



Metadata management for the BBC's 2010 World Cup site using OWLIM

Marin Dimitrov (Ontotext)

European Semantic Technology Conference 2010

Contents

- About Ontotext
- BBC's dynamic semantic publishing platform
- Key OWLIM features
- Metadata management for the BBC's 2010 World Cup site using OWLIM

About Ontotext

- Semantic technology provider
 - **Est. in 2000** as part of Sirma Group
 - Spin-off since 2008 (VC investment acquired in Jul/2008)
 - Offices in **Bulgaria** and **USA**
 - **55 employees** and multiple contractors
- Unique technology portfolio
 - High performance **RDF Databases**
 - **Semantic Annotation & Search**
 - **Linked Data** Management & RDF based data integration
 - **Web Mining**

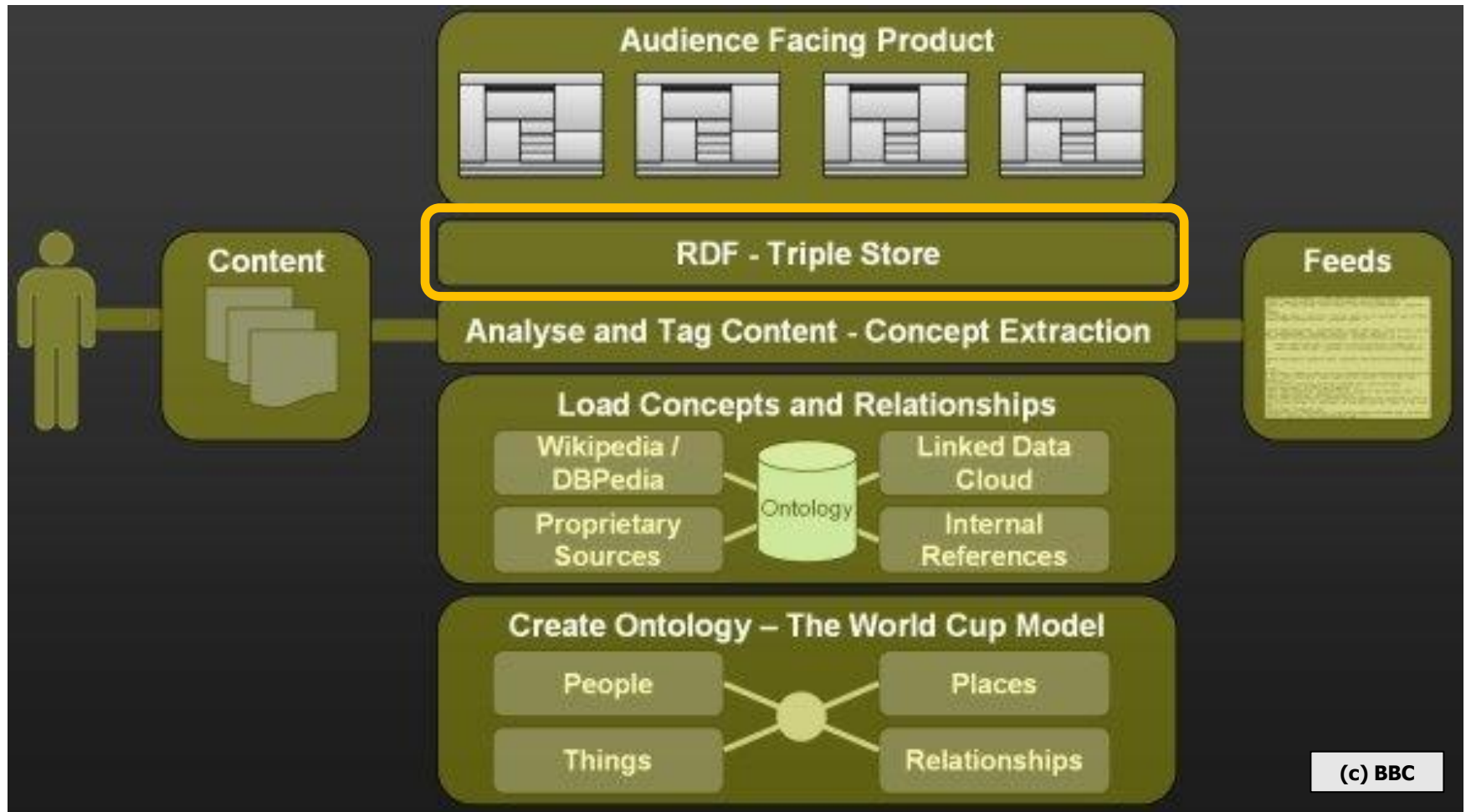
BBC's dynamic semantic publishing

- High-performance dynamic publishing
 - Metadata for rich content relationships & semantic navigation
 - Automated metadata driven web pages – dynamically generated by querying the aggregated metadata
- Ontological models
 - Domain specific (about football / World Cup) and generic (e.g. FOAF)
 - Journalist-authored stories associated to ontology concepts

