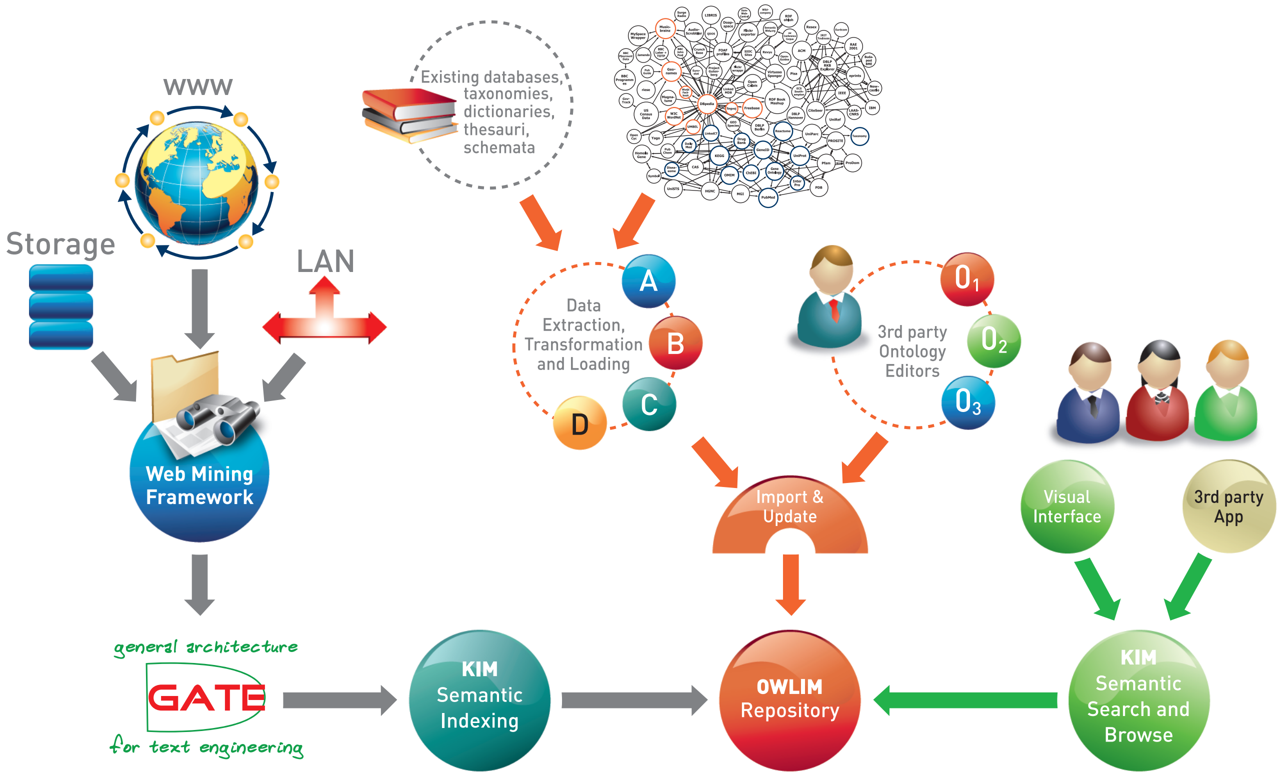
Mobile Recipes Served Up by OWLIM p.4

Linked Data for the Enterprise

Every organization struggles to get the right information to the right people when they need it – whether these people are internal analysts, executive decision-makers, or customers. But often the right data is in multiple formats, various locations and being accessed by different systems.

Enterprises need a better approach to find information where it lives, link it based on meaning, search it efficiently, and manage changes more easily.

This approach is *Linked Data for the Enterprise*, powered by semantics, built by Ontotext.



The Ontotext Value Proposition

Ontotext's unique interdisciplinary expertise provides a single access point into all technologies that turn the promise of semantics and linked data into reality.

Our technology unveils implicit relations and finds hidden links, matching facts scattered across huge volumes of diverse information. It delivers more answers in less time and with less effort.

