ontotext

Динамичното семантично публикуване в Би Би Си (E*mpowering Dynamic Semantic Publishing at the BBC)*

CESAR, META-NET Meeting, Sofia

May 2012

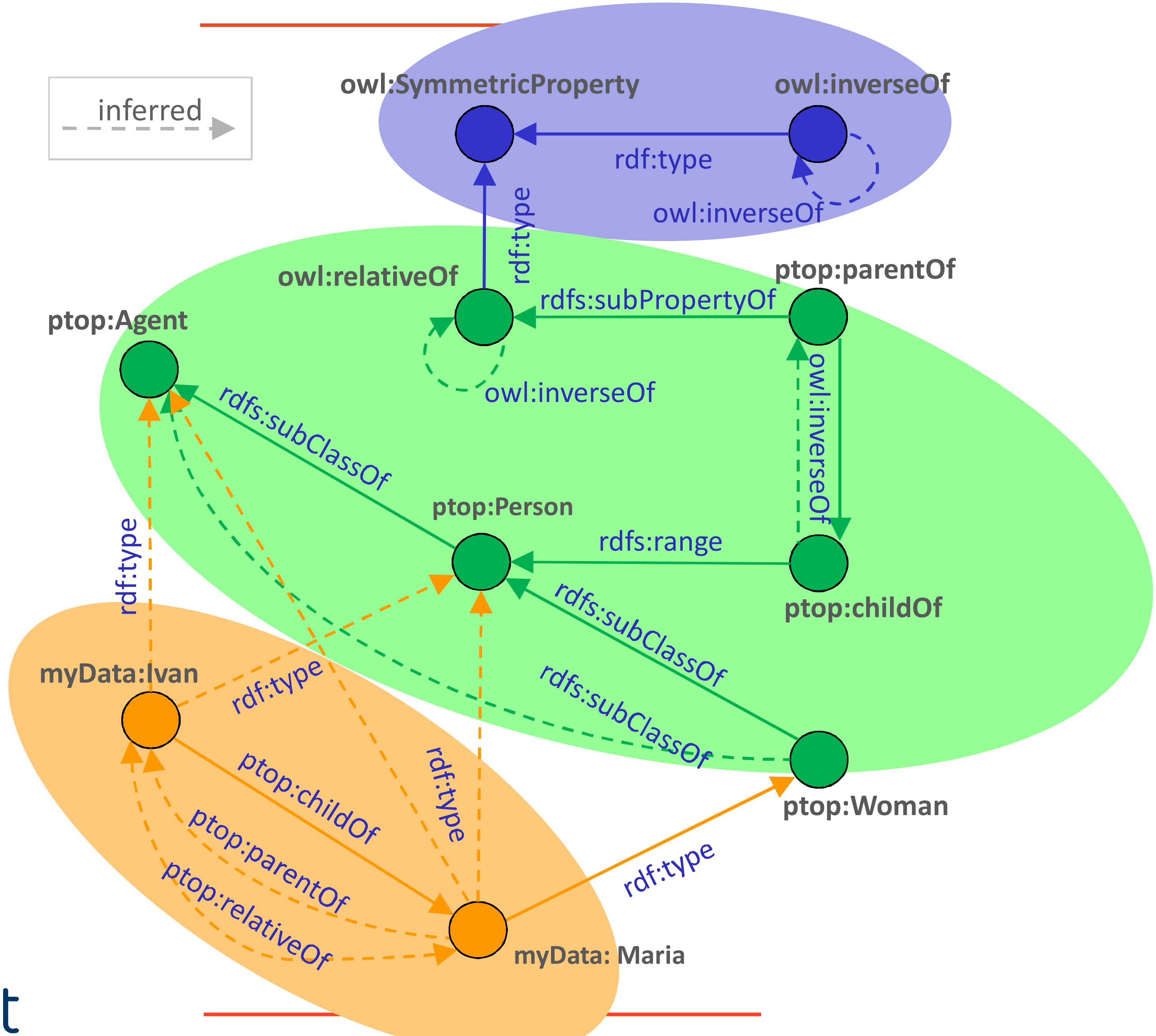
Presentation Outline

- Ontotext
- Linked data
- BBC’s Business case
- The solution

Ontotext Brief

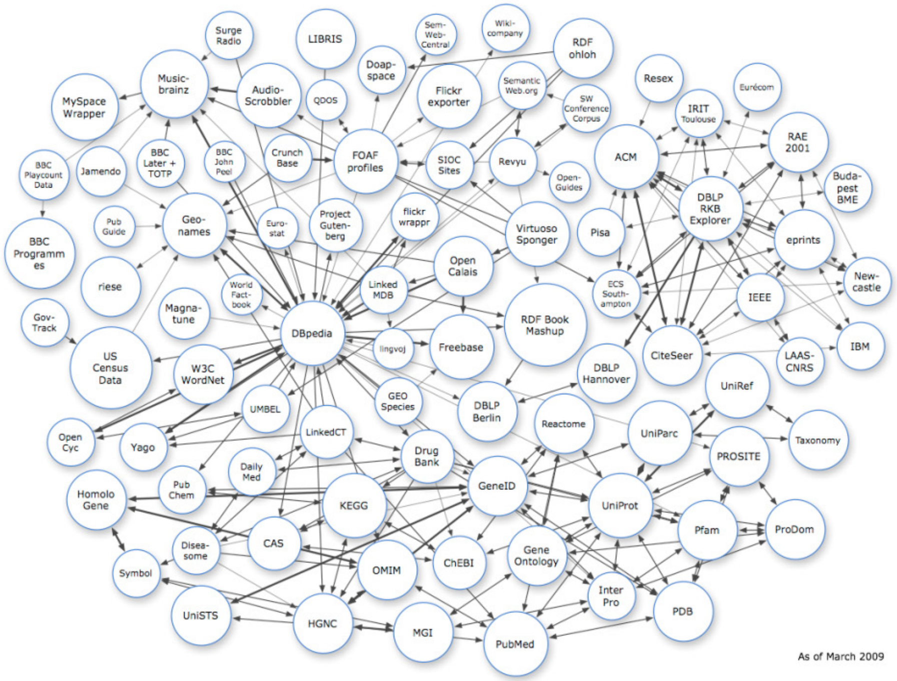
- **Semantic technology developer**
  - Established in year 2000 as part of Sirma Group
  - 65+ employees
  - Offices in Bulgaria (Sofia and Varna), USA (Fairfield, CT), London
- **Global leader** in semantic databases and search
  - Competing with ORACLE, IBM, Google and few specialized companies
- Delivered the highest profile SemTech application
  - **The BBC's 2010 World Cup web site**
- **Customers** include: BBC, AstraZeneca, Korea Telecom
  - Press Association, Telecom Italia, Natural Resources Canada, The National Archive, UK, The British Museum, UK Parliament

