LDBC SPB – News Monitoring Scenario Experiment

Semantic Publishing Benchmark (SPB) with real news and big open data at FactForge

Venelin Kotsev, Atanas Kiryakov
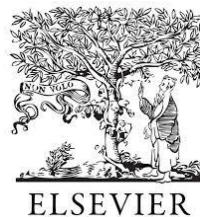
LDBC 9th TUC, SAP Waldorf, Feb 2017

Ontotext Introduction

- *Unique technology mix*: Graph DB engine + Text mining

- *Robust technology*: We run BBC.CO.UK/SPORT and parts of FT.COM

- We serve the most knowledge intensive enterprises

BBC

FT FINANCIAL TIMES

EUROMONEY

Top-5 business information agency

GETTY

HOUSES OF PARLIAMENT INFORMATION & COMMUNICATIONS TECHNOLOGY

OXFORD UNIVERSITY PRESS

WILEY Publishers Since 1807

ELSEVIER

DK

THE BRITISH MUSEUM

AstraZeneca

ontotext

Presentation Outline

- Introduction to Semantic Publishing Benchmark (SPB)

- Use cases: Relation discovery and Media monitoring

- FactForge: Open data and news about people and organizations

- SPB News Monitoring Variant

ontotext

Introduction to the original LDBC Semantic Publishing Benchmark (SPB)

FIFA World Cup 2010

Health

World Cup 2010 in South Africa 'triggered baby boy boom'

🕐 7 November 2015 | Health

Researchers believe people in South Africa were less stressed during the World Cup in 2010, which increased the ratio of boys born

AFP/Getty Images

Top Stories

US university chief quits in race row

University of Missouri President Tim Wolfe resigns amid criticism of his handling of racial issues.

🕐 19 minutes ago

Jordan policeman kills two US military

🕐 38 minutes ago

Athletics doping chiefs urge Russia ban

🕐 22 minutes ago

BBC's DSP Approach

"The goal is to be able to more easily and accurately aggregate content, find it and share it across many sources. From these simple relationships and building blocks you can dynamically build up incredibly rich sites and navigation on any platform."

John O'Donovan,

Chief Technical Architect, BBC

http://www.bbc.co.uk/blogs/bbcinternet/2012/04/sports_dynamic_semantic.html

Jem Rayfield, Lead Architect, BBC

ontotext

BBC's DSP Approach

- **The SPB was inspired by the Dynamic Semantic Publishing**
  - First applied at BBC's FIFA Worldcup 2010 website
  - Next at website of BBC and the official one for the 2012 Olympics
  - Now continuously used at BBC Sport and by many others (e.g. FT)

- **The BBC's primary use case: rich, deep, dynamic websites**
  - Text-mining automatically annotates articles with entities
  - Editor curates the metadata before storing it in a triplestore
  - Thematic web pages are generated on-the-fly through SPARQL

ontotext

LDBC SPB – Semantic Publishing Benchmark

- **A benchmark for RDF Databases**

- **Scenario: maintenance of a metadata catalogue for media assets**
  - Simulates metadata consumption & management operations
  - Measures the performance of both types of operations

- **Developed by: Ontotext and FORTH**
  - with support from BBC and OpenLink

- **References and documentation**
  - http://ldbcouncil.org/benchmarks/spb
  - https://github.com/ldbc/ldbc_spb_bm/tree/master/doc

ontotext
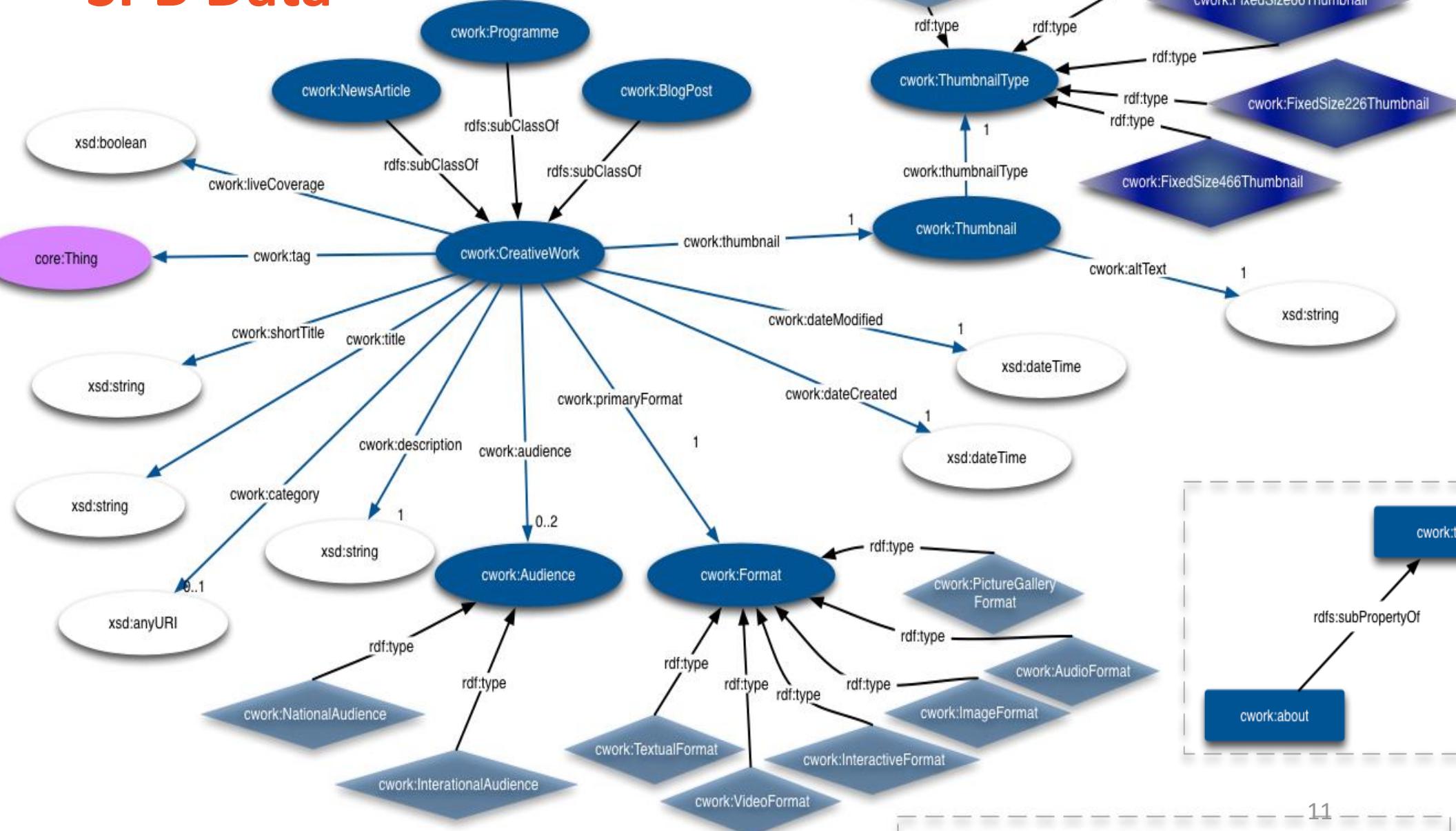
LDBC SPB – Semantic Publishing Benchmark (2)

- **Triplestore holds both reference data and content metadata**

  – **Reference data**: master data in Dynamic Semantic Publishing

  – E.g. taxonomies and entity datasets used to describe "creative works"