Ontotext focuses on technology leadership that is sustainable, practical, and readily available now.

THE BUSINESS CASE: A Shorter Adoption Curve for Long-Term Benefits

- Flexible data schemas adapt to changing information more easily
- Search across multiple sources without increasing complexity
- Text mining saves manual indexing & data entry time, without loss of depth and quality
- Our solutions can be applied in cooperation with existing data, storage, search, and publishing systems

OUR PROMISE: Sustainable Value

- We deliver semantic technologies at enterprise scale and reliability
- Ontotext has expertise in the full range of semantic technologies
- The up-front investment is lower than you might expect
- Open standards protect your investment
- We are committed to training, knowledge transfer, and support

Use cases

Mobile Recipes Served Up by OWLIM

Edamam, New York, NY

Edamam has assembled a unique food database that integrates data from public data sources, like calorie and nutrition information, with recipes from leading food websites. The combination is a powerful tool for consumers to make intelligent food decisions - all served up by the OWLIM triple store (p. 6).

Selected by VentureBeat and IDG as an innovative technology company, Edamam launched its first consumer product – a universal recipe search – at DEMO2012.

"Transforming the organic and implied knowledge about food into structured data was a huge challenge. Ontotext was instrumental in solving this problem" "Edamam is about eating better. We harness technology to organize food knowledge and give it back to the people so they can make smarter choices about food. We aim to bring the joy of food and cooking back into people's lives."

Victor Penev, Edamam CEO

Open Policy Improves Search for Government Regulations

LMI, Washington, DC

LMI is a not-for-profit management consulting firm serving U.S. government departments and agencies. LMI has partnered with Ontotext to combine Ontotext's semantic annotation and search capabilities (p.7) with OWLIM (p.6) to create an advanced document search system called Open Policy.

OpenPolicy offers end-users faceted searching capabilities alongside semantic keywords and full text search in an easy-to-use browser interface. The combined search capabilities vastly improve access to key regulatory information for government users at all levels - from field operations to policy specialists.

Live at BBC Sports and Olympics 2012 Sites

Following the successful use of **Dynamic Semantic Publishing** (DSP) in the **BBC FIFA World Cup 2010 website**, BBC is extending the use of Ontotext technology for their redesigned Sports website and for the highly anticipated London Olympics 2012 website. "*DSP uses linked data technology to automate the aggregation, publishing and re-purposing of interrelated content objects according to an ontological domain-modelled information architecture, providing a greatly improved user experience and high levels of user engagement*"*

The use of semantic technology in the BBC's Web publishing process improves cost efficiency and product features: "*Replacing a static publishing mechanism with a dynamic request-by-request solution that uses a scalable metadata/data layer will remove the barriers to creativity for BBC journalists, designers and product managers, allowing them to make the very best use of the BBC's content.*"*

Ontotext's OWLIM semantic repository (p.6) and the Concept Extraction Service (p.7) developed by Ontotext using the GATE architecture, have an important role for the success of the DSP.

* from blog post on "Sports Refresh: Dynamic Semantic Publishing", http://www.bbc.co.uk/blogs/bbcinternet/2012/04/sports_dynamic_semantic.html

Jem Rayfield,
Lead Technical Architect
at the BBC's Future Media
department

Pharma Company Outsources Linked Data Management

Linked Life Data: Increasing Research Scope without Integration Headaches

UCB, Brussels, Belgium

UCB (www.ucb.com) is a global biopharmaceutical company focused on the discovery and development of innovative medicines and solutions to transform the lives of people living with severe diseases of the immune system or of the central nervous system. With more than 8,000 people in about 40 countries, the company generated revenue of EUR 3.2 billion in 2011. UCB selected Ontotext and the Linked Life Data (LLD) service to provide support for the linked data cloud and leverage its capabilities to help support research questions faced in the drug discovery process.