Linking Data Across Different Servers



Linking Open Data (LOD)

- Linking Open Data W3C SWEO Community project
  http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData



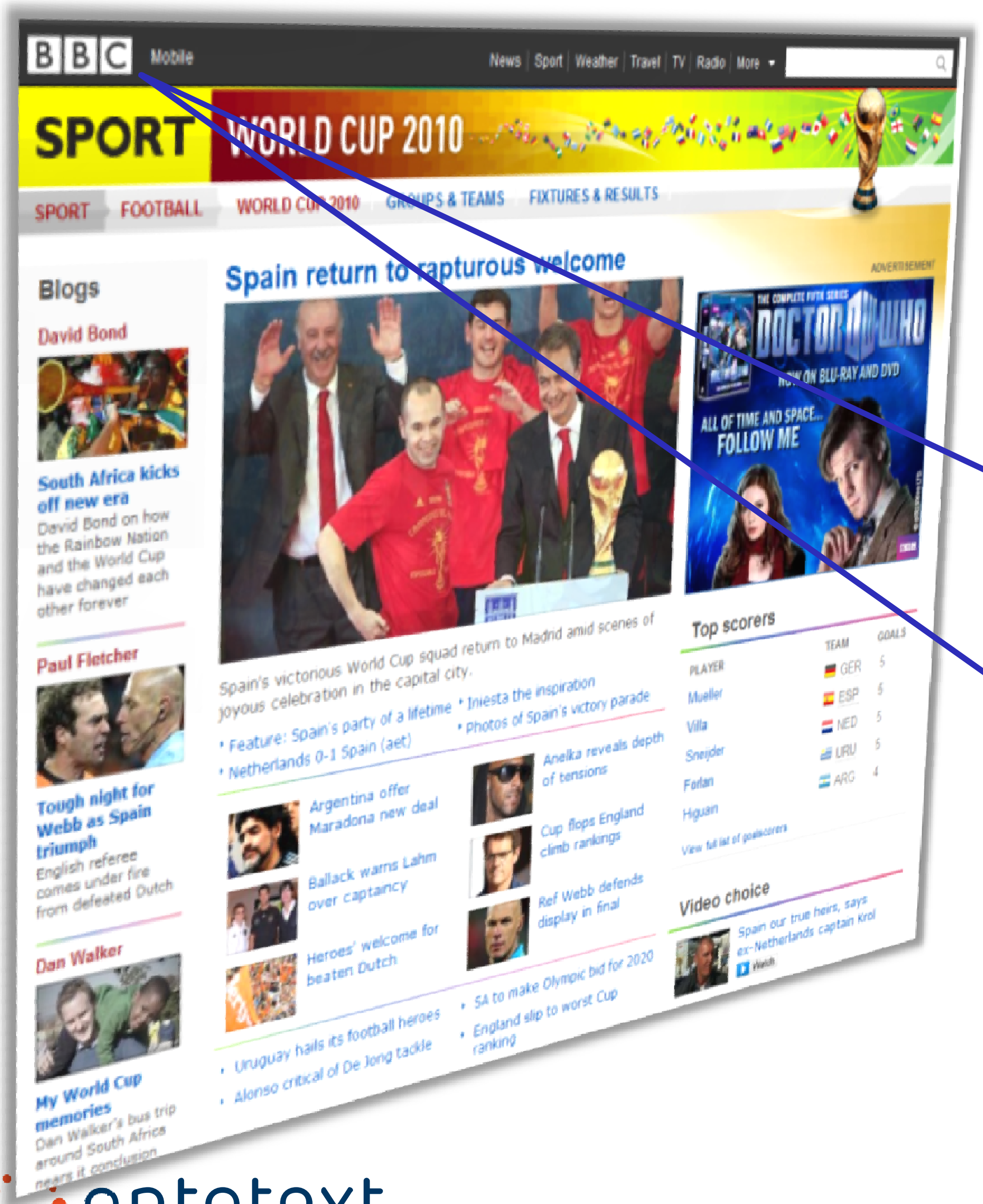
As of March 2009

- Initiative for publishing “linked data” which includes 400+ interlinked datasets and about 50B facts

Presentation Outline

- Ontotext
- Linked data
- **BBC’s Business case**
- The solution

BBC World Cup 2010 Website



BBC Mobile News | Sport | Weather | Travel | TV | Radio | More

SPORT WORLD CUP 2010

SPORT FOOTBALL WORLD CUP 2010 GROUPS & TEAMS FIXTURES & RESULTS

Blogs

David Bond

South Africa kicks off new era
David Bond on how the Rainbow Nation and the World Cup have changed each other forever

Paul Fletcher

Tough night for Webb as Spain triumph
English referee comes under fire from defeated Dutch

Dan Walker

My World Cup memories
Dan Walker's bus trip around South Africa nears it conclusion

Spain return to rapturous welcome

Spain's victorious World Cup squad return to Madrid amid scenes of joyous celebration in the capital city.