- **The main interactions with the repository are:**

  – **updates**, adding new metadata or updating the reference knowledge

    ▪ For a big news/sport/media site there are 10s of small transactions/second

  – **aggregation queries** retrieve content according to various criteria

    ▪ At peak moments site like BBC.CO.UK/SPORT gets more than million of requests per hour which translates into more than 100 queries/second (after some caching)

ontotext

SPB Data

- **The dataset can be generated at different scale factors**
  - 64M statements (SF1), 128M, 256M, 512M, 1B (SF5), …

- **Fixed size Reference data: 22M triples**
  - 7M statement **spatial data** from GeoNames about Europe
  - 14M statements from DBPedia for **Companies, Persons, Events**

- **The Metadata can be scaled up:**
  - There are 25 annotations per "creative work" (digital media asset)
    - On average 9 tags describing assets by linking them to topics or entities
  - 1B dataset contains metadata for 40M assets

ontotext

SPB Data

LDBC SPB – Changes in v.2.0

- **Much bigger reference dataset: from 170k to 22M**
  - 1M persons, 650k locations, 85k companies, 50k events

- **Interconnected reference data: more than 0.5M links**
  - **owl:sameAs links** between DBPedia locations and Geonames; more on the next slide

- **More comprehensive usage of inference**
  - Still the simplest possible flavor of OWL is used
    - **rdfs:subClassOf, rdfs:subPorpertyOf, owl:TransitiveProperty, owl:sameAs**
  - Transitive closure over company control and geographic nesting
  - Allows for simpler queries through usage of super-property

- **Retrieval of relevant content through links in the reference data**

ontotext

Interconnected Reference Dataset

- **Substantial volume of connections between entities**

- **Geonames comes with hierarchical relationship, defining nesting of locations,** `gn:parentFeature`

- **DBPedia inter-entity relationships between entities*:**

	To Company	**To Person**	**To Place / gn:Feature**	**To Event**
Company	40,797	26,675	218,636	18
Person	89,506	1,324,425	3,380,145	145,892
Event	5,114	154,207	140,579	35,442

ontotext

LDBC-SPB Requirements

- **Support for quadruples; triples plus context/named graphs**
  - Support for TRIG or NQ RDF serialization syntax

- **SPARQL Query 1.1 support**

- **SPARQL Update 1.1 support**
  - READ COMMITTED transaction isolation
  - Consistent handling of updates, without batching

- **SPARQL Protocol 1.1 (known as "SPARQL End-point")**

- **Inference support – OWL 2 RL semantics is sufficient**

- **Specific features: FTS, Geo-spatial, owl:sameAs**
  - Not a requirement, but optimized support for those helps performance

ontotext

SPB WORKLOADS: Editorial Agents

- **Simulate the work performed by journalists or editors**

  – E.g. enriching journalistic assets with meta-data: description, creation date, location etc.

- **Run simultaneously**

- **Provide a constant stream of update operations**

- **Editorial operations:**

  – INSERT, DELETE, UPDATE (Insert + Delete)

ontotext

SPB WORKLOADS: Aggregation agents

- **Simulate the interactions of end-users or semi-automated tools**

  – E.g. queries generated by an application that dynamically generates web pages for wide range of topics (teams, players, events, etc.) at BBC Sport website

- **Run simultaneously**

- **Provide a constant workload of queries from two available query mixes:**

  – Basic query mix of 12 queries

  – Advanced query mix of 25 queries

ontotext

SPB WORKLOADS: Aggregation agents

- **Basic (interactive) query mix (12 Queries) :**
    - Search queries
    - Full-text search queries
    - Aggregation queries
    - Geospatial queries

- **Advanced (interactive + analytical) query mixes (25 Queries):**
    - Analytical queries
    - Faceted search queries
    - Drill-down queries

ontotext

Interactive Workload

- **Q1: All info about the recent creative works tagged with thing T**

- **Q2: Retrieve details about a given creative work**

- **Q3: Retrieve a list of blog post and news, matching some criteria**

- **Q4: Creative works tagged with thing T, ordered by creation date**

- **Q5: The most popular topics, under some CW criteria**

- **Q6: Creative works mentioning location within a geospatial rectangle**

- **Q7: Creative works from specific period, ordered**

- **Q8: Creative works by word contained in its title (FTS)**

- **Q9: Retrieve 10 similar creative works, based on tags**

ontotext

Interactive Workload (ctd)

- **Q10: Retrieve CWs that mention locations in the same province (A.ADM1) as the specified one**

  - There is additional constraint on time interval (5 days)

- **Q11: Retrieve the most recent CWs that are tagged with entities, related to a specific popular entity**

  - Relations can be inbound and outbound; explicit or inferred

- **Q12: Retrieve all information about CWs**

  - Using SPARQL GRAPH clause

ontotext

Q10: News from the region

```
SELECT ?cw ?title ?dateModified {
  <http://dbpedia.org/resource/Sofia> geo-ont:parentFeature ?province .
  ?province geo-ont:featureCode geo-ont:A.ADM1 .
  {
    ?location geo-ont:parentFeature ?province .
  } UNION {
    BIND(?province as ?location) .
  }
  ?cw a cwork:CreativeWork ;
      cwork:tag ?location ;
      cwork:title ?title ;
      cwork:dateModified ?dateModified .

  FILTER(?dateModified >= "2011-05-14T00:00:00.000"^^<http://www.w3.org/2001/XMLSchema#dateTime>
      &&  ?dateModified < "2011-05-19T23:59:59.999"^^<http://www.w3.org/2001/XMLSchema#dateTime>)
}
LIMIT 100
```

ontotext

Q11: News on about related entities

```
SELECT DISTINCT ?cw ?title ?description ?dateModified ?primaryContent
{
  {
    <http://dbpedia.org/resource/Teresa_Fedor> ?p ?e .
  }
  UNION
  {
    ?e ?p <http://dbpedia.org/resource/Teresa_Fedor> .
  }

  ?e a core:Thing .

  ?cw cwork:tag ?e ;
    cwork:title ?title ;
    cwork:description ?description ;
    cwork:dateModified ?dateModified ;
    bbc:primaryContentOf ?primaryContent .
}
ORDER BY DESC(?dateModified)
LIMIT 100
```

ontotext

DATA GENERATOR

- **Reference data and ontologies**
  - Real ontologies and reference datasets provided by the BBC
  - Reference data includes LOD datasets: DBpedia and Geonames

- **Metadata generation**
  - Parallel data generation
  - Deterministic
  - Scalable - scales to billions of triples

- **Generated datasets simulate the activity of a publishing organization for a period of time**
  - Media assets are enriched by metadata called 'Creative Works'

ontotext

Metadata for "Popular Entities"

- **The Data Generator uses a set of "popular entities"**

  – Those are referred to in 30% of the content-to-entity relations/tags

  – Heuristics used to produce more realistic data distributions, e.g. story lines

- **Popular entities are those with top 5% RDF Rank**

  – Entities which are more often used to tag content also have better connectivity in the Reference Data

- **In the future RDF-ranks can be used for other purposes also**

  – E.g. ordering, disambiguation, etc.

ontotext

LDBC-SPB Test Driver

- **The SPB test driver:**

  – Open Source

  – Available on GitHub: https://github.com/ldbc/ldbc_spb_bm

- **Runs Editorial and Aggregation agents simultaneously**

  – Parallel execution

  – Provide a steady update stream during query workload

- **Validates Results**

  – Query results completeness and correctness

  – Update operations; batching is not allowed

- **Gathers and reports performance metrics**

ontotext

SPB Official Results (Basic Interactive Query Mix)