UCB is using Linked Life Data to integrate dozens of public biological, chemical and medical data sources into its internal research stream for drug discovery. The knowledge base is a valuable source of research data that may be used to generate and validate complex hypothesis and does this while offloading the data integration cost from the UCB researchers.

"Uncovering the answers to complex questions is increasingly challenging especially as the key elements are often broadly distributed across multiple data sources and consist of independent but related concepts. LLD plays a critical role in the compilation of public data and then brings these data sources together using a Linked Data approach which in turn facilitates efficient integration with our internal data. Ontotext's vision for the future of this product is very much in line with our own goals and we value this partnership."

Phil Scordis, Director, Informatics, UCB

Museums Publish Data with OWLIM

The British Museum, London, UK; Yale Center for British Art, New Haven, CT, USA

The British Museum (BM) and Yale Center for British Art (YCBA) are some of the leading cultural heritage (CH) institutions that want to publish their collections as Linked Open Data (LOD). While both institutions have online search (BM Collection Search, YCBA Collection Search), the search applications are totally different and not easy to integrate.

Publishing as LOD allows the data to be reused in various scenarios, consumed by various applications and used by developers for creative purposes. Both institutions use the **CIDOC CRM ontology** that allows universal representation of cultiural heritage (CH) artifacts, including museum collections, painting galleries, archaeological finds, etc. Ontotext is helping CH institutions with harmonization of CRM mappings, in order the make data representation more uniform and more useful.

Both BM and YCBA have selected OWLIM (p.6) for their semantic repository. OWLIM was selected because of its high performance for both data loading and querying, strict standards compliance, OWL reasoning and rule inferencing that are useful for CRM, especially regarding complex search. An added benefit is that OWLIM supports the new W3C standard SPARQL 1.1 Federated Query that allows a user to interrogate several semantic repositories at once. Federation can help research communities and projects to collaborate on a certain topic, even when the relevant works and data are scattered in different collections. Federation supports both Union scenarios (e.g. finding all works by a certain author no matter in which collection they reside) and Join scenarios (e.g. assembling data about a work or author across several repositories).

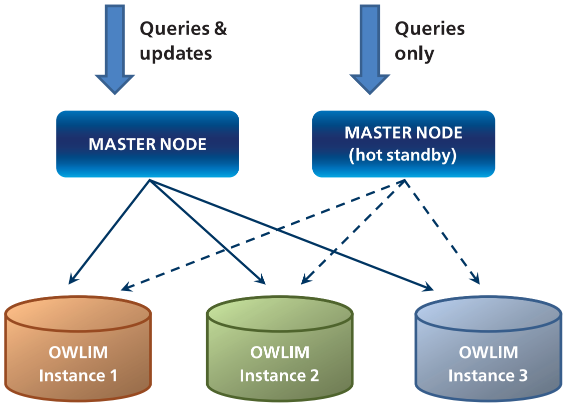
Enterprise-Ready Data Management

What: **RDF Database with Reasoning**

Semantic repositories are database management systems (DBMS) that use RDF data representation and can infer non-explicit information.

OWLIM is a family of semantic repositories with the following key characteristics:

- **Native RDF engines**, implemented in **Java**
- Fully performant through both **Sesame** and **Jena**
- Inference support for RDFS, **OWL 2 RL** and QL
- Cluster support, providing high resilience, automatic fail-over and **linearly scalable query performance**

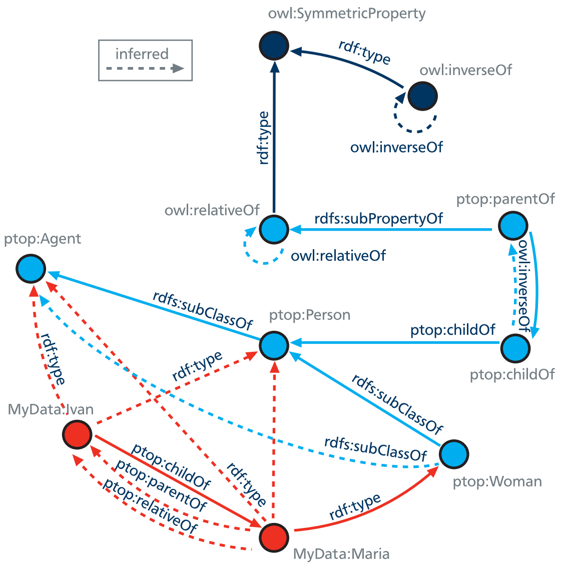


Why: **Easier Data Integration and Querying**

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

Semantic repositories allow one to:

- Query data without knowing the vocabulary used to assert it, given even a course-grained schema mapping
- Retrieve relationships of unknown types; a query pattern like "John ?x Mary" is impractical in relational DBMS
- Flexibly modify the schema(s) during normal operation



Why Us: **Proven Usability, Robustness and Performance**

OWLIM is the only RDF engine that provides **transparent reasoning** support throughout the entire life cycle of the data (loading, updates, querying). It also provides hybrid querying capabilities that combine SPARQL with efficient full-text search, **geo-spatial** constraints and **ranking** of query results.

OWLIM has been selected after thorough evaluation for a range high-profile projects (see pages 4 and 5), including:

- **BBC's World Cup** project, now extended to BBC Sports and 2012 Olympics
- **UniProt**'s SPARQL end-point

OWLIM provides **optimized owl:sameAs** support that delivers dramatic improvements in performance and usability when huge volumes of data from multiple sources are integrated (see FactForge, p.8).

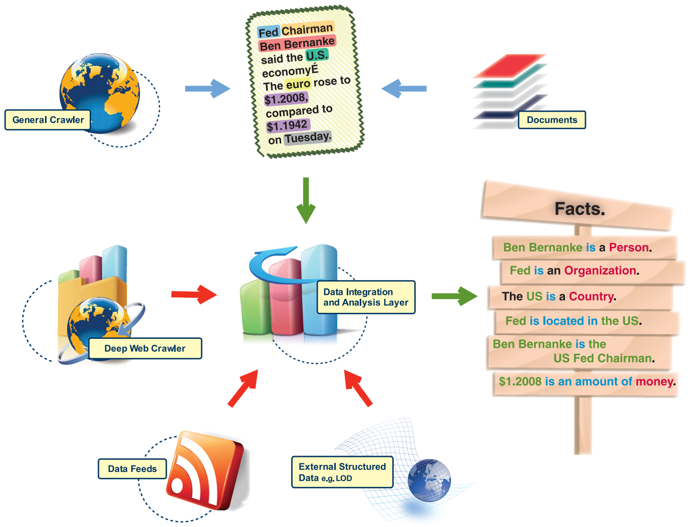
OWLIM's superior performance was justified by all recent independent benchmarking efforts http://www.ontotext.com/owlim/references. OWLIM version 5.0 is **43% faster** than its predecessor in scenarios involving frequent updates in the context of heavy query loads and uses up to **70% less storage** space.

Content Management and Web Mining

What: **Interlink, Search and Navigate Your Information**

KIM is the original commercial platform for **Semantic Annotation** that offers:

- Ontology-based **information extraction** from unstructured documents
- Linking named entities in text to structured data and **metadata generation**
- Customizable data representation and search interfaces


General Crawler
Fed Chairman Ben Bernanke said the U.S. economyÉ The euro rose to $1.2008, compared to $1.1942 on Tuesday.
Documents
Deep Web Crawler
Data Integration and Analysis Layer
Facts.
Ben Bernanke is a Person.
Fed is an Organization.
The US is a Country.
Fed is located in the US.
Ben Bernanke is the US Fed Chairman.
$1.2008 is an amount of money.
Data Feeds
External Structured Data e.g. LOD

What: **Get the Data from the Web**

Ontotext's Web Mining Framework (WMF) has all the tools that one needs to acquire the data, wherever it comes from or appears:

- **Focused crawling** of specific sections or selected information from web pages
- **Screen scraping** of structured online data with high precision (e.g. job boards)
- Data extraction, transformation, merging and **de-duplication**

It is a platform for **full-lifecycle web mining applications** that is optimized for autonomous, continuous 24/7 collection of large volumes of data.