- Feature: Spain's party of a lifetime
- Netherlands 0-1 Spain (aet)
- Iniesta the inspiration
- Photos of Spain's victory parade

Argentina offer Maradona new deal

Ballack warns Lahm over captaincy

Heroes' welcome for beaten Dutch

Anelka reveals depth of tensions

Cup flops England climb rankings

Ref Webb defends display in final

- Uruguay hails its football heroes
- Alonso critical of De Jong tackle
- SA to make Olympic bid for 2020
- England slip to worst Cup ranking

ADVERTISEMENT

Top scorers

PLAYER	TEAM	GOALS
Mueller	GER	5
Villa	ESP	5
Sneijder	NED	5
Forlan	URU	5
Higuan	ARG	4

View full list of goalscorers

Video choice

Spain our true heirs, says ex-Netherlands captain Krol

Watch

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."***
>
> -- **John O'Donovan**, *Chief Technical Architect, Journalism and Knowledge, BBC Future Media & Technology*

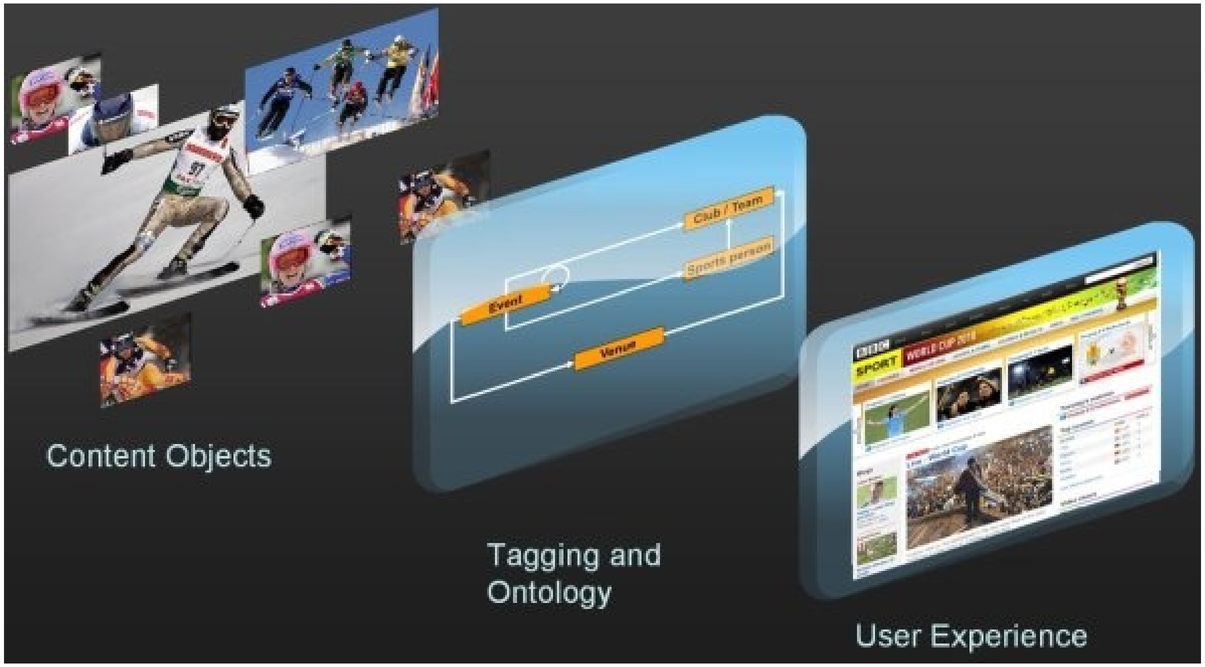
ontotext

The World Cup Website Scenario

*“**The World Cup site is a large site** with over 700 aggregation pages (called index pages) designed to lead you on to the thousands of story pages and content …*

*…**we are not publishing pages, but publishing content** as assets which are then organised by the metadata dynamically into pages, but could be re-organised into any format we want much more easily than we could before.*

*… The index pages are published automatically. This process is what assures us of the highest quality output, but still **saves large amounts of time** in managing the site and makes it possible for us to **efficiently run so many pages** for the World Cup.“*



John O'Donovan,
Chief Technical Architect, BBC Future Media & Technology
http://www.bbc.co.uk/blogs/bbcinternet/2010/07/the_world_cup_and_a_call_to_ac.html

BigOWLIM Powered the BBC's World Cup Web Site

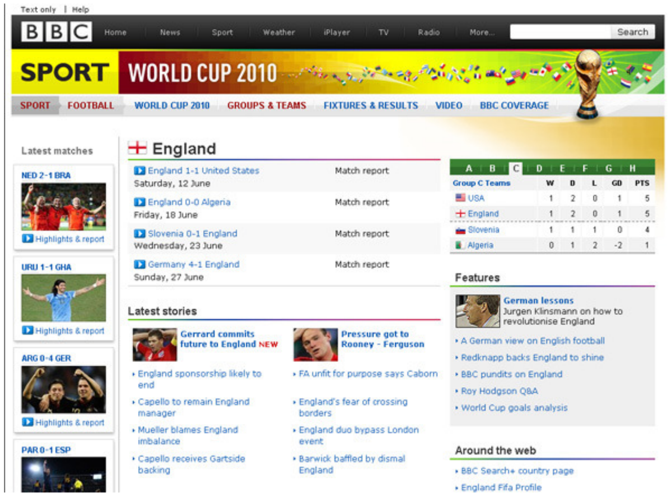
"A RDF triplestore and SPARQL approach was ***chosen over and above traditional relational database*** *technologies due to the requirements for interpretation of metadata with respect to an ontological domain model."*

Jem Rayfield,

Senior Technical Architect, BBC News and Knowledge

"It Begins ..."