BBC's dynamic semantic publishing (2)

- Text analysis
 - Journalist-authored stories are analysed and ontology concepts are automatically extracted
- RDF-ization of external data
 - External datasources and feeds are RDF-ized and mapped to the World Cup ontology
- Metadata repository (triplestore)
 - Agile modelling & reduced query complexity
 - Inference of implicit facts (forward chaining)
 - Resilient cluster
 - 1-2 million SPARQL queries per day

BBC's dynamic semantic publishing (3)



OWLIM – a scalable RDF database

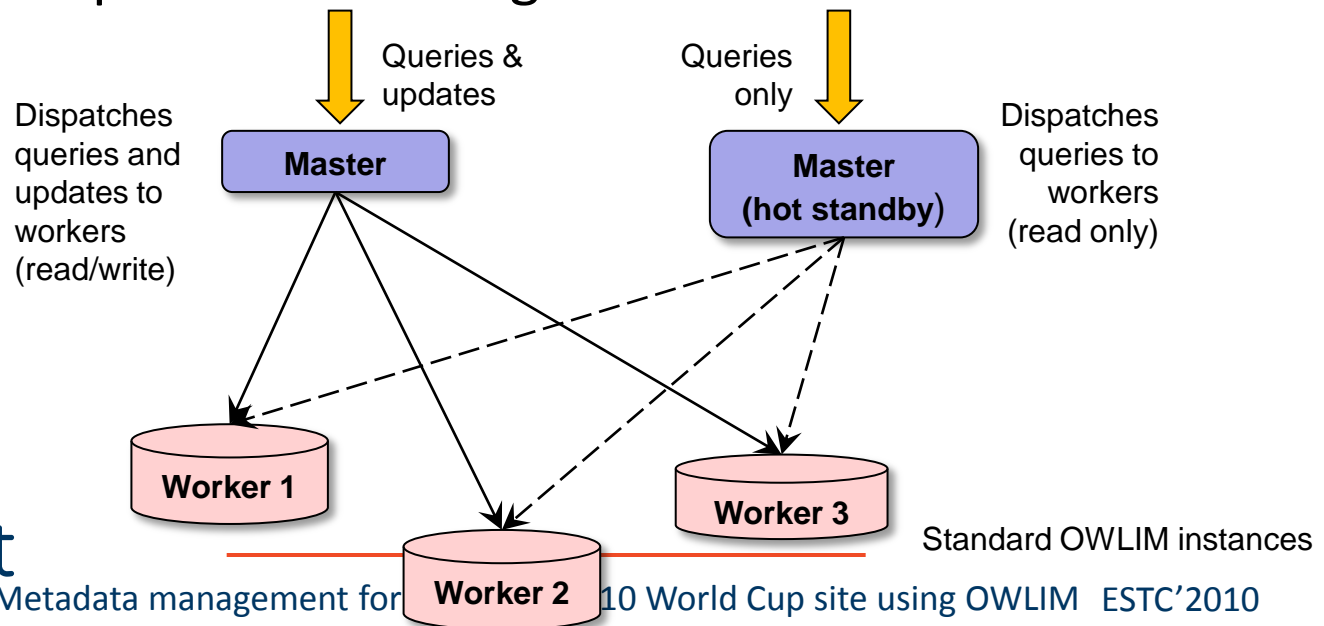
- Fast and scalable materialisation
- High-performance reasoning
 - Full **RDFS**, OWL-Horst, **OWL 2 RL** and **OWL 2 QL**^(new)
 - Restricted OWL Lite
 - ... or specify custom rule-sets
- Optimised handling of equivalence classes (owl:sameAs)
- Fast retraction of statements
- Query optimisations

OWLIM – a scalable RDF database (2)

- Fully compatible with Sesame 2, Jena adapter^(new)
- Replication cluster
- RDF search / full-text search
- Geo-spatial extensions^(new)
- Expressive consistency checks
- Very fast RDF Rank
- RDF Priming (customizable spreading activation)
- Available on Amazon EC2^(new)
 - Amazon AutoScaling integration