Engine	Scale	R/W Agents	Hardware	Load time (sec)	Reads /sec	Updates /sec
GraphDB SE 6.3	64M	8/2	1xXeon (6 cores), 64GB, SSD	2 766	143	11
Virtuoso OS 7.5	64M	22/2	2xXeon (2x6 cores) 192GB, HDD	1 270	149	157
GraphDB SE 6.3	256M	8/2	1xXeon (6 cores), 64GB, SSD	19 655	56	10
Virtuoso OS 7.5	256M	22/2	2xXeon (2x6 cores) 192GB, HDD	5 436	81	93
Virtuoso OS 7.5	256M	30/3	AWS r3.8xlarge, 32 vCPU, 60GB, SDD	6 120	115	110
Virtuoso OS 7.5	1B	22/2	2xXeon (2x6 cores) 192GB, HDD	24 035	32	73
Virtuoso OS 7.5	1B	30/3	AWS r3.8xlarge, 32 vCPU, 60GB, SDD	12 142	46	55

ontotext

Sizing SPB in the AWS Could

Within the LDBC project we performed experiments to find:
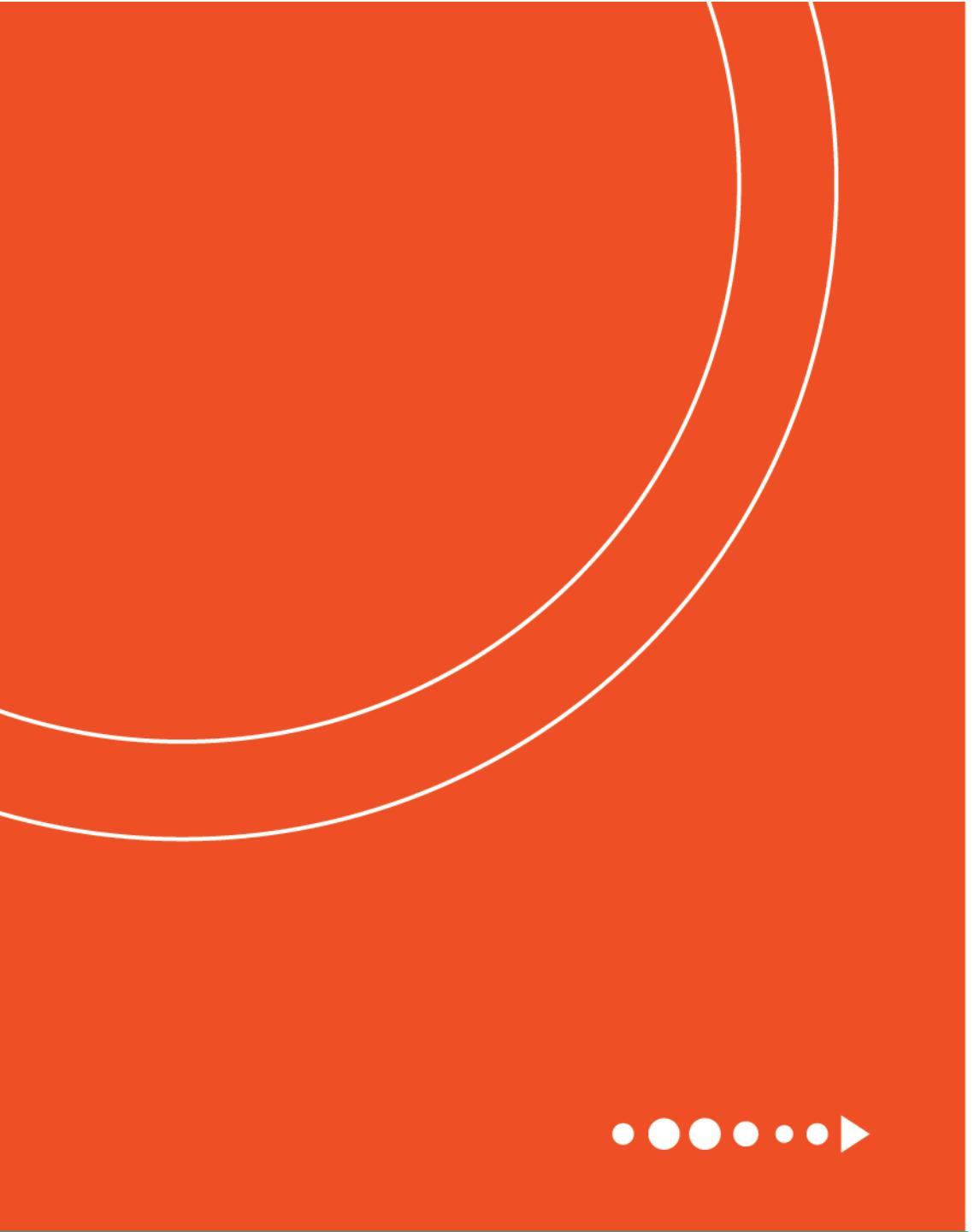
- **The most cost-effective AWS instance types for SPB**

- **Sample AWS costs for a production-grade cluster**

- **Full results published at:**
  http://ldbcouncil.org/blog/sizing-aws-instances-semantic-publishing-benchmark

Some of the key findings:

- **$1 spent on c3.2xlarge allows for handling 140 000 queries**
  - With 8GB of RAM and 8 vCPUs this is the best instance for SPB 50M (2M docs)

- **$81/day cost cluster that can sustain 100 reads queries/sec. and 10 updates/sec.**
  - 2 masters and 6 workers, all of them c3.2xlarge instance type
  - Such cluster can handle the load of the triplestore at BBC's FIFA Worldcup website

ontotext

What is SPB Good for?