A comment at ReadWriteWeb's post on the subject



Latest matches

NED 2-1 BRA
Highlights & report

URU 1-1 GHA
Highlights & report

ARG 0-4 GER
Highlights & report

PAR 0-1 ESP

England

England 1-1 United States — Match report
Saturday, 12 June

England 0-0 Algeria — Match report
Friday, 18 June

Slovenia 0-1 England — Match report
Wednesday, 23 June

Germany 4-1 England — Match report
Sunday, 27 June

Latest stories

Gerrard commits future to England NEW

Pressure got to Rooney - Ferguson

- England sponsorship likely to end
- Capello to remain England manager
- Mueller blames England imbalance
- Capello receives Gartside backing
- FA unfit for purpose says Caborn
- England's fear of crossing borders
- England duo bypass London event
- Barwick baffled by dismal England

A B C D E F G H

Group C Teams	W	D	L	GD	PTS
USA	1	2	0	1	5
England	1	2	0	1	5
Slovenia	1	1	1	0	4
Algeria	0	1	2	-2	1

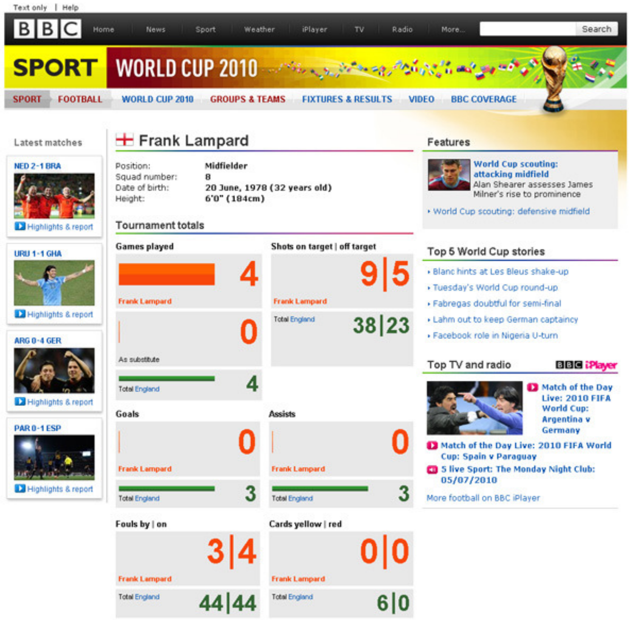
Features

German lessons
Jurgen Klinsmann on how to revolutionise England

- A German view on English football
- Redknapp backs England to shine
- BBC pundits on England
- Roy Hodgson Q&A
- World Cup goals analysis

Around the web

- BBC Search+ country page
- England Fifa Profile



Latest matches

NED 2-1 BRA
Highlights & report

URU 1-1 GHA
Highlights & report

ARG 0-4 GER
Highlights & report

PAR 0-1 ESP
Highlights & report

Frank Lampard

Position: Midfielder
Squad number: 8
Date of birth: 20 June, 1978 (32 years old)
Height: 6'0" (184cm)

Tournament totals

Games played
4 Frank Lampard
0 As substitute
4 Total England

Shots on target | off target
9|5 Frank Lampard
38|23 Total England

Goals
0 Frank Lampard
3 Total England

Assists
0 Frank Lampard
3 Total England

Fouls by | on
3|4 Frank Lampard
44|44 Total England

Cards yellow | red
0|0 Frank Lampard
6|0 Total England

Features

World Cup scouting: attacking midfield
Alan Shearer assesses James Milner's rise to prominence

- World Cup scouting: defensive midfield

Top 5 World Cup stories

- Blanc hints at Les Bleus shake-up
- Tuesday's World Cup round-up
- Fabregas doubtful for semi-final
- Lahm out to keep German captaincy
- Facebook role in Nigeria U-turn

Top TV and radio

Match of the Day Live: 2010 FIFA World Cup: Argentina v Germany

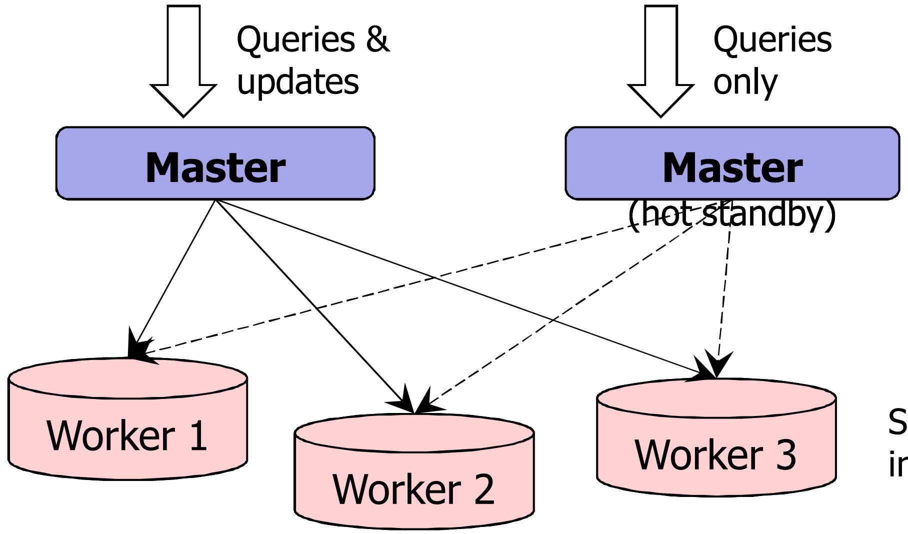
Match of the Day Live: 2010 FIFA World Cup: Spain v Paraguay