Why: **More Efficient Information Management**

Efficient acquisition, normalisation and interlinking of enterprise information requires a wholistic semantic approach, where all the components work with a single conceptual model (ontology) and access a shared semantic repository to retrieve relevant information and to deliver results.

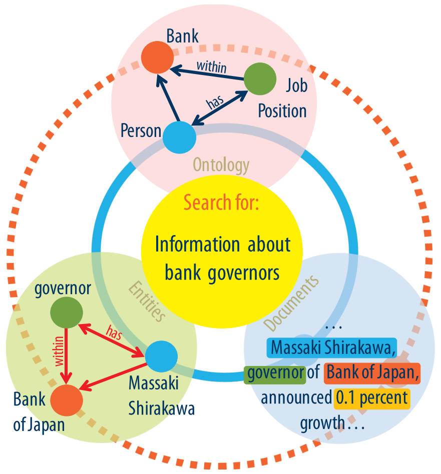
Why Us: **Proven 360° Semantic Technology from Single Vendor**

We can offer top-notch expertise and technology all the way from data integration through text mining to semantic search. All that integrated with the best RDF database on the market (p.6) and proven in multiple real high-profile projects (pp.4-5).

KIM's text analysis modules and algorithms are at the heart of many of Ontotext's commercial projects, such as:

- Concept Extraction for the BBC's Dynamic Semantic Publishing system (p.4)
- Entity Recognition and Integration for Open Building's online architectural database
- Faceted Search for LMI's Open Policy document search system (p.4)

Using WMF we can **build within a year a market intelligence database** for a G7 country. We already did it for jobs in the UK, cars in the USA, hotel rates in Europe. Or to collect and integrate 5 types of data about food from 20+ sources in EDAMAM (p.4).



Linked Data Management

Reason-able Views to LOD

Linking Open Data (LOD) is a W3C Project, which 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

The main **challenges** for the adoption of the linked data are that:

- LOD is **hard to comprehand**, making structured queries against 200 different schemata is tough
- LOD is generally **not reliable**, no consistency guarantees
- Querying data distributed on the Web is **slow**, because the "remote joins" are slow

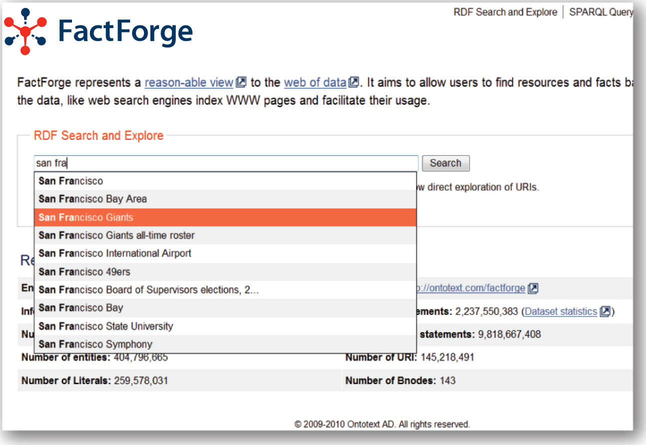
The reason-able views concept:

- Makes it easier and less risky to use part of the LOD data for specific purposes
- Selects, cleans up and integrates selected datasets and ontologies in a compound dataset
- Loads the compound dataset in a single semantic repository

Sample Queries:

An extensive set of sample queries provided with the reasonable-view aims to:

- **Guarantee data consistency** in the same way in which unit tests guarantee software quality
- **Lower the cost of entry**, demonstrating useful patterns of joining data from multiple datasets



RDF Search and Explore | SPARQL Query

FactForge

FactForge represents a reason-able view to the web of data. It aims to allow users to find resources and facts b
the data, like web search engines index WWW pages and facilitate their usage.

