FactForge
Data Service and the Value of Inferred Knowledge

Mariana Damova, PhD

European Open Data Forum
June 2012
Ontotext

- Top-5 provider of core Semantic Technology
- Established in year 2000; offices in Bulgaria, UK, USA
- Active both in research and commercial projects

**360° semantic technology – unique portfolio:**

- **Semantic Databases**: high-performance RDF DBMS, scalable reasoning
- **Semantic Search**: text-mining (IE), metadata generation, Information Retrieval (IR)
- **Web Mining**: focused crawling, screen scraping, data fusion
- **Linked Data Management and Data Integration**

*Good recognition in the SemTech community*

- Ontotext pages are ranked #1 for “semantic annotation” and “semantic repository” at GYM, #3 for “linked data management” at Google

*Several joint ventures and subsidiaries*

- Innovantage: leading online recruitment intelligence provider in UK
Ontotext Clients (selected)

**British Broadcasting Corporation (BBC)**
- Run its World Cup 2010 sites on top of OWLIM
- Since Mar’12 BBC Sports and 2012 Olympics sections are driven by OWLIM and a Concept Extraction service developed by Ontotext

**Press Association (UK)**
- Analysis of Sports news
- Concept extraction
- Linked data generation

**Top-3 USA media** (not allowed to name)

**The National Archives (UK)** contracted Ontotext to implement semantic KB and semantic search for the Government Web Archive

**British Museum (UK)** Ontotext leads the development of Phase 3 of ResearchSpace project on collaborative research in cultural heritage; British Museum’s public SPARQL end-point is powered by OWLIM

**de Bibliothek (Holland)** aggregation of data from 150 library databases
Linked Open Data is maturing

LOD cloud grows by billions of triples yearly

Technologies and guidelines about

how to produce linked data fast

how to assure their quality

how to provide vertical oriented data services

LOD2, LATC, baseKB
This talk is about reasoning and coping with diversity of the data on the web of data.
Outline

• FactForge (beta)
• Reference Layer
• Access Modes
• Querying
  – Airports around London
  – US city – a subject of a Novel
  – US city – contact Information
• Challenges
• Conclusion
FactForge (beta)

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

**Repository overview**

- Engine: OWLIM SE 5.0
- Inference ruleset: factforge.pr (more)
- Number of statements: 3,126,004,378
- Number of expl. statements: 1,860,804,539
- Number of entities: 538,242,324
- Number of Literals: 269,814,590
- Number of statements: 3,126,004,378
- Number of expl. statements: 1,860,804,539
- Number of entities: 538,242,324
- Number of Literals: 269,814,590

the largest body of heterogeneous general knowledge on which inference has been performed

- powered by **OWLIM 5.0**
- supporting **SPARQL 1.1**
Datasets

**REASON-ABLE VIEW of LOD datasets**

- Number of explicit statements: 1,796,673,630
- Implicit statements: 1,3
-Retrievable statements: 14,928,925,039

- DBpedia 3.7
- Freebase
- Wordnet 3.0
- Geonames
- NY Times
- CIA FactBook
- Lexvo
- MusicBrainz
- Lingvoj

Materialization is performed with respect to the semantics of OWL-Horst optimized
Reference Layer

PROTON – light weight upper level ontology
~500 classes, ~150 properties
http://www.ontotext.com/proton-ontology

Linking at schema level:
(1) using rdfs:subClassOf and rdfs:subPropertyOf statements;
(2) using OWL expressions where there is a difference in the conceptualization
(3) using inference rules if additional individuals are necessary in the repository to support the mapping
Access modes

**RDF Search** - retrieve ranked list of URIs related to literals, which contain specific keywords

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

Repository overview

- **Engine**: OWLIM SE 5.0
- **Information**: [http://ontotext.com/factforge](http://ontotext.com/factforge)
- **Inference ruleset**: factforge-pie (more)
- **Number of statements**: 3,066,787,172 (Dataset statistics)
- **Number of expl. statements**: 1,756,673,630
- **Number of entities**: 507,683,596
- **Number of uris**: 203,294,583
- **Number of literals**: 266,753,728
- **Number of nodes**: 61,032,945