- **Engine shutout**

  – Implemented on both GraphDB and Virtuoso

  – MarkLogic and BlazeGraph expressed interest to implement it

  – There are no assumptions about inference strategy or special-purpose plug-ins

- **Performance engineering and sizing**

  – It's easy to replace the datasets and the queries

- **Quality Assurance**

  – avoid performance degradation

- **Training**

  – Broad variety of queries

  – Realistic & understandable dataset allow experiments with Linked Open Data

ontotext
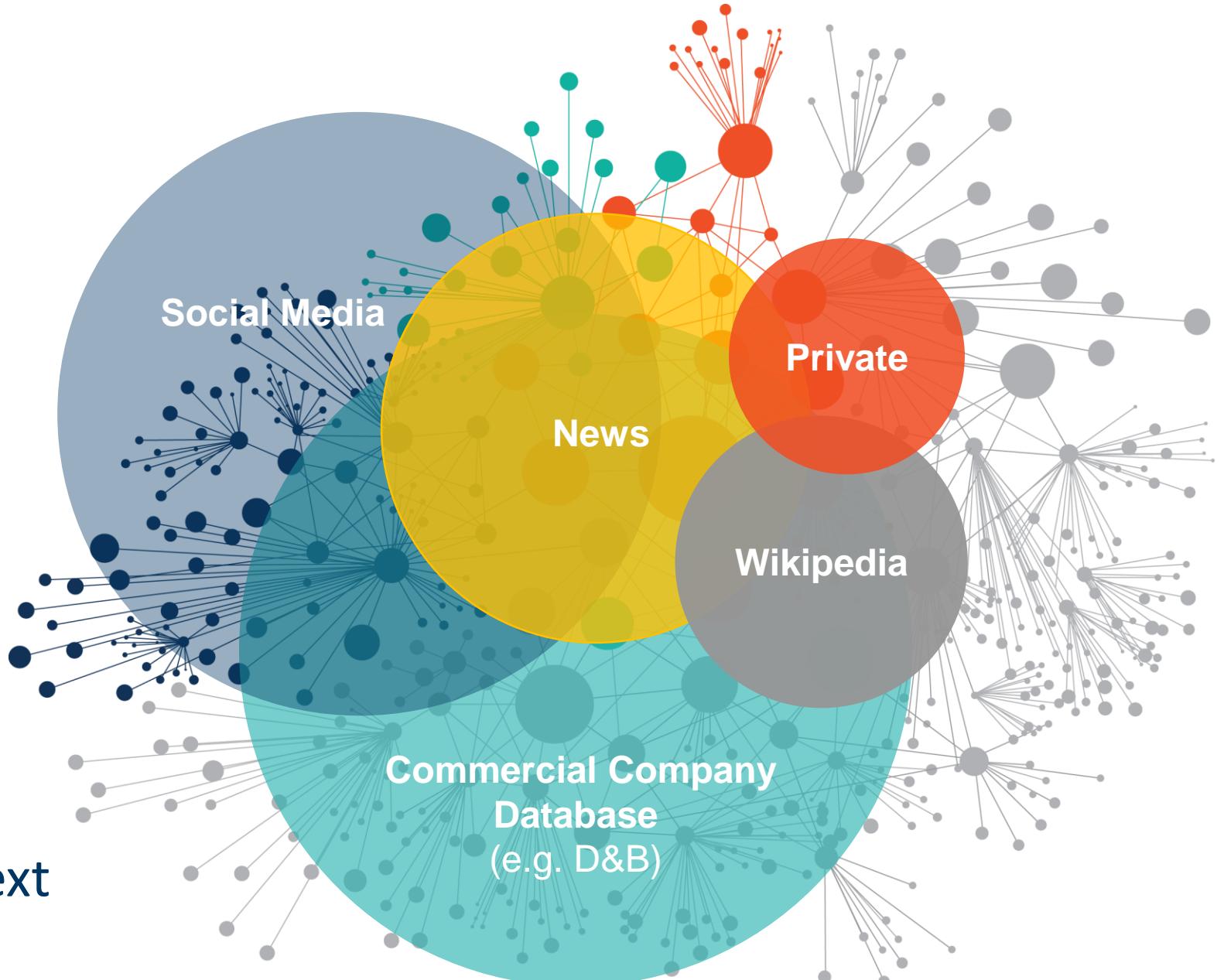
Use cases: Relation discovery and Media monitoring
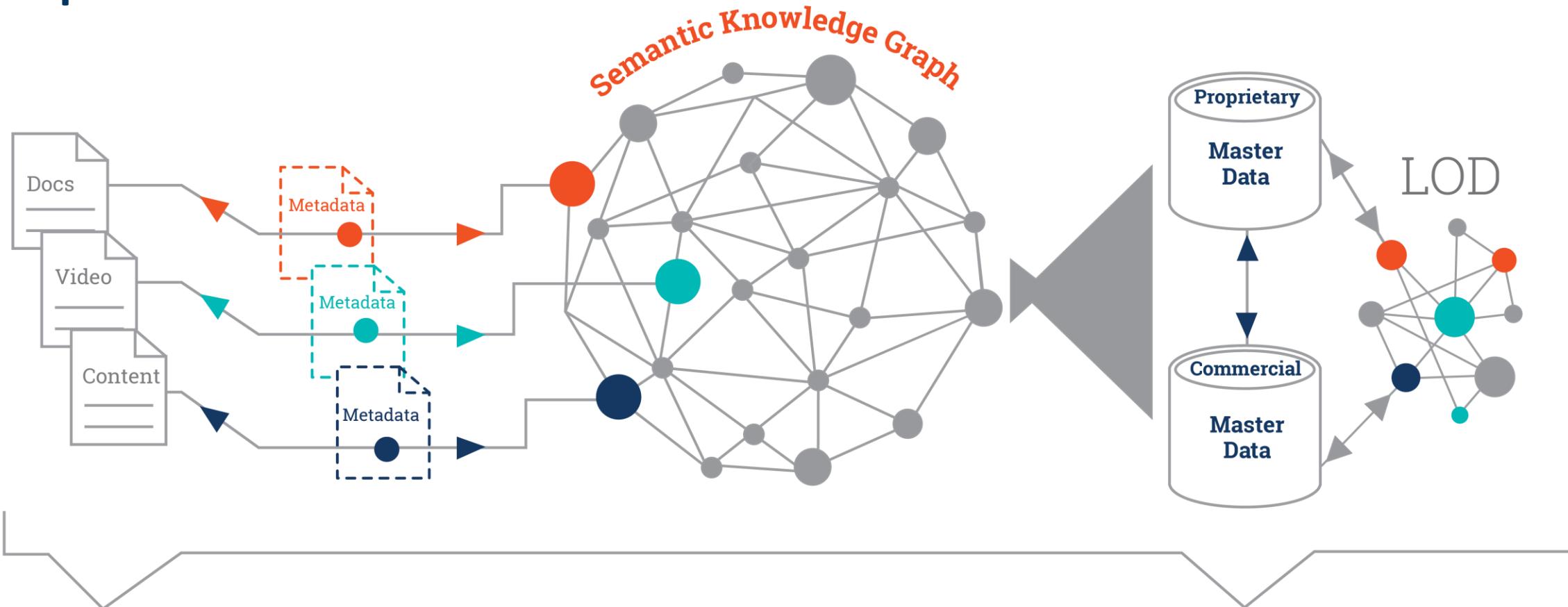
Link data!
Reveal more!

- Link diverse data in a Knowledge Graph

- Analyze News and Social Content

- Extract facts and link content to data

- Interpret data in context of big linked data

Social Media

News

Private

Wikipedia

Commercial Company Database
(e.g. D&B)

ontotext

Content Analytics & Exploration Platform

GraphDB

Linked Open Data

Semantic Knowledge Graph

Docs

Video

Content

Metadata

Metadata

Metadata

Proprietary

Master Data

Commercial

Master Data

LOD
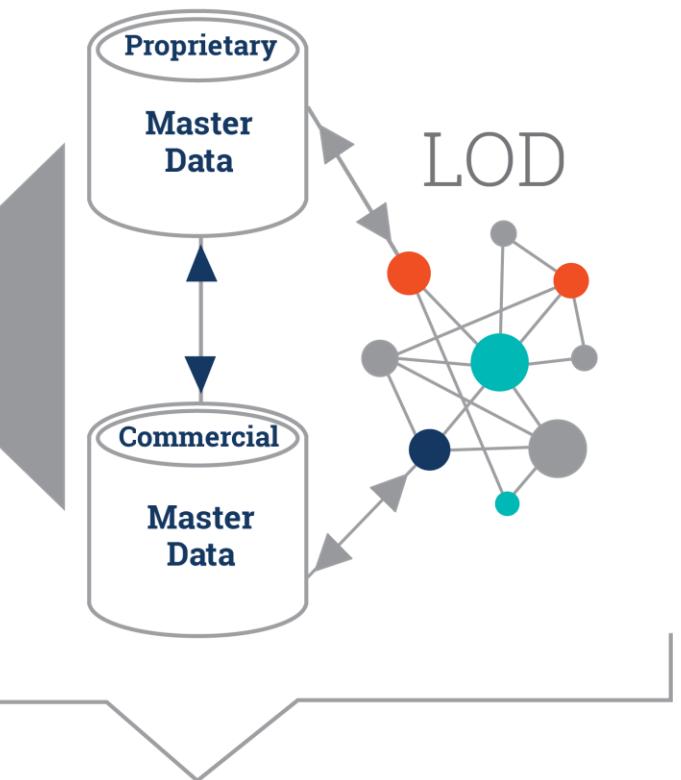
Automated Tagging
Content Publishing
Personalized Recommendation
Regulatory Compliance