RDF Search and Explore

san fra

Search

San Francisco
San Francisco Bay Area
San Francisco Giants
San Francisco Giants all-time roster
San Francisco International Airport
San Francisco 49ers
San Francisco Board of Supervisors elections, 2...
San Francisco Bay
San Francisco State University
San Francisco Symphony

w direct exploration of URIs.

Re

En ://ontotext.com/factforge

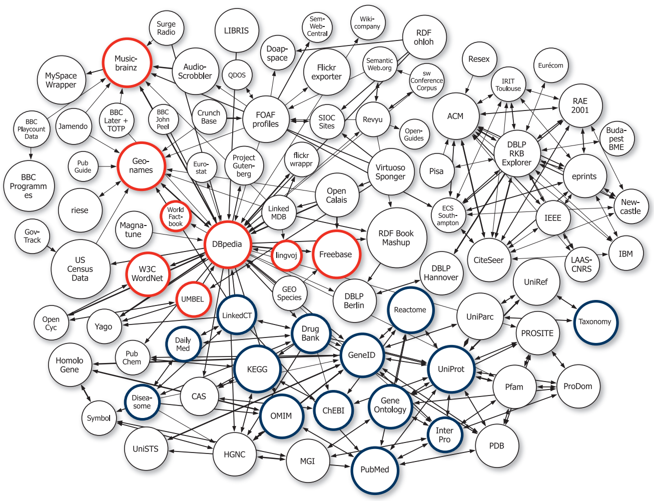
Inf ements: 2,237,550,383 (Dataset statistics)

Nu statements: 9,818,667,408

Number of entities: 404,796,665 **Number of URI:** 145,218,491

Number of Literals: 259,578,031 **Number of Bnodes:** 143

© 2009-2010 Ontotext AD. All rights reserved.



The FactForge Data Service

The Fast Track To The Centre of the Data Web
http://factforge.net

FactForge is a reason-able view including several of the central datasets of the LOD project. The schemata are mapped to the PROTON upper-level ontology (http://www.ontotext.com/proton-ontology). OWL 2 RL reasoning is performed to "materialize" the facts that could be inferred from these data.

THE DATA:

- **Datasets included**: DBPedia, Geonames, UMBEL, Wordnet, Freebase, CIA World Factbook, Lingvoj, MusicBrainz (marked in red on the LOD map above)
- **Ontologies:** PROTON, DC, SKOS, FOAF, RSS and the dataset's proprietary ontologies
- **Size**: 1.7B explicit and 1.4B inferred statements were indexed; the total number of retrievable statements is **14 billion**

THE ACCESS:

- **Go-to-resource**: incremental URI auto-suggest
- **Keyword search**: RDF Search, returning a ranked list of RDF snippets
- **Exploration**: traversing the data, one resource at a time
- **Structured queries**: evaluation of queries in SPARQL
- **Remote server access**: SPARQL end-point

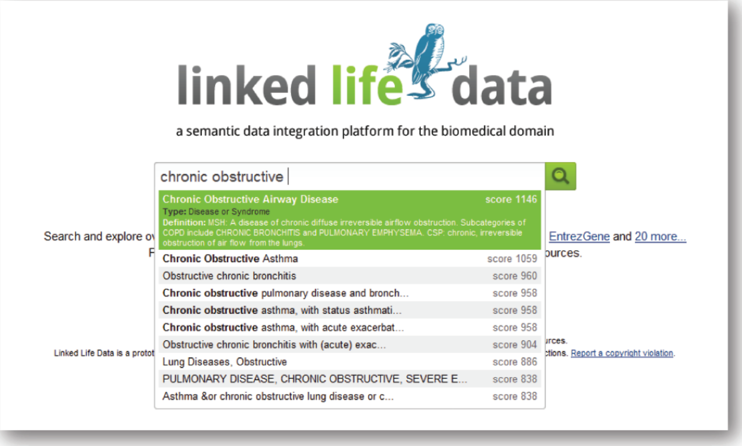
Semantics for Life Sciences

Linked Life Data - the Platform and the Public Service

The Linked Life Data (LLD) platform from Ontotext offers a novel service that allows companies to **easily integrate dozens of public biomedical data sources** into internal research and data analysis processes. Using LLD enterprises can fully outsource the hosting, the integrating and the maintenance of very big RDF repositories, as presented in the UCB use case (p.5).

