Building (Production) Semantic Applications
Ontotext

- **Leading semantic technology** provider
  - **Top-5** core semantic technology developer
  - Supplying **components** to vendors and solution developers
- Established in year 2000
- Venture capital funding in 2008
- Head quarters in Sofia (Bulgaria); further offices in Fairfield, CT (USA), London (UK), Innsbruck (AT) and Varna (Bulgaria)
- Overall staff: ~70
Ontotext Offering

• **Unique technology** portfolio:
  
  – **Semantic Databases**: high-performance RDF DBMS, SPARQL 1.1 support, scalable RDFS & OWL reasoning
  
  – **Text mining and Semantic Search**: text-mining, IE, IR
  
  – **Web Mining**: focused crawling, screen scraping, data fusion
  
  – **Linked Data Management**: integration, publishing, “reason-able views”

• Part of **research projects** with total budget £100M
  
  – Partnering with SAP, IBM, Wikimedia, Google Labs, BT, Telefonica, KT, BBC, RAI, and leading European universities
OWLIM

- OWLIM is a family of **semantic repositories** for RDF(S) & OWL
  - **OWLIM-Lite**: in-memory, fastest, scales to ~100 million statements
  - **OWLIM-SE**: file-based, sameAs & query optimizations, scales to 20 billion statements
  - **OWLIM-Enterprise**: replication cluster deployment for resilience and high performance parallel query-answering

- OWLIM provides
  - Management, integration & analysis of heterogeneous data
  - Light-weight, **high-performance reasoning**
OWLIM Performance

• OWLIM-SE is the only engine that can reason with more than 10 billion statements.
• OWLIM-SE’s query performance is at least as good as any engine than can handle semantics on 1 Billion statements.
• OWLIM-SE successfully passes Lehigh University Benchmark (LUBM 90000) – over 20 billion explicit & implicit statements.
• OWLIM-SE is openly demonstrated via:
  – FactForge (factforge.net) – ‘general knowledge’ from LOD Cloud.
  – Linked Life Data (linkedlifedata.com) - 25 of the most popular life-science datasets.
FactForge and LinkedLifeData

**FactForge**
- 1.2B explicit
- .9B indexed inferred
- 10B retrievable statements

**LinkedLifeData**
- 2.7B explicit
- 1.4B inferred
- 4.1B indexed statements
Sample FactForge Query

SELECT * WHERE {
  ?Person dbp-ont:birthPlace ?BirthPlace ;
  rdf:type opencyc:Entertainer ;
  owlim:hasPageRank ?RR .
  ?BirthPlace geo-ont:parentFeature dbpedia:Germany .
} ORDER BY DESC(?RR) LIMIT 100

- Who are the most important German entertainers
- This query involves data from DBPedia, Geonames, and UMBEL (OpenCyc)
- It involves inference over types, sub-classes, and transitive relationships
- Ranking the results by 'importance' – RDF rank
FactForge Text Search Query and UI

- Get the descriptions of entities that have a FOAF topic containing ‘American’ and something that is similar to ‘Life’

```
SELECT * WHERE {
}
```

- Full-Text Search over literals and URIs
- Web-based auto-complete
Efficiency Demonstrations

• Proving cost-efficiency is easy with Amazon’s EC2

• Benchmarks of OWLIM Enterprise’s Replication Cluster
  – 100,000 queries for $1 total cost!
  – 5M queries/hour in a 100 node cluster

• Processing TNA’s archive of Government web sites
  – 42 TB of web data
  – 1.3B files
  – 160M unique documents analyzed and indexed
  – 10B facts extracted and managed in an RDF database
  – 40K GBP to process it twice in one month at EC2
BBC World Cup 2010 Website

Delivering content... not pages!

“..... we believe this is the first large scale, mass media site to be using concept extraction, RDF and a Triple store to deliver content.”
“A RDF triplestore and SPARQL approach was chosen over and above traditional relational database technologies due to the requirements for interpretation of metadata with respect to an ontological domain model.”

Jem Rayfield,
Senior Technical Architect, BBC News and Knowledge

“It Begins …”

Comment at ReadWriteWeb’s post on the subject
Next DSP Application

- 200+ Countries
- 400-500 Disciplines
- 10000+ Athletes

"maximising BBC content and data for the audience"
XYZ announced profits in Q3, planning to build a $120M plant in Bulgaria and more and more and more and more and more text.
Building Semantic Applications
BBC Annotation in Graffiti

Man killed as four lorries crash


A man dies in a crash between four lorries, leading to the closure of the M1 and 12-mile tailbacks.

Tag story > Select locations
Select the locations that appear in this story.