Professional Services Consultancy

Data Integration
Master Data Management
Information Discovery
Open Data Publishing

ontotext

Our approach to Big Data

1. **Integrate data from many sources**

   – Build a Big Knowledge Graph that integrates relevant data from proprietary databases and taxonomies plus millions of facts of Linked Data
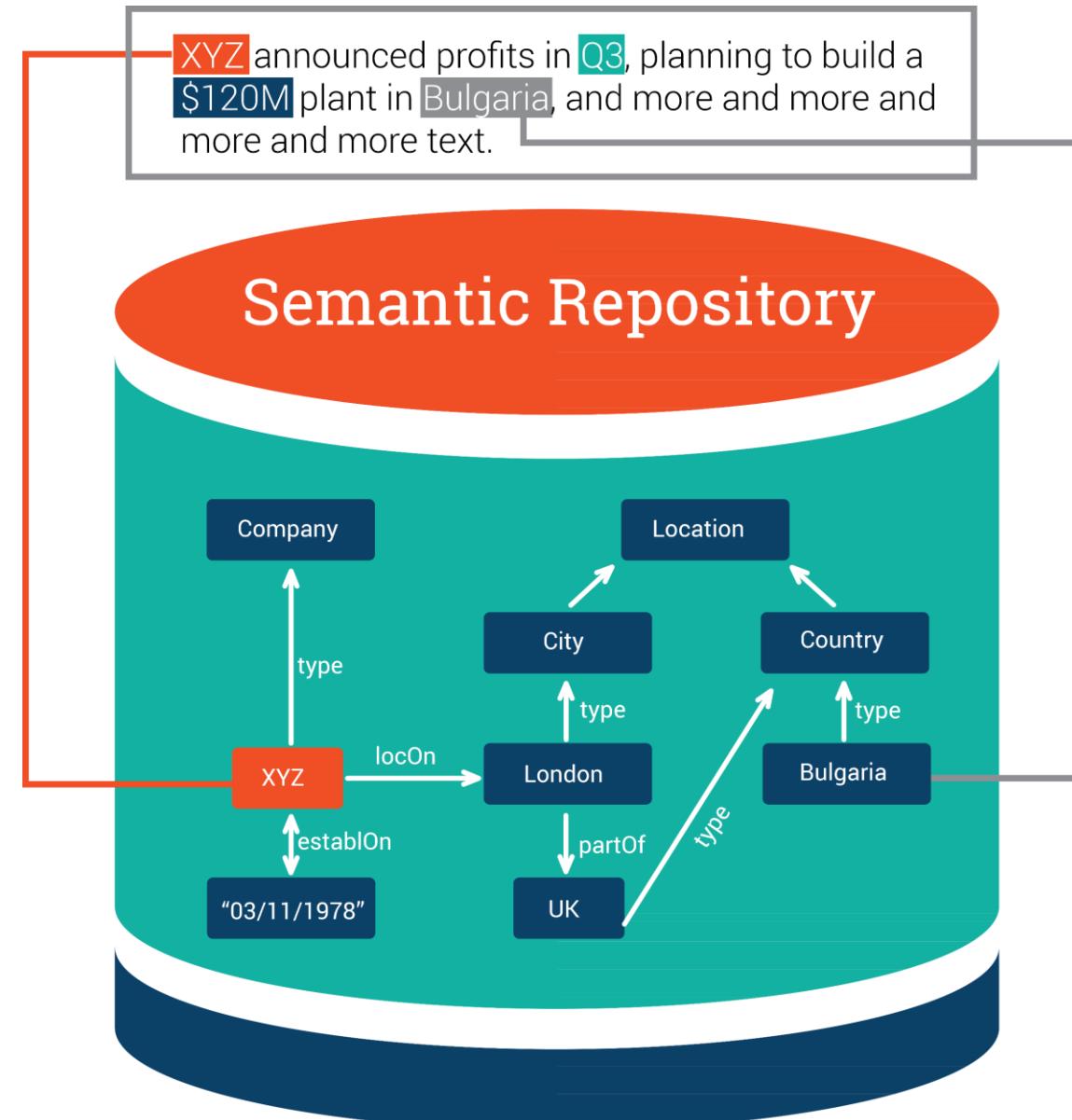
2. **Infer new facts and unveil relationships**

   – Performing reasoning across different data sources

3. **Interlink text and with big data**

   – Using text-mining to automatically discover references to concepts and entities

4. **Use graph database for metadata management, querying and search**

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 — type — XYZ
XYZ — locOn — London
XYZ — establOn — "03/11/1978"
City — type — London
London — partOf — UK
Location — City
Location — Country
Country — type — Bulgaria
UK — type — Country