Results for "Copenhagen Denmark" (10 of 10)

- Copenhagen (Denmark) 20120
- Copenhagen (Denmark) 20120
- Copenhagen (Denmark) 20120
- Copenhagen (Denmark) 20120
- Copenhagen (Denmark) 20120
**Access modes (condt)**

**Exploration** - traversing the data, one resource at a time

---

Copenhagen (Denmark) RDF Park

Copenhagen is the capital and largest city of Denmark, with an urban population of 1,199,224 (as of 1 January 2011) and a metropolitan population of 1,918,052 (as of 1 April 2011)....

Source: http://dbpedia.org/resource/Copenhagen

Same as: dbp:open.copenhagen, dbpedia:Denmark, wsw:geonames.org/6210425, w:tte:traversing the data, one resource at a time
### Access modes (condt)

**Exploration** - traversing the data, one resource at a time, inspecting inferred knowledge

- **locatedIn** - Denmark, Northern Europe
- **Geonames types/FeatureCodes** P.PPL
- **parentFeature** - Denmark, Europe

---

**Example Resource:**

- **type**: City
- **location**: Copenhagen, Denmark
- **population**: 1.15M

---

**Geonames Data:**

- **id**: 2619424
- **url**: [Geonames](http://www.geonames.org/2619424)

---

**Image:**

- A screenshot of the resource in the Geonames database with details such as location and population.
Access modes (condt)

**Exploration** - traversing the data, one resource at a time, inspecting inferred knowledge

- locatedIn - Europe
- subRegionOf - Europe
- hasContactInfo - website via Freebase
- containsLocation

![FactForge](image)

Statements in which the resource exists as a subject.

<table>
<thead>
<tr>
<th>Predicate</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>rdf:type</td>
<td>pfo:Location, psy:Entity, pto:Location, pto:Object, rdfs:Resource</td>
</tr>
<tr>
<td>pto:locatedIn</td>
<td>dbpedia Europe, <a href="http://sws.geonames.org/6295630/">http://sws.geonames.org/6295630/</a></td>
</tr>
<tr>
<td>psy:mainLabel</td>
<td>Northern Europe@en</td>
</tr>
<tr>
<td>pto:hasContactInfo</td>
<td>fb:m:09xstj</td>
</tr>
<tr>
<td>pto:subRegionOf</td>
<td>dbpedia Europe, <a href="http://sws.geonames.org/6295630/">http://sws.geonames.org/6295630/</a></td>
</tr>
</tbody>
</table>
| pawr:containsLocation      | dbpedia Hammers, dbpedia/Sakskebing, dbpedia/Vøagå, dbpedia/Sivndal, dbpedia/Notby, dbpedia/Froya,_Norway, dbpedia/Jalsbratad, dbpedia/Hulikvall_County, dbpedia/Sand,_Norway, dbpedia/Milddi, dbpedia/Ostreid, dbpedia/Edsberg,_Sollentuna, dbpedia/Sandurn, dbpedia/Arkila, dbpedia/Masmo, dbpedia/Koylangane, dbpedia/Heyenhall_(station), dbpedia/Bernstoffe_aben, dbpedia/Stångårna, dbpedia/Stångårts, dbpedia/Allhelgonakyrkan, dbpedia/Montebello_(station), dbpedia/Overskölen, dbpedia/Galbyen, dbpedia/Stororden,
Access modes (condt)

SPARQL endpoint

European Data Forum
Access modes (condt)

RelFinder
Querying

Using LOD concepts

```sql
SELECT *
WHERE {
  ?Person dbp-ont:birthPlace ?BirthPlace ;
  rdf:type dbp-ont:Politician ;
  ?BirthPlace geo-ont:parentFeature dbpedia:Germany .
}
```

Using the intermediary layer

```sql
SELECT *
WHERE {
  ?Person prot:birthPlace ?BirthPlace ;
  rdf:type prot:Politician ;
  ?BirthPlace prot:subRegionOf dbpedia:Germany .
}
```
Find Airports near London