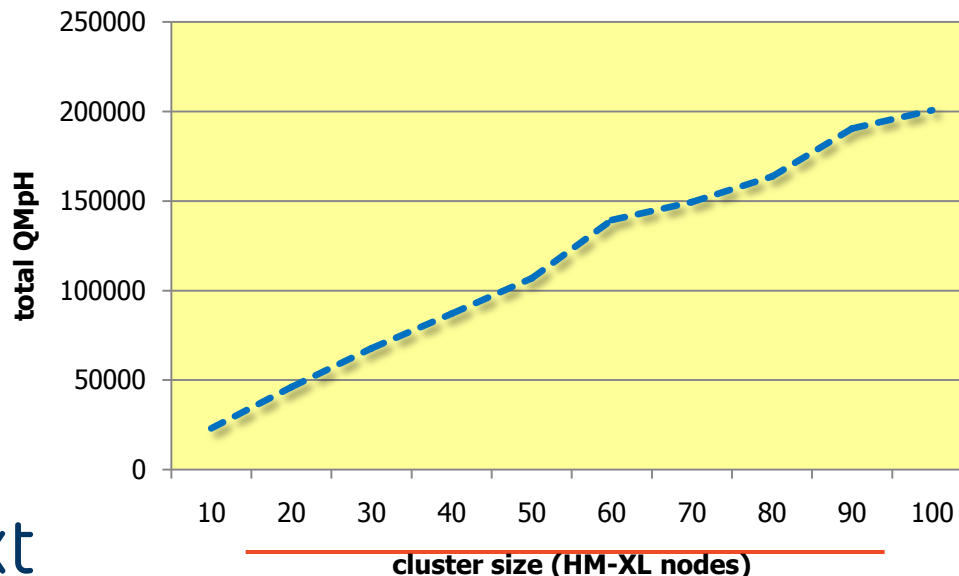
OWLIM – Replication Cluster

- Improved scalability & resilience
 - Load-balancing of queries
 - Query performance of the cluster = sum of the throughputs of all the nodes
 - Updates are multiplexed across all instances
 - Graceful performance degradation when a node fails



OWLIM @ AWS – millions of queries on the Cloud

- Large scale test of an OWLIM cluster on Amazon EC2
 - Up to 100 nodes
 - Almost **linear scalability** of query performance
 - **5 million SPARQL queries per hour**
 - BSBM benchmark, 100M dataset, HM-2XL master, HM-XL workers
 - Cost efficiency – 100,000 queries per \$1 on EC2



Using OWLIM for the BBC World Cup site

- Replication Cluster
 - 2 datacenters, 2 masters, 6 workers
 - **Redundancy** of all masters and workers
- Information stored in the OWLIM database
 - **Ontologies** (domain specific or general)
 - **Factual knowledge** (about teams, players, games, etc.)
 - **Metadata** about the authored content (extracted references to the ontologies)
- **Constant updates** to reflect the real-time content stream
 - hundreds per hour

Using OWLIM for the BBC World Cup site (2)

- **SPARQL queries** to generate dynamic web pages from the metadata stored in OWLIM
 - **1-2 million** SPARQL queries per day
- **Reasoning** to infer new facts
 - Forward chaining
 - Complexity in-between RDFS and OWL 2 RL
- **Fast retraction** of deleted statements
 - Ontologies are marked as read-only to limit the scope of update propagation

Key OWLIM advantages for this scenario

- **Fast queries**
 - Forward chaining / materialisation is mandatory
- **Fast materialisation** of new facts
 - Real-time content stream, cannot be batched
- **Fast retraction** of deleted statements
 - Allows immediate updates of the factual knowledge
 - avoids full re-computation of the materialised closure
- **Replication cluster**
 - Query scalability
 - Failover (when a worker node goes down)

BBC World Cup 2010 website

“... this is the first large scale, mass media site to be using concept extraction, RDF and a Triple store to deliver content.”

John O'Donovan, Chief Technical Architect, Journalism and Knowledge, BBC Future Media & Technology



The screenshot shows the BBC website's 'SPORT' section for the 'WORLD CUP 2010'. The main navigation bar includes 'SPORT', 'FOOTBALL', 'WORLD CUP 2010', 'GROUPS & TEAMS', 'FIXTURES & RESULTS', 'VIDEO', and 'BBC COVERAGE'. The page is focused on the 'England' team. It features a 'Latest matches' section with results for England 1-1 USA, England 0-0 Algeria, and Germany 4-1 England. A 'Group C Teams' table is visible, showing England's record. The 'Latest stories' section includes headlines like 'Gerrard commits future to England' and 'Pressure got to Rooney - Ferguson'. A 'Features' section highlights 'German lessons' and 'A German view on English football'. The 'Around the web' section lists external links.

“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



The screenshot shows the BBC website's 'SPORT' section for the 'WORLD CUP 2010', specifically the profile page for Frank Lampard. The navigation bar is the same as the previous screenshot. The page features a 'Latest matches' section. The main content area is dedicated to Frank Lampard, showing his position as a 'Midfielder', squad number '9', and date of birth '20 June, 1978 (32 years old)'. It includes 'Tournament totals' for 'Games played' (4) and 'Goals' (0). A 'Shots on target | off target' section shows 9|5 for Lampard and 38|23 for the total England team. Other statistics include 'Assists' (0), 'Fauls by | on' (3|4), and 'Cards yellow | red' (0|0). The 'Features' section includes 'World Cup scouting: attacking midfielder' and 'World Cup scouting: defensive midfielder'. The 'Top 5 World Cup stories' and 'Top TV and radio' sections are also visible.

Conclusions

- Enterprises are starting to understand the advantages of Semantic Technologies
 - **Cost efficient** data integration of heterogeneous data sources
 - **Agile** data modelling & reduced query complexity
 - **Reasoning** to discover implicit knowledge
- OWLIM is a scalable, enterprise ready RDF database

Q & A

Questions?

twitter @ontotext