ontotext
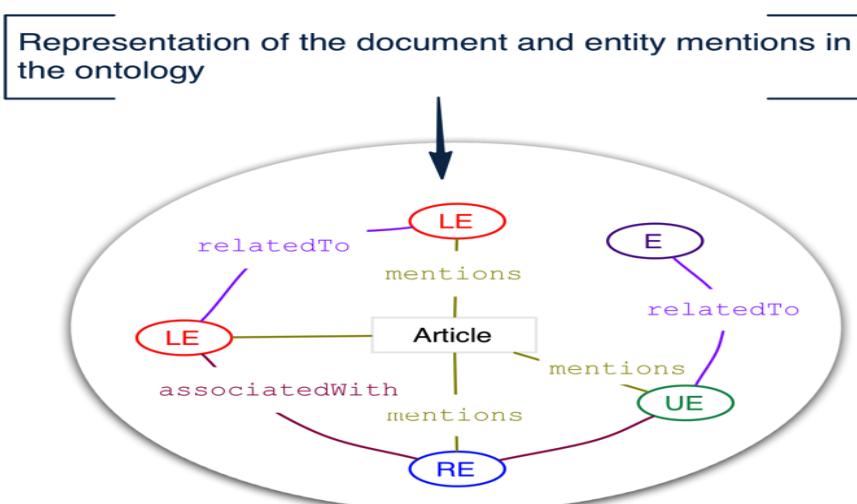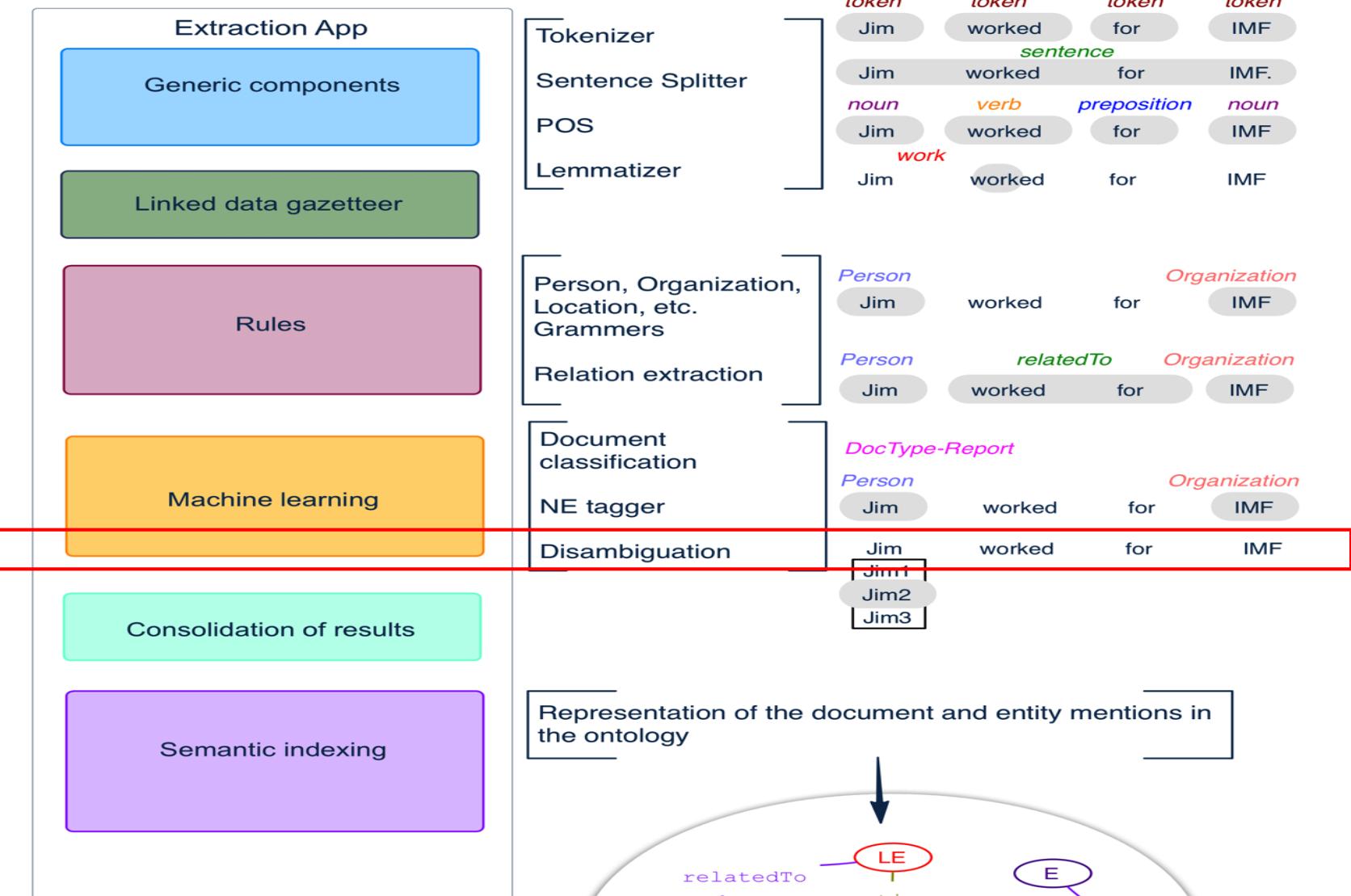
Analyzing Text

- **Full spectrum of NLP weaponry**
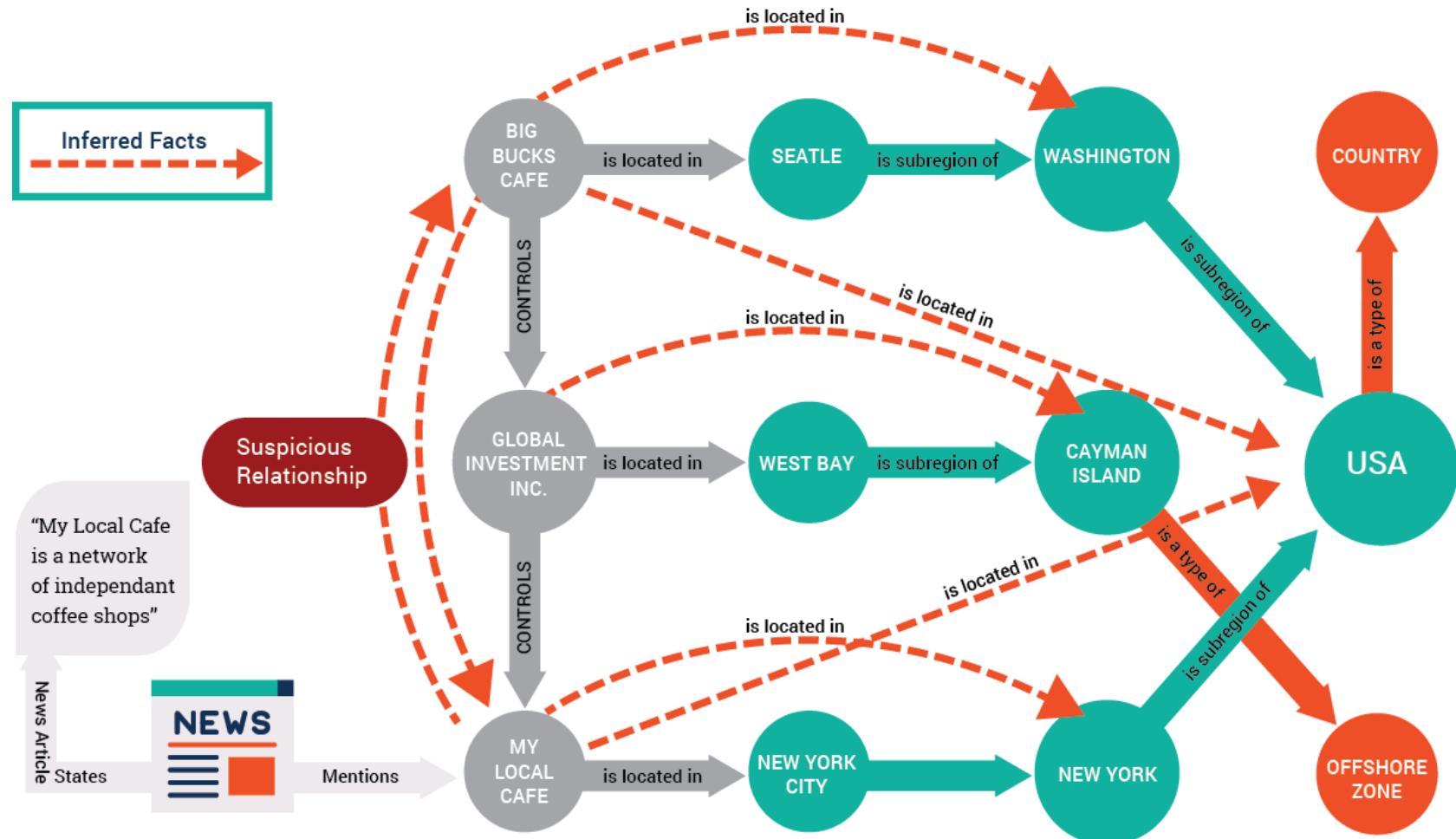
- **Semantic indexing**
  - Tag references with entity IDs
  - Generate semantic metadata descriptions of documents
  - Store metadata in GraphDB

Relation Discovery Case

- **Find suspicious relationships like:**
  - Company in USA
  - Controls another company in USA
  - Through a company in an off-shore zone