**SPARQL Query**

<table>
<thead>
<tr>
<th>airport</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpedia.London_Heathrow_Airport</td>
<td>London Heathrow Airport@en</td>
</tr>
<tr>
<td>dbpedia.London_City_Airport</td>
<td>London City Airport</td>
</tr>
<tr>
<td>dbpedia.raf_Northolt</td>
<td>Royal Air Force Northolt 90px@en</td>
</tr>
<tr>
<td>dbpedia.Antwerp_International_Airport</td>
<td>Antwerp International Airport@en</td>
</tr>
<tr>
<td>dbpedia.Croydon_Airport</td>
<td>Croydon Airport@en</td>
</tr>
<tr>
<td>dbpedia.London_Biggin_Hill_Airport</td>
<td>London Biggin Hill Airport@en</td>
</tr>
<tr>
<td>dbpedia.Bisham_Airfield</td>
<td>Bisham Airfield@en</td>
</tr>
<tr>
<td>dbpedia.London_Heliport</td>
<td>London Heliport@en</td>
</tr>
<tr>
<td>dbpedia.Heston_Aerodrome</td>
<td>Heston Aerodrome@en</td>
</tr>
<tr>
<td>dbpedia.Stapleford_Aerodrome</td>
<td>Stapleford Aerodrome@en</td>
</tr>
<tr>
<td>dbpedia.North_Weald_Airfield</td>
<td>North Weald Airfield@en</td>
</tr>
<tr>
<td>dbpedia.Staig_Lane_Aerodrome</td>
<td>Staig Lane Aerodrome@en</td>
</tr>
</tbody>
</table>

Standard LOD vs. PROTON query
13 vs. 20 results
DBpedia vs. DBpedia and Geonames

European Data Forum
June 2012 #17
Find airports near London - Results comparison

### SPARQL Query

Results for your query (13) - Edit query

<table>
<thead>
<tr>
<th>airport</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpedia.London_Heathrow_Airport</td>
<td>London Heathrow Airport@en</td>
</tr>
<tr>
<td>dbpedia.London_City_Airport</td>
<td>London City Airport</td>
</tr>
<tr>
<td>dbpedia.RAF_Northolt</td>
<td>Royal Air Force Northolt 90px@en</td>
</tr>
<tr>
<td>dbpedia.Antwerp_International_Airport</td>
<td>Antwerp International Airport@en</td>
</tr>
<tr>
<td>dbpedia.Croydon_Airport</td>
<td>Croydon Airport@en</td>
</tr>
<tr>
<td>dbpedia.London_Biggin_Hill_Airport</td>
<td>London Biggin Hill Airport@en</td>
</tr>
<tr>
<td>dbpedia.Elstree_Airfield</td>
<td>Elstree Airfield@en</td>
</tr>
<tr>
<td>dbpedia.London_Heliport</td>
<td>London Heliport@en</td>
</tr>
<tr>
<td>dbpedia.Heston_Aerodrome</td>
<td>Heston Aerodrome@en</td>
</tr>
<tr>
<td>dbpedia.Stapleford_Aerodrome</td>
<td>Stapleford Aerodrome@en</td>
</tr>
<tr>
<td>dbpedia.North_Wkaid_Airfield</td>
<td>North Wkaid Airfield@en</td>
</tr>
<tr>
<td>dbpedia.Stag_Lane_Aerodrome</td>
<td>Stag Lane Aerodrome@en</td>
</tr>
</tbody>
</table>