Suggested locations

These locations appear in the story:

Milton Keynes, Borough of Milton Keynes, GB
Milton Ernest, GB
Milton Regis, County of Kent, GB
Milton Damerel, County of Devon, GB
Milton, County of Oxfordshire, GB
Milton of Muir, Aberdeen City, GB
Dynamic Semantic Publishing (World Cup)
The World Cup Website Statistics

- More than a **million queries to OWLIM per day**
  - Caching was used in the architecture to allow for handling tens of millions of requests to the web server

- **Hundreds of updates per hour**

- On of a cluster of several machines
  - Typical DB servers with assembly cost below $10,000
BBC Deployment Challenges

• High resilience
  – replication cluster deployment across two data-centres
  – resilient to the loss of a worker node, master node or data-centre

• High performance
  – Parallel query answering across all available worker nodes in both data-centres

• Rapidly changing data
  – New data is inserted and existing data modified/deleted a few hundred times an hour

• Non-trivial inference
  – Expressive OWL inference where logical consistency is maintained on every commit operation (several hundred times an hour)
Data replication is used to:

- Improve scalability of concurrent query requests
- Resilience – failover, online configuration

How does it work?

- Every user write request is pushed in a transaction queue
- Each data write request is multiplexed to all repository instances
- Each read request is dispatched to one instance only
- To ensure load-balancing, each read requests is sent to the instance with smallest execution queue at this point in time
Replication Cluster - Behaviour

- The total **loading/modification performance** of the cluster is equal to that of one instance
- The **data scalability** of the cluster is determined by the amount of RAM of the weakest instance
- The **query performance** of the cluster represents the sum of the throughputs that can be handled by each of the instances
- Failover:
  - In case of failure of one or more instances, the **performance degradation** is graceful
  - The cluster is **fully operational** even when only one instance working
- Cluster can be **reconfigured when running**
Replication Cluster - Types of Nodes

- Two types of nodes
- Flexible topologies possible
- Resilience to failure of workers and masters

**Diagram:**
- **Master**
  - Dispatches queries and updates to workers (read/write)
  - Queries & updates
- **Master (hot standby)**
  - Dispatches queries to workers (read only)
- **Worker 1**
- **Worker 2**
- **Worker 3**

OWLIM-Enterprise master nodes
OWLIM-Enterprise worker nodes (derivative of OWLIM-SE)

Queries &
updates
Queries only

Building Semantic Applications
Full Text Search

• Alternative information access method (different indices)
• Find information based on string elements (tokens)
• Two approaches
  – Node Search – simple, but very fast token matching
  – RDF Search – integrates Lucene to index 'RDF Molecules' with powerful query expressions
• Both methods integrate with SPARQL to allow hybrid searching:
  – Example: “Show all instances of renaissance artist based in Netherlands whose name begins ‘Rem’ and rank according to most well-linked”
RDF Rank

- OWLIM-SE includes a plug-in that allows for efficient calculation of a modification of PageRank over RDF graphs
- Computation of rank values is fast, e.g.
  - 400M LOD statements takes **310 sec** (27 iterations)
- Results are available through a system predicate
- Example: get the 100 most important nodes in the RDF graph

```
SELECT ?n {?n rank:hasRDFRank ?r}
ORDER BY DESC(?r) LIMIT 100
```
Notifications

• The client can subscribe for notifications for incoming statements matching desired statement patterns

• The patterns are then used to filter incoming statements
  – Notify the subscriber about those statements that help form a new solution of at least one of the graph patterns
  – Inferred statements are treated in the same way

• The subscriber should not rely on any particular order or distinctness of the statement notifications
  – Both inserted and deleted statements are notified

• High performance 'in-process' notifications

• Remote notification mechanism
Geo-spatial Extensions

- Allows for the high performance evaluation of geo-spatial queries
- Uses the WGS84 ontology
- Find points within rectangles, polygons or circles
- Compute great circle distance between points
- Example: (ordered) nearest airports to London:

```
SELECT DISTINCT ?airport {
  ?airport omgeo:nearby(?lat1 ?long1 "50mi");
  a dbp-ont:Airport ;
  geo:lat ?lat2 ; geo:long ?long2 .}
ORDER BY ASC(omgeo:distance(?lat1, ?long1, ?lat2, ?long2))
```
Financial Publishing Case Study

Building Semantic Applications
Financial Publishing Case Study

Article Title

Consectetuer adipiscing elit, sed diam nonumy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volupat. Ut wisi enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Duis autem vel eum irure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et justo odio dignissim qui blandit praesent luptatum.

Ut wisi enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Organisations mentioned in this article

Client 1
Client 2
Client 3

Law firms mentioned in this article

Firm 1
Firm 2
Firm 3

Lawyers mentioned in this article

Lawyer 1 + IFLR rank
Lawyer 2 + IFLR rank
Lawyer 3 + IFLR rank

Clicking one of these will open the listing for this firm in the same window.

If we are extracting cases, and they mention specific lawyers, then clicking a lawyer name would return a list of these cases.

I believe this is not included in the PoC though.
Financial Publishing Case Study

Firm name

Consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum irure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisi. At vero eros et accumsan et justo odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Nam liber tempor cum soluta

Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum irure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisi at vero eros et accumsan et justo odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Nam liber tempor cum soluta

Notable clients

| Client 1 |
| Client 2 |
| Client 3 |

Related articles

| Article 1 |
| Article 2 |
| Article 3 |

Clicking a client will display a list of all firms that have represented them – i.e. Firms Ontotext have extracted from the MS Word questionnaires.

Clicking loads this article