- **Show news relevant to these companies**

Linking News to Big Knowledge Graphs

- The DSP platform links text to knowledge graphs

- One can navigate from news to concepts, entities and topics, and from there to other news


Try it at **http://now.ontotext.com**

Semantic Media Monitoring

For each entity:

- popularity trends

- relevant news

- related entities

- knowledge graph information

Try it at http://now.ontotext.com

ontotext

FactForge: Open data and news about people and organizations

http://factforge.net

FactForge: Data Integration

DBpedia (the English version)	496M
Geonames (all geographic features on Earth)	150M
owl:sameAs links between DBpedia and Geonames	471K
Company registry data (**GLEI**)	3M
Panama Papers DB (#LinkedLeaks)	20M
Other datasets and ontologies: WordNet, WorldFacts, FIBO	
News metadata (2000 articles/day enriched by NOW)	473M
Total size (1152M explicit + 322M inferred statements)	**1 475M**

News Metadata
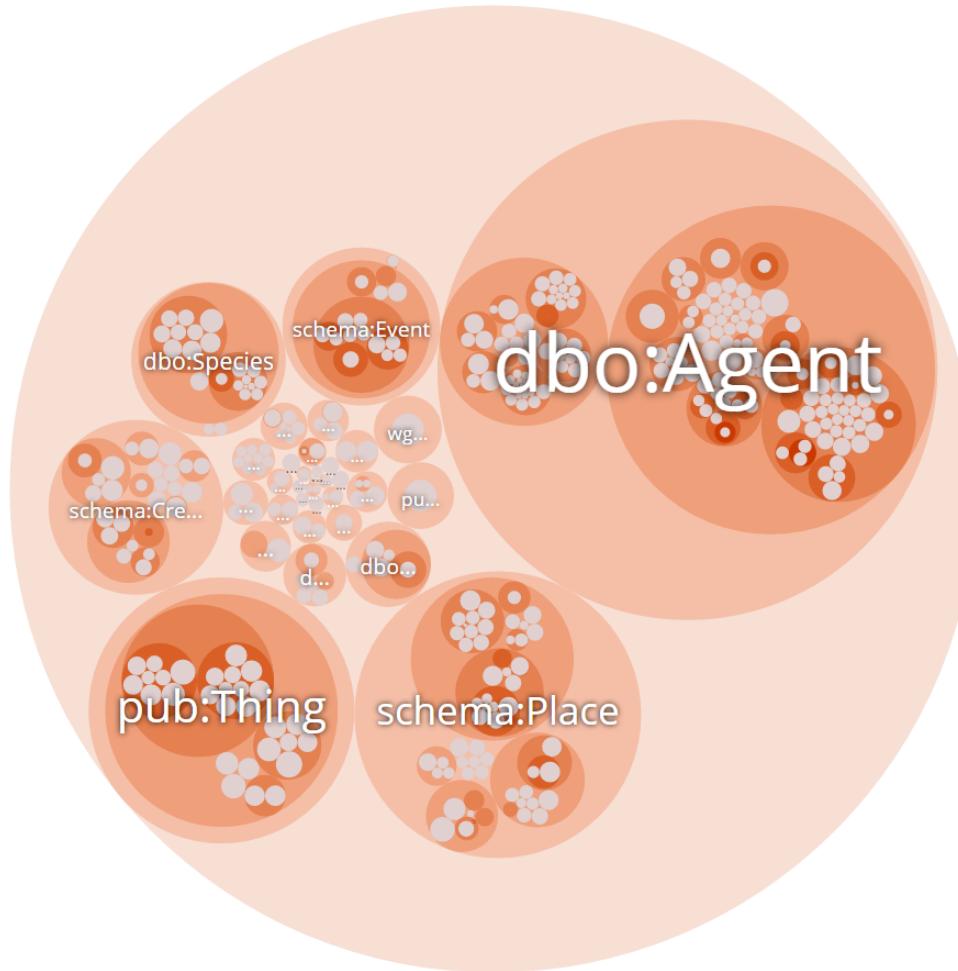
- **Metadata from Ontotext's Dynamic Semantic Publishing platform**
  - News stream from Google
  - Automatically generated as part of the NOW.ontotext.com semantic news showcase

- **News stream from Google since Feb 2015, about 60k news/month**
  - ~70 tags (annotations) per news article

- **Tags link text mentions of concepts to the knowledge graph**
  - Technically these are URIs for entities (people, organizations, locations, etc.) and key phrases

ontotext

News Metadata
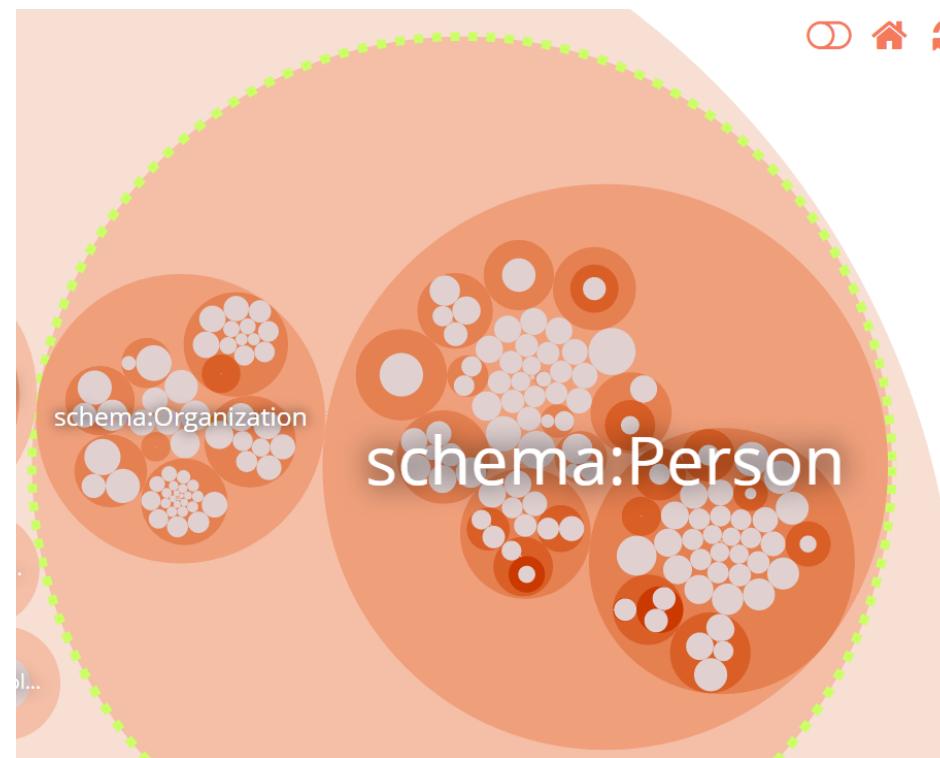
Category	Count
International	52 074
Science and Technology	23 201
Sports	20 714
Business	15 155
Lifestyle	11 684
	122 828

Mentions / entity type	Count
Keyphrase	2 589 676
Organization	1 276 441
Location	1 260 972
Person	1 248 784
Work	309 093
Event	258 388
RelationPersonRole	236 638
Species	180 946

ontotext

Class Hierarchy Map (by number of instances)