### SPARQL Query

Results for your query (20) - Edit query

<table>
<thead>
<tr>
<th>airport</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpedia.London_Heathrow_Airport</td>
<td>London Heathrow Airport@en</td>
</tr>
<tr>
<td>dbpedia.London_City_Airport</td>
<td>London City Airport</td>
</tr>
<tr>
<td>dbpedia.RAF_Northolt</td>
<td>Royal Air Force Northolt 90px@en</td>
</tr>
<tr>
<td>dbpedia.Antwerp_International_Airport</td>
<td>Antwerp International Airport@en</td>
</tr>
<tr>
<td>dbpedia.Croydon_Airport</td>
<td>Croydon Airport@en</td>
</tr>
<tr>
<td>dbpedia.London_Biggin_Hill_Airport</td>
<td>London Biggin Hill Airport@en</td>
</tr>
<tr>
<td>dbpedia.Elstree_Airfield</td>
<td>Elstree Airfield@en</td>
</tr>
<tr>
<td>dbpedia.London_Heliport</td>
<td>London Heliport@en</td>
</tr>
<tr>
<td>dbpedia.Heston_Aerodrome</td>
<td>Heston Aerodrome@en</td>
</tr>
<tr>
<td>dbpedia.Stapleford_Aerodrome</td>
<td>Stapleford Aerodrome@en</td>
</tr>
<tr>
<td>dbpedia.North_Wkaid_Airfield</td>
<td>North Wkaid Airfield@en</td>
</tr>
<tr>
<td>dbpedia.Stag_Lane_Aerodrome</td>
<td>Stag Lane Aerodrome@en</td>
</tr>
<tr>
<td><a href="http://sws.geonames.org/0301624/">http://sws.geonames.org/0301624/</a></td>
<td>Northolt</td>
</tr>
<tr>
<td>dbpedia.Stag_Lane_Aerodrome</td>
<td>Stag Lane Aerodrome@en</td>
</tr>
<tr>
<td><a href="http://sws.geonames.org/6296597/">http://sws.geonames.org/6296597/</a></td>
<td>Biggin Hill</td>
</tr>
<tr>
<td><a href="http://sws.geonames.org/6691396/">http://sws.geonames.org/6691396/</a></td>
<td>London Heathrow Terminal 1</td>
</tr>
<tr>
<td><a href="http://sws.geonames.org/6691397/">http://sws.geonames.org/6691397/</a></td>
<td>London Heathrow Terminal 2</td>
</tr>
<tr>
<td><a href="http://sws.geonames.org/6691398/">http://sws.geonames.org/6691398/</a></td>
<td>London Heathrow Terminal 3</td>
</tr>
<tr>
<td><a href="http://sws.geonames.org/6691399/">http://sws.geonames.org/6691399/</a></td>
<td>London Heathrow Terminal 4</td>
</tr>
</tbody>
</table>

Using Geospatial index of OWLIM
City – a subject of a science fiction author
OWLIM 5.0 and SPARQL 1.1

Exemplary queries:

GROUP BY, min
- Minimal and maximal population counts of European countries

Federated Query between FactForge and LinkedLifeData
- Drugs that cure the disease from which died Alexandre Graham Bell

Literal index over dates
- World governors in office between 1980 and 2005

Literal index over digits
- European countries with population above 20 MLN

Geospatial index
- Show the distance from London of airports located at most 50 miles away from it
Challenges and usage

• Clean data
  – Clean up input data

• At model level
  – Contradiction detection
  – Consistency checking

• Curation and upgrading methodology

FactForge has been used as data layer infrastructure in FP7 projects, like RENDER. FactForge has been used in tasks of linked data generation from unstructured data, metadata enrichment of structured data, providing linkages to the entire LOD cloud, for example The National Archive.
Acknowledgements

Partial funding

Colleagues
Atanas Kiryakov, CEO of Ontotext
Zdravko Tashev, Ontotext
Ivan Peikov, Ontotext
Rouslan Velkov, Ontotext
Kiril Simov, Ontotext
Barry Bishop, Ontotext
Barry Norton, Ontotext
Marin Dimitrov, Ontotext
Alex Simov, Ontotext
Jordan Dichev, Ontotext
Konstantin Penchev, Ontotext

Links
http://ff-dev.ontotext.com
http://www.ontotext.com/owlim
http://www.ontotext.com/factforge

Email:
info@factforge.net
Thank you for your attention!

mariana.damova@ontotext.com