5 live Sport: The Monday Night Club: 05/07/2010

More football on BBC iPlayer

Statistics

- More than a **million of queries to OWLIM per day**
  - Caching was used in the architecture t allow for handling 10s of millions of requests to the web server
- **Hundreds of updates per hour**
- Out of a cluster of several machines
  - Typical DB servers with assembly cost below $10,000



Dispatches queries and updates to workers (read/write)

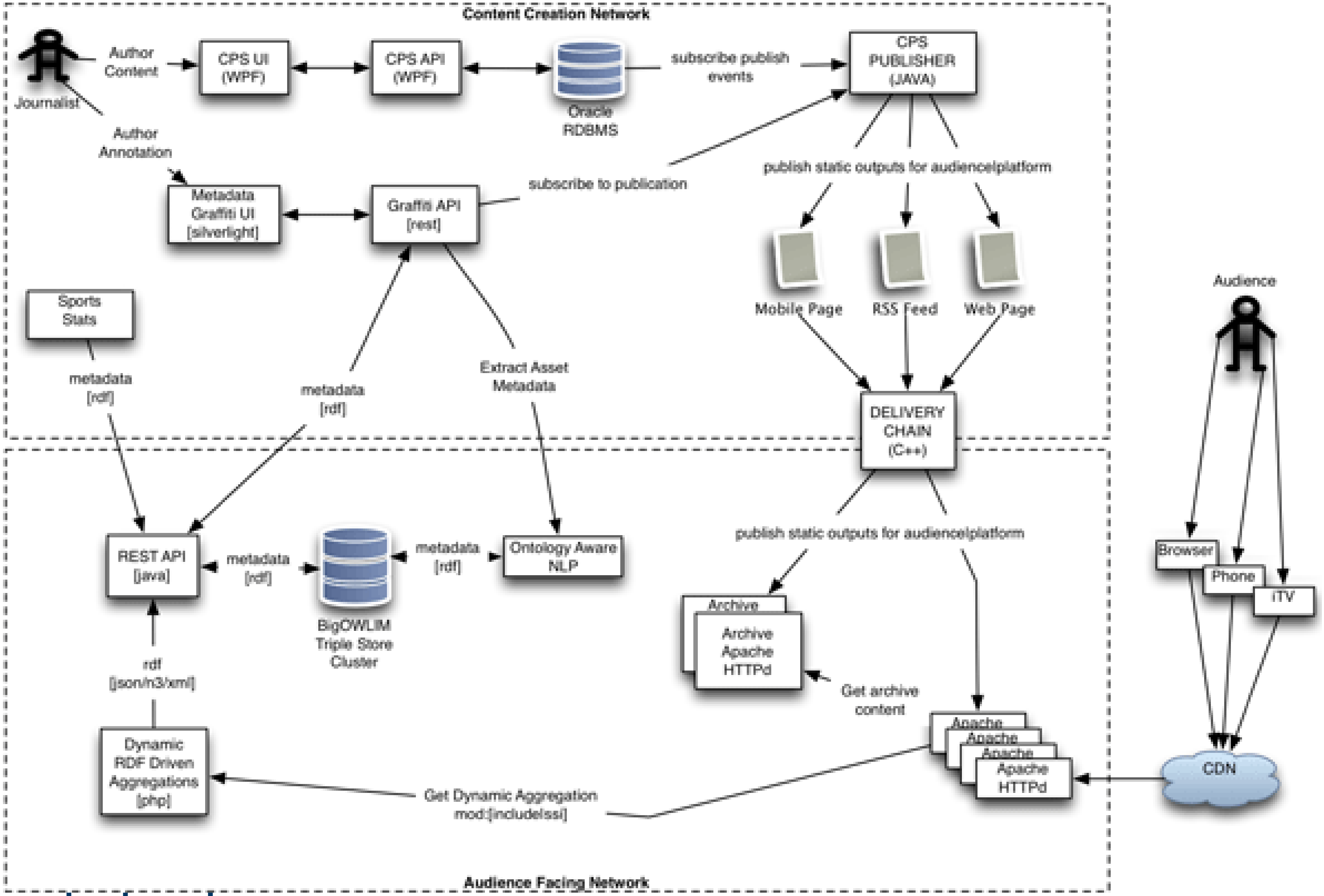
Dispatches queries to workers (read only)

Standard BigOWLIM instances

2012 Update

- Ontotext implements also the “Concept Extraction Serivce”
- The technology is rolled out to be used in BBC Sport and Olympics 2012 websites
- Major press release from BBC can be found at http://www.bbc.co.uk/blogs/bbcinternet/2012/04/sports_dynamic_semantic.html
  - Some of the materials on the next slides come from there