The Linked Life Data service is a public showcase of the LLD semantic warehousing platform. The service semantically integrates and aligns more than **26 diverse data sources** by identifying similar concepts. The warehouse aggregates more than 1 billion biomedical objects described by more than 5 billion RDF statements generated using reasoning and text analysis methods. In LLD users can exploit Interactive Relationship Discovery in billions of biomedical entities.

Each month LLD serves about 100 thousand SPARQL queries and more than 1.2 million linked data resource requests. Give it a try at http://www.linkedlifedata.com



a semantic data integration platform for the biomedical domain

chronic obstructive

Chronic Obstructive Airway Disease score 1146
Type: Disease or Syndrome
Definition: MSH: A disease of chronic diffuse irreversible airflow obstruction. Subcategories of COPD include CHRONIC BRONCHITIS and PULMONARY EMPHYSEMA. CSP: chronic, irreversible obstruction of air flow from the lungs.

Chronic Obstructive Asthma	score 1059
Obstructive chronic bronchitis	score 960
Chronic obstructive pulmonary disease and bronch...	score 958
Chronic obstructive asthma, with status asthmati...	score 958
Chronic obstructive asthma, with acute exacerbat...	score 958
Obstructive chronic bronchitis with (acute) exac...	score 904
Lung Diseases, Obstructive	score 886
PULMONARY DISEASE, CHRONIC OBSTRUCTIVE, SEVERE E...	score 838
Asthma &or chronic obstructive lung disease or c...	score 838

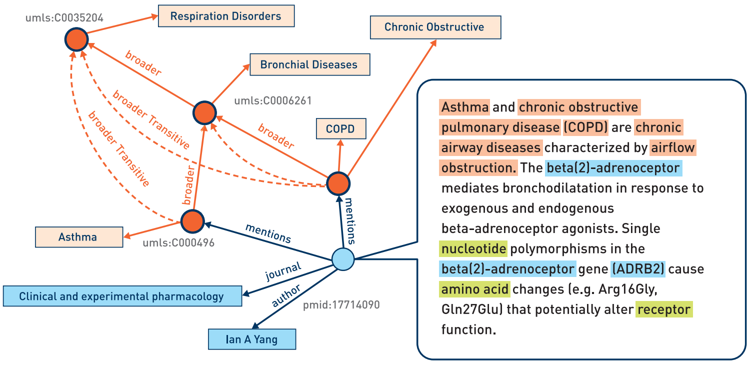
Search and explore ov EntrezGene and 20 more...
F ources.

Linked Life Data is a protot rces.
ctions. Report a copyright violation.

Semantic Biomedical Tagger

The Semantic Biomedical Tagger recognizes 135 biomedical entity types and semantically links them to LLD. This process is also known as semantic annotation and helps overcome natural language related ambiguities when expressing notions and their computational representation in a formal language. The key benefits of implementing semantic annotations are:

- Enriching the unstructured information with a context that is further linked to the domain structured knowledge
- Allowing results that are not explicitly related to the original search
- Processing complex filter and search operations



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.

Ontotext's Life Sciences team provides **professional services** for data modelling, database optimization, and complex information extraction that requires specific biological expertise. Find more at http://www.ontotext.com/life-sciences

Training

Ontotext training provides knowledge and expertise to current and prospective users of semantic technology and Ontotext products. We have experience in **tailor-made courses** designed to meet the requirements of specific client organizations and teams - from short executive trainings to extensive courses for technicians responsible for the operations and the administration of critical environments.