Left: The big picture
Right: dbo:Agent class (2.7M organizations and persons)

ontotext

Sample queries at http://factforge.net

- **F1: Big cities in Eastern Europe**

- **F2: Airports near London**

- **F3: People and organizations related to Google**

- **F4: Top-level industries by number of companies**

Available as Saved Queries at http://factforge.net/sparql

Note: Open Saved Queries with the folder icon in the upper-right corner

ontotext

Relationship Discovery Examples

Offshore control example

- **Query: Find companies, which control other companies in the same country, through company in an off-shore zone**

- **How it works:**

  - Establish control-relationship

  - Establish a company-country mapping

  - Establish an "off-shore criteria"

  - SPARQL it

ontotext

Off-shore company control example

```
SELECT *
FROM onto:disable-sameAs
WHERE {
        ?c1 fibo-fnd-rel-rel:controls ?c2 .
        ?c2 fibo-fnd-rel-rel:controls ?c3 .
        ?c1 ff-map:orgCountry ?c1_country .
        ?c2 ff-map:orgCountry ?c2_country .
        ?c3 ff-map:orgCountry ?c1_country .

        FILTER (?c1_country != ?c2_country)
        ?c2_country ff-map:hasOffshoreProvisions true .
}
```

ontotext

Media Monitoring Examples

Semantic Media Monitoring/Press-Clipping

- **We can trace references to a specific company in the news**

    – This is pretty much standard, however we can deal with syntactic variations in the names, because state of the art Named Entity Recognition technology is used

    – What's more important, we distinguish correctly in which mention "Paris" refers to which of the following: Paris (the capital of France), Paris in Texas, Paris Hilton or to Paris (the Greek hero)

- **We can trace and consolidate references to daughter companies**

- **We have comprehensive industry classification**

    – The one from DBPedia, but refined to accommodate identifier variations and specialization (e.g. company classified as dbr:Bank will also be considered classified as dbr:FinancialServices)

ontotext

Media Monitoring Queries

- **F5: Mentions in the news of an organization and its related entities**

- **F7: Most popular companies per industry, including children**

- **F8: Regional exposition of company – normalized**

ontotext

News Popularity Ranking: Automotive

Rank	Company	News #
1	General Motors	2722
2	Tesla Motors	2346
3	Volkswagen	2299
4	Ford Motor Company	1934
5	Toyota	1325
6	Chevrolet	1264
7	Chrysler	1054
8	Fiat Chrysler Automobiles	1011
9	Audi AG	972
10	Honda	717

Rank	Company incl. mentions of child companies	News #
1	General Motors	4620
2	Volkswagen Group	3999
3	Fiat Chrysler Automobiles	2658
4	Tesla Motors	2370
5	Ford Motor Company	2125
6	Toyota	1656
7	Renault-Nissan Alliance	1332
8	Honda	864
9	BMW	715
10	Takata Corporation	547

ontotext

SPB News Monitoring Variant

Motivation for the SPB News Monitoring

- **The reference data in SPB is too small**

  – This is for purpose: to match the needs of the small scale factors

  – However, this way it doesn't match the size of the reference data used in such applications

- **Data generation converts graph centrality into news popularity**

  – SPB 2.0 implements a naive approach towards "popular entities"

  – The data generator takes RDFRank in DBPedia and directly converts it into news popularity. In fact those are independent and often different

    ▪ There are entities which are quite central in DBPedia, without being mentioned often in news

    ▪ SAP is a good example

ontotext

Motivation for the SPB News Monitoring (II)

- **Check SPB performance against FactForge, where:**

    - We have much bigger reference data 22M vs. 700M

    - Real news stream, real story-lines

    - Real tags linking mentions in the text to their true matches in the

- **Questions**

    - Overall query performance comparison

    - Check the real distributions of:

        - Number of entity mentions per article

        - How many entities are really mentioned

        - Popular entities

ontotext

SPB256 vs. FF Data Comparison

	SPB256	FactForge	Comment
News articles	8 821 539	406 672	
News/day over 2 years	12 390	571	SPB volumes are relevant only for the biggest news aggregators
Tags	75 961 660	25 848 188	
Tags / news	8.61	63.56	
Mentions of known entities	75 961 660	16 805 697	
Mentions / news	8.61	41.32	Much bigger number of entities and concepts and entities mentioned in a article
Distinct entities mentioned	1 792 805	536 898	Much smaller fraction of the entities are actually referred
Total entities - Ref. Data.	1 822 627	3 606 581	

ontotext

Queries implemented in SPB News Monitoring

- **Q1: All info about the recent creative works tagged with thing T**

- **Q2: Retrieve details about a given creative work**

- **Q3: Retrieve a list of blog post and news, matching some criteria**

- **Q4: Creative works tagged with thing T, ordered by creation date**

- **Q5: The most popular topics, under some CW criteria**

- **Q6: Creative works mentioning location within a geospatial rectangle**

- **Q7: Creative works from specific period, ordered**

- **Q8: Creative works by word contained in its title (FTS)**

- **Q10: Retrieve CWs that mention locations in the same province as the specified one**

- **Q11: Retrieve CWs that are tagged with entities, related to a specific popular entity**

ontotext

SPB256 vs. FF Query Performance

	SPB256		SPB-FF		Diff %		Non-empty Results		Avg. adj. empty		Diff %
	Avg.	Max	Avg.	Max	Avg.	Max	SPB256	SPB-FF	SPB256	SPB-FF	
Q1	33	6 889	15	1 095	120	529	86.39%	57.47%	38	26	46
Q2	5	398	5	840	0	-111	99.97%	52.54%	5	10	-90
Q3	68	7 718	7	792	871	874	84.16%	7.19%	81	97	-20
Q4	52	1 045	15	1 960	247	-88	81.95%	57.49%	63	26	143
Q5	96	12 338	86	4 243	12	191	59.50%	18.39%	161	468	-190
Q6	73	11 530	1 721	17 109	-2 258	-48	63.10%	30.37%	116	5 667	-4 798
Q7	6	413	7	15 392	-17	-3 627	99.99%	29.17%	6	24	-300
Q8	45	11 504	11	2 149	309	435	100.00%	83.58%	45	13	242
Q9	267	12 300	0	0			99.97%		267		
Q10	590	12 486	22	2 908	2 582	329	53.87%	5.48%	1 095	401	173
Q11	101	7 510	856	20 473	-748	-173	98.54%	96.92%	102	883	-762
Q12	141	11 604	0	0			99.96%				
Avg. QPS	50.09		29.05				86%	44%			

ontotext

SPB256 vs. FF Query Performance Analysis

- **This is still work in progress**

- **We see comparable query performance**
  - Except couple of queries that would require further tuning to make them comparable
  - Results are comparable for GraphDB … This may not be the case for other engines

- **We expected the big differences to appear on Queries Q9-Q11**
  - Q9 still has to be implemented in SPB-FF
  - Q10 (Regional news): SPB-FF is faster, because there are fewer news to be returned
  - Q11 (News on related entities): it is 8x slower at SPB-FF, because there is much bigger number of related entities there

ontotext

SPB News Monitoring: Next Steps

- **Adapt all the queries from the Basic Interactive query mix**

- **Implement the update transactions to allow for benchmarking Aggregation and Editorial workloads combined**

- **Consider proposing changes to the Semantic Publishing Benchmark:**

  - Data Generator: adjusted data distributions

  - New workloads and query mixes matching media monitoring use case

ontotext

Thank you!

Experience the technology with NOW: Semantic News Portal

http://now.ontotext.com

and play with open data at

http://factforge.net

ontotext