2012 Architecture

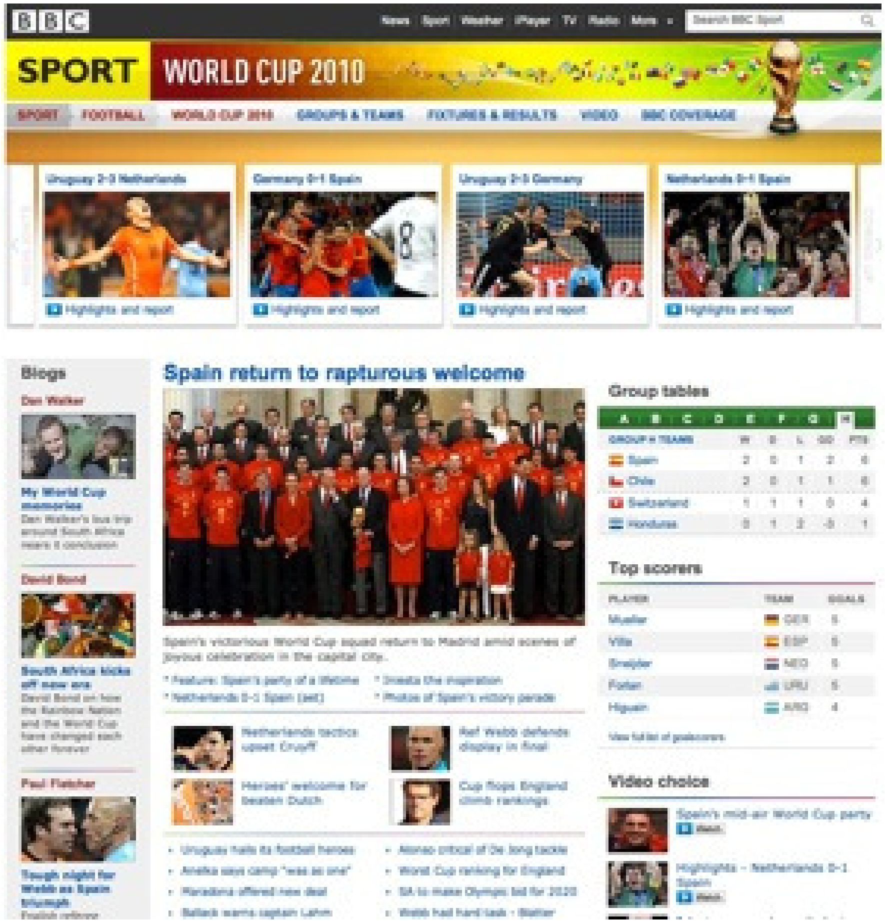


The Challenge

- Recognize and identify People, Teams, Tournaments and Locations in Sports News
- Use Linked data as primary data source
- Deal with high ambiguity of candidates
- Calculate Confidence factor of the recognition
- Calculate Relevance factor of the association of an entity with an article
- Do all this for under a second per document
- Involve journalists in a correction feedback mechanism
- Adapt recognition quality on the basis of this feedback

Final Objective

- Identify relevant entities associated with an article
- To create aggregated pages for teams, players or events



BBC

News Sport Weather iPlayer TV Radio More

Search BBC Sport

SPORT WORLD CUP 2010

SPORT FOOTBALL WORLD CUP 2010 GROUPS & TEAMS FIXTURES & RESULTS VIDEO BBC COVERAGE

Uruguay 2-3 Netherlands
Highlights and report

Germany 0-1 Spain
Highlights and report

Uruguay 2-3 Germany
Highlights and report

Netherlands 0-1 Spain
Highlights and report

Blogs

Dan Walker

My World Cup memories
Dan Walker's bus trip around South Africa nears its conclusion

David Bond

South Africa kicks off new era
David Bond on how the Rainbow Nation and the World Cup have changed each other forever

Paul Fletcher

Tough night for Webb as Spain triumph
English referee

Spain return to rapturous welcome

Spain's victorious World Cup squad return to Madrid amid scenes of joyous celebration in the capital city.

- Feature: Spain's party of a lifetime
- Iniesta the inspiration
- Netherlands 0-1 Spain (aet)
- Photos of Spain's victory parade

Netherlands tactics upset Cruyff

Ref Webb defends display in final

Heroes' welcome for beaten Dutch

Cup flops England climb rankings

- Uruguay hails its football heroes
- Alonso critical of De Jong tackle
- Anelka says camp "was as one"
- World Cup ranking for England
- Maradona offered new deal
- SA to make Olympic bid for 2020
- Ballack warns captain Lahm
- Webb had hard task - Blatter

Group tables

A B C D E F G H

GROUP H TEAMS	W	D	L	GD	PTS
Spain	2	0	1	2	6
Chile	2	0	1	1	6
Switzerland	1	1	1	0	4
Honduras	0	1	2	-3	1

Top scorers

PLAYER	TEAM	GOALS
Muller	GER	5
Villa	ESP	5
Sneijder	NED	5
Forlan	URU	5
Higuain	ARG	4