In addition to the bespoke training, at present we offer a single regular, open training course entitled "Semantic Technology with OWLIM". Dates for this course are scheduled regularly in London.

Semantic Technology with OWLIM

This course begins with the fundamentals of various semantic technologies and leads up to providing the practical knowledge required to build and support semantically enabled applications based on OWLIM. It is intended both for those new to semantic technologies and those who already have some experience.

The full duration of the course is three days, but attendants can book only those days they feel are necessary, depending upon their experience and interest. People with a broad experience in semantic technologies and standards can skip the first day. The third day is focused mostly on the operational aspects of setting up and maintaining a robust OWLIM deployment.

The maximum number of attendantees is 15, in order to facilitate a hands-on experience and personal attention. The agenda of the course is presented below.

DAY 1: Introduction to Semantic Technologies

- RDF, RDF Schema Vocabulary (RDFS) and OWL
- Automated Reasoning and Ontologies
- SPARQL
- Representing your data in RDF, Linked Data

DAY 2: OWLIM for users and application developers

- Overview of OWLIM, installation with Tomcat
- Getting started application and Sesame utilities
- Full-text search, geo-spatial extensions and RDF Rank
- Notifications, query modifiers and RDF Priming

DAY 3: OWLIM for administrators and operations staff

- Rules, indexing and operating behaviour
- Jena, owl:sameAs optimisation, performance tuning
- Administration tasks, security
- Replication cluster

One can find further information and registar for upcomming courses at
http://www.ontotext.com/training

Partners

TECHNOLOGY PARTNERS

bpeng
BUSINESS PROCESS ENGINEERING

Rovereto, Italy

TopQuadrant™

Mountain View, CA, USA

general architecture
GATE
for text engineering

Sheffield, UK

STRUCTURED DYNAMICS

Coralville, IA, USA

innovantage_

Cardiff, UK

fluid Operations

Walldorf, Germany

REGIONAL PARTNERS

ONTOBA™

United Kingdom

semantrix
Intelligent Information Access

Australia & New Zealand

Saltlux
Communicating Knowledge

South Korea

MONDECA

Paris, France

RESEARCH PROJECTS

Ontotext participates in the following projects funded under the Seventh Framework Program (FP7) of the European Commission:

AnnoMarket - to deliver an innovative, accessible, and open marketplace for pay-as-you-go, cloud-based text analysis resources
http://www.annomarket.com/

CUBIST - combines essential features of Semantic Technologies and Business Intelligence
http://www.cubist-project.eu/

EUCLID - providing a comprehensive educational curriculum, supported by multi-modal learning materials and highly visible eLearning distribution channels, tailored to the real needs of data practitioners
http://www.euclid-project.eu/

Khresmoi - to build a multi-lingual multi-modal search and access system for biomedical information and documents
http://www.khresmoi.eu/

MOLTO - cross-lingual search and ontology-language interoperability. Asking questions in a natural way.
http://www2.molto-project.eu/

RENDER - developing methods, techniques, software and datasets that will leverage diversity as a crucial source of innovation and creativity
http://www.render-project.eu/

TRENDMINER - to deliver innovative, portable open-source real-time methods for cross-lingual mining and summarisation of large-scale stream media
http://www.trendminer-project.eu/

MEMBERSHIPS

W3C®

STI · INTERNATIONAL

Ontotext USA Inc.
2490 Black Rock Turnpike #331
Fairfield
Connecticut 06825-2400, USA

Contact: Mr. Matthew Petrillo
Tel. +1 (718) 785 9692
Toll-free for N. America:
+1 (866) 972 6686

Ontotext AD
47A Tzarigradsko Chaussee, Floor 4
1504 Sofia, Bulgaria

Contact: Ms. Kamelia Atanasova
Tel: (+359 2) 974 61 60
Fax: (+359 2) 975 32 26

info@ontotext.com
www.ontotext.com