View full list of goalscorers

Video choice

Spain's mid-air World Cup party
Watch

Highlights – Netherlands 0-1 Spain
Watch

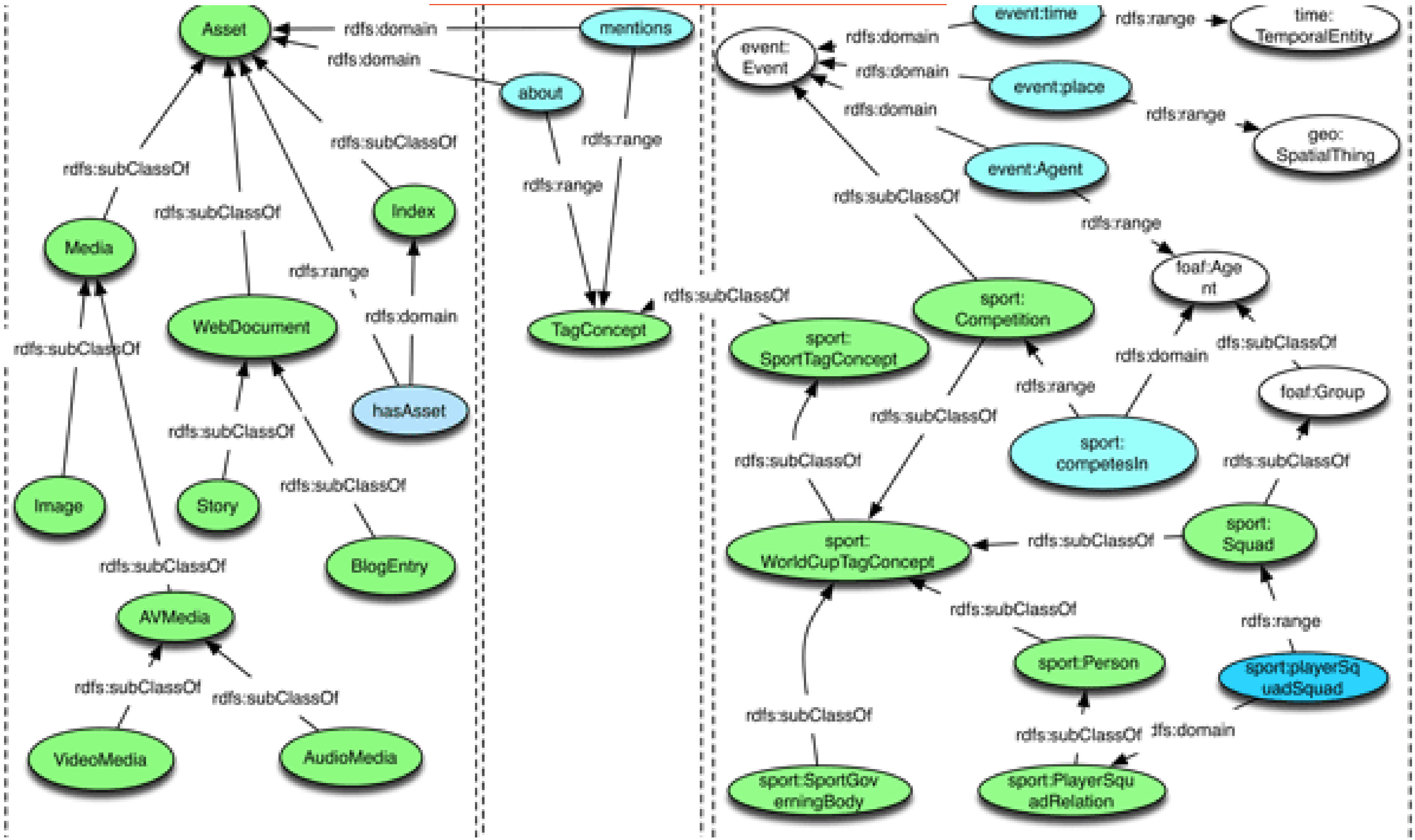
Know It All Data Sets

- BBC-developed ontology of sports entity classes and relationships
- Extensive data set of teams, players and types of sports
- Semi-automatically distilled sub set of GeoNames
- Tens of millions of RDF triples served by an OWLIM Enterprise Cluster

Presentation Outline

- Ontotext
- Linked data
- BBC’s Business case
- **The solution**

BBC Ontologies



Find the Candidates

- Finding mentions of entities from the knowledge base: not an issue
- Ontotext used proprietary LKB Gazetteer
- Nowadays a completely new Linked Data Gazetteer is in use
- Due to the millions of entities – high rate of over generation and high ambiguity
- Up to 20 candidates for a mention in some cases

Disambiguation Approach

- Disambiguation based on multitude of features in the vicinity of the mention and training of a Max Entropy classifier
- Graph based disambiguation using relatedness of entities
- Geospatial awareness used to disambiguate entities

Entity Confidence

- Disambiguate entities on the basis of text context in the vicinity of an entity
- Confidence scores for each candidate entity e
- Ontological graph context of the vicinity of an entity
- Frequencies of entities in the corpus and document.
- We do it with 75-90% measured in terms of F1 score
- Linear model: *h(x) = argmax_y f(x,y)* w,* where
  - *f(x,y)* is a feature function
  - *w* is a model parameter vector

Entity Relevance

- Rank entities with respect to their relatedness to the article
- We achieve 75% accuracy
- We consider:
  - Frequencies in the document and in the corpus
  - Mentions in the title (a separate field)

– For each entity *e* and each field *f* we calculate the local frequency of *e* in *f*. Also we calculate the global frequency of *e* in the corpus. The final relevance score is a combination of the local and global frequencies.

– Approach is very close to the Zaragoza et al 2004, see here: Microsoft Cambridge at TREC–13: Web and HARD tracks, Zaragoza, Craswellet, Taylor, Saria and Robertson 2004

Entity Relevance - More

The pseudo frequency of an entity e in the filed f for document d is:

$$pf(d,f,e) = f(d,f,e)/(1+B(f)*(len(d,f)/len(f)-1)),$$

where

- *f(d,f,e)* - is the real frequency of an entity *e* in the filed *f* of the document *d*,
- *len(d,f)* - is the length of the field *f* in *d*;
- len(f) - is the average length of the field *f* in all documents;
- *B(f)* - is a length normalising factor between 0 and 1.

Entity Relevance – Even More

The pseudo frequency of an entity e in the document d is:

$$pf(d,e) = W(title)*pf(d,title,e) + W(body)*pf(d,body,e),$$

where

- *W(title)* and *W(body)* are parameters greater than 0

The irrelevance of an entity *e* with respect to a document *d* is

$$irr(d,e) = -log(pf(d,e)/(K+pf(d,e))) - alpha*log(df(e)),$$

where

- *K* is a parameter greater than 0;
- *alpha* is the factor of global importance, it is greater than 0
- *df(e)* is the document frequency of the entity *e*.

Geospatial Disambiguation

- **Geospatial distance** - a feature of OWLIM
- **Super region** – GeoNames hierarchy and containment relations, e.g. parentFeature
- **RDF Rank**
- **Human approval score** (on the basis of curated documents)
- **Class/code based priority** – fine grained ontology may allow a rule or machine learning prioritization of classes and entities based on learning we already have.
- **Asset geo association** - some entities could be disambiguated by using the asset domain association. BBC UK local sports is more likely to talk about national entities.

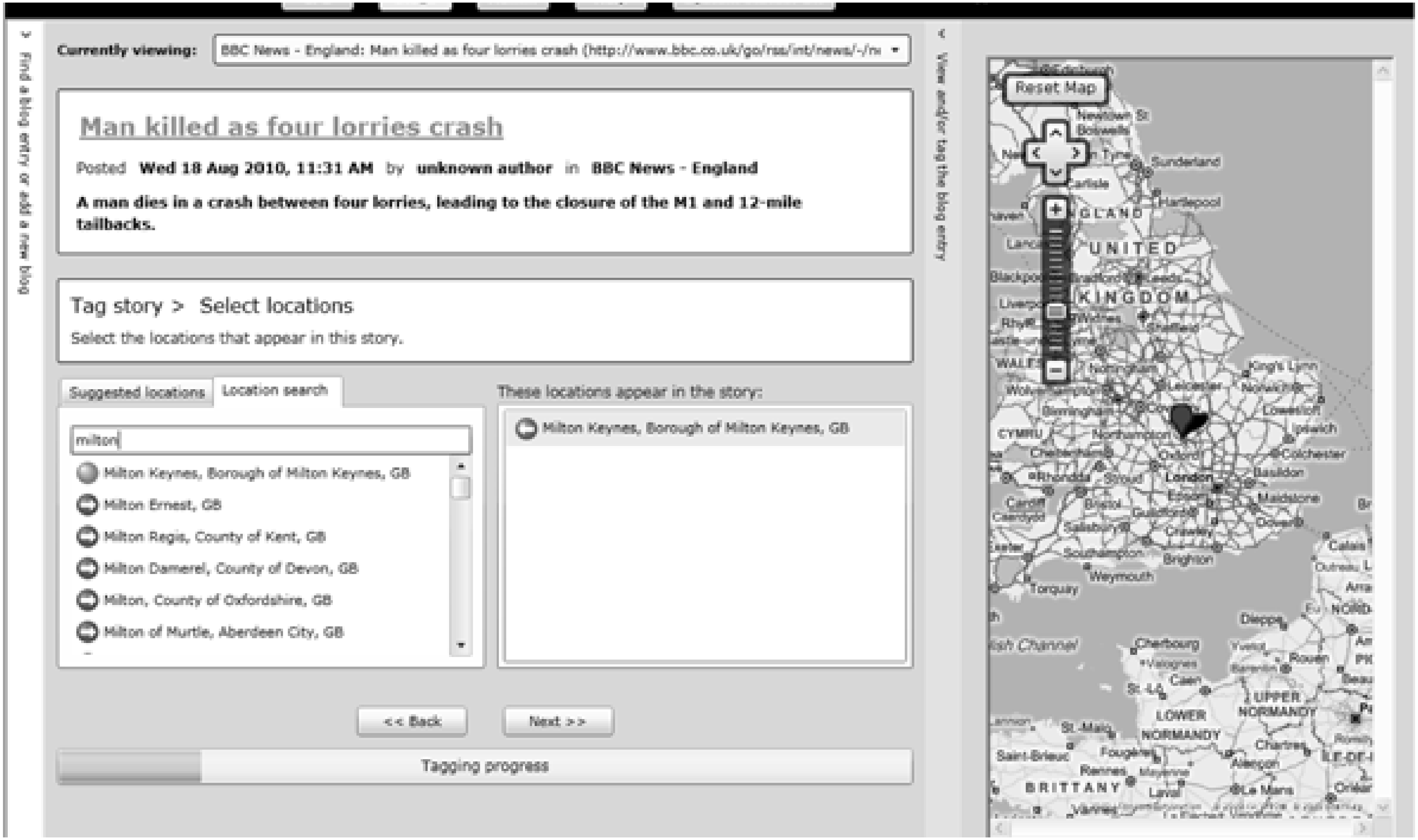
Geospatial Disambiguation

- **Geospatial distance** - a feature of OWLIM
- **Super region** – GeoNames hierarchy and containment relations, e.g. parentFeature
- **RDF Rank**
- **Human approval score** (on the basis of curated documents)
- **Class/code based priority** – fine grained ontology may allow a rule or machine learning prioritization of classes and entities based on learning we already have.
- **Asset geo association** - some entities could be disambiguated by using the asset domain association. BBC UK local sports is more likely to talk about national entities.

Adaptation

- Journalists get sorted suggestions from the automatic extraction
- They correct the suggestions by adding, removing or re-ordering
- A metadata fingerprint of the articles with their entity mentions is stored in OWLIM
- Each 2 weeks the models are being retrained with these new examples

Curation Interface



Currently viewing: BBC News - England: Man killed as four lorries crash (http://www.bbc.co.uk/go/rss/int/news/-/n

Man killed as four lorries crash

Posted **Wed 18 Aug 2010, 11:31 AM** by **unknown author** in **BBC News - England**

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 | Location search

milton

- 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 Murtle, Aberdeen City, GB

These locations appear in the story:

- Milton Keynes, Borough of Milton Keynes, GB

In Production

- Currently the resulting pipeline is a part of the new Dynamic Semantic Publishing platform behind the BBC Sports web site
- It is shaped as an updatable and parallelizable Concept Extraction Service
- Through adaptation of the models, the BBC will also apply it for the news about the 2012 Olympics

Examples

- **Chelsea Football Club**: All the content objects associated to the concept *"Chelsea“*
  http://www.bbc.co.uk/sport/football/teams/chelsea
- **Tom Daley**: All the content objects associated to the concept *"Tom Daley“*
  http://www.bbc.co.uk/sport/olympics/2012/athletes/02025fcb-457d-4a77-8424-f5b8fe49b87f
- **Team GB**: All the content objects associated to the concept *"Team GB“*
  http://www.bbc.co.uk/sport/olympics/2012/countries/great-britain

Thank you!

We develop core semantic technology

Ontotext invested 300 person-years, partnered with 100 leading groups, created some of the most popular tools, and delivered multiple solutions.

We know what works and what doesn’t

Ontotext set many benchmarks and advanced the frontiers of the semantic databases.

We invented the “semantic annotation” – linking text with data

Now we are prepared to

interlink your data, your content, and the web