



making sense of text and data

**Analytics on Big Knowledge  
Graphs Deliver Entity Awareness  
and Help Data Linking**

Semantics 2018, Vienna

# Presentation Outline

- **Ontotext Introduction**
- Technology and Portfolio
- Cognitive Analytics Meet Big Knowledge Graphs
- Big Company Data: Knowing, Matching and Cleaning
- Product Roadmap



ontotext



## Vision

- **Global business information will be key for competitiveness tomorrow**
- **Adequate business decisions require global information!**
  - ✓ Analytics cannot deliver deep market/business insights based only on proprietary data
  - ✓ Broader context and signals are needed
- **Merging data requires concept and entity awareness**
  - ✓ Entity matching across databases requires rich knowledge about the entity
  - ✓ Entity recognition in text requires even more context
- **Ontotext makes this possible**

# Mission

We help enterprises to **identify meaning** across:

- **Diverse databases & unstructured data**

We combine:

- **Proprietary & Global data**
- **Graph databases & Text mining**
- **Symbolic reasoning & Machine learning**



# History and Essential Facts

- **Started in year 2000 as Semantic Web pioneer**
  - ✓ Part of Sirma Group: ~400 persons, listed at Sofia Stock Exchange
  - ✓ Got spun-off and took VC investment in 2008
- **R&D Center in Sofia, 80% sales in USA and UK**
  - ✓ Over 400 person-years invested in R&D
  - ✓ Multiple innovation awards: Washington Post, BBC, FT, ...
- **Member of multiple industry bodies**
  - ✓ W3C, EDMC, ODI, LDBC, STI, DBPedia Foundation



# Best known for GraphDB

*“Despite all of this attention the market is dominated by Neo4J and **Ontotext (GraphDB)**, which are graph and RDF database providers respectively. These are the longest established vendors in this space (both founded in 2000) so they have a longevity and experience that other suppliers cannot yet match. How long this will remain the case remains to be seen.”*

Bloor Group report

**Graph Databases, April 2015**

<http://www.bloorresearch.com/technology/graph-databases/>



# Fancy Stuff and Heavy Lifting

- We do advanced analytics:  
We predicted BREXIT
  - ✓ 14 Jun 2016 whitepaper:  
**#BRExit Twitter Analysis: More Twitter Users Want to Split with EU and Support #Brexit**  
<https://ontotext.com/white-paper-brexit-twitter-analysis/>
- But most of the time we do the heavy lifting of data integration and information extraction
  - ✓ Enabling data scientists can do fancy things

The Economist @TheEconomist · 14 Jun 2016  
#Brexit could be a "disaster for science" econ.st/1Xn3hIR

The EU provides 10% of research funding in British universities

Total research funding from EU, 2006-15, £bn

Source: "Examining Implications of Brexit for the UK Research Base", by D Hook and M. Szomszor, May 2016

25 346 163

Ontotext @ontotext Following

Replying to @TheEconomist

Hi, we did a semantic analysis of over 1.5mil. tweets about #Brexit, here it is what it shows!

More Twitter Users Want to Split with EU and Support #Bre...  
Find out if the british want to split with EU according to twitter analysis from several hastags related to the subject. Find relationships and polarity.  
ontotext.com

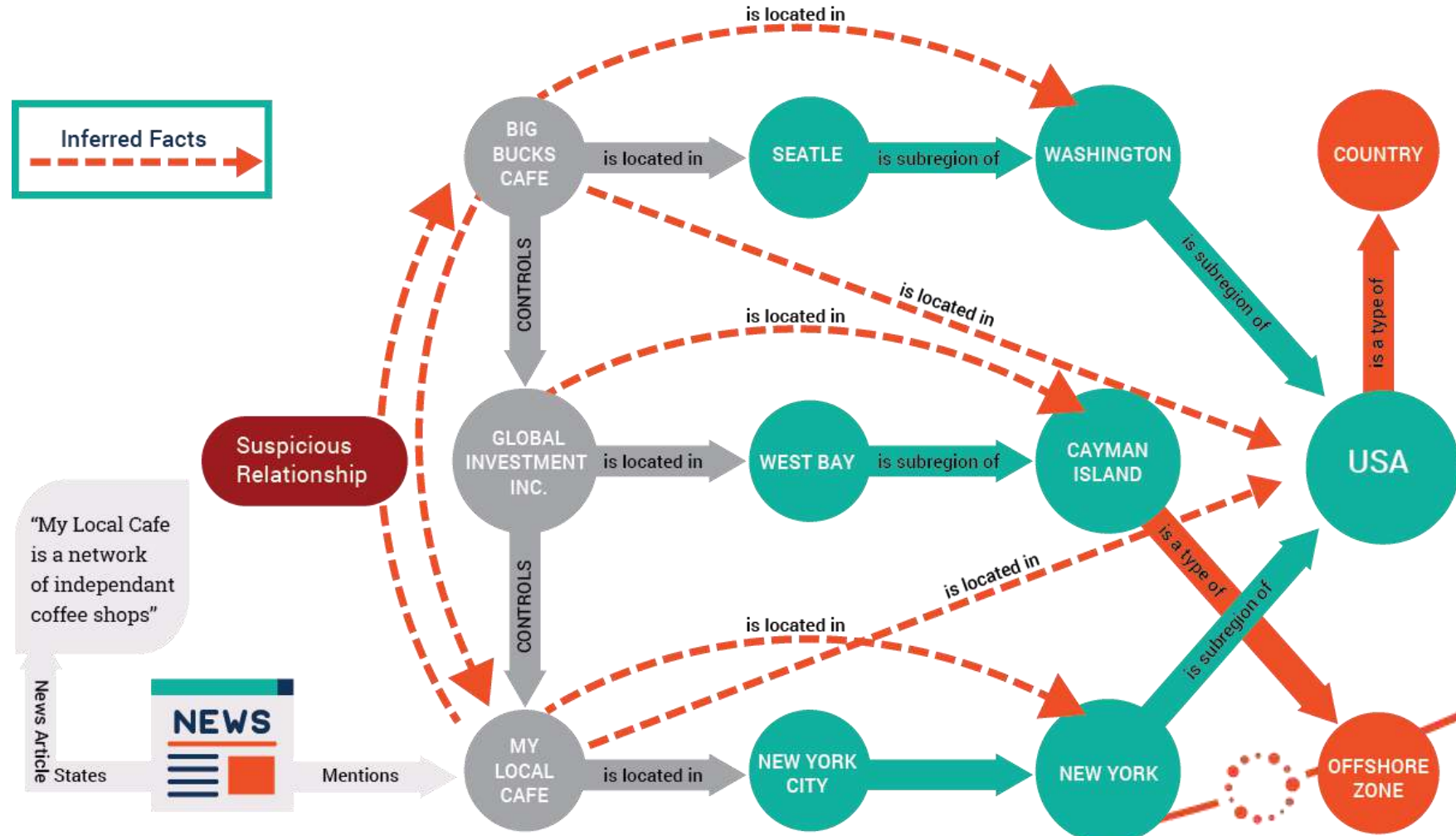
4:32 PM - 14 Jun 2016

# Discovery in Knowledge Graphs

## ○ Find suspicious patterns like:

- ✓ Company in USA
- ✓ Controls another company in USA
- ✓ Through a company in an off-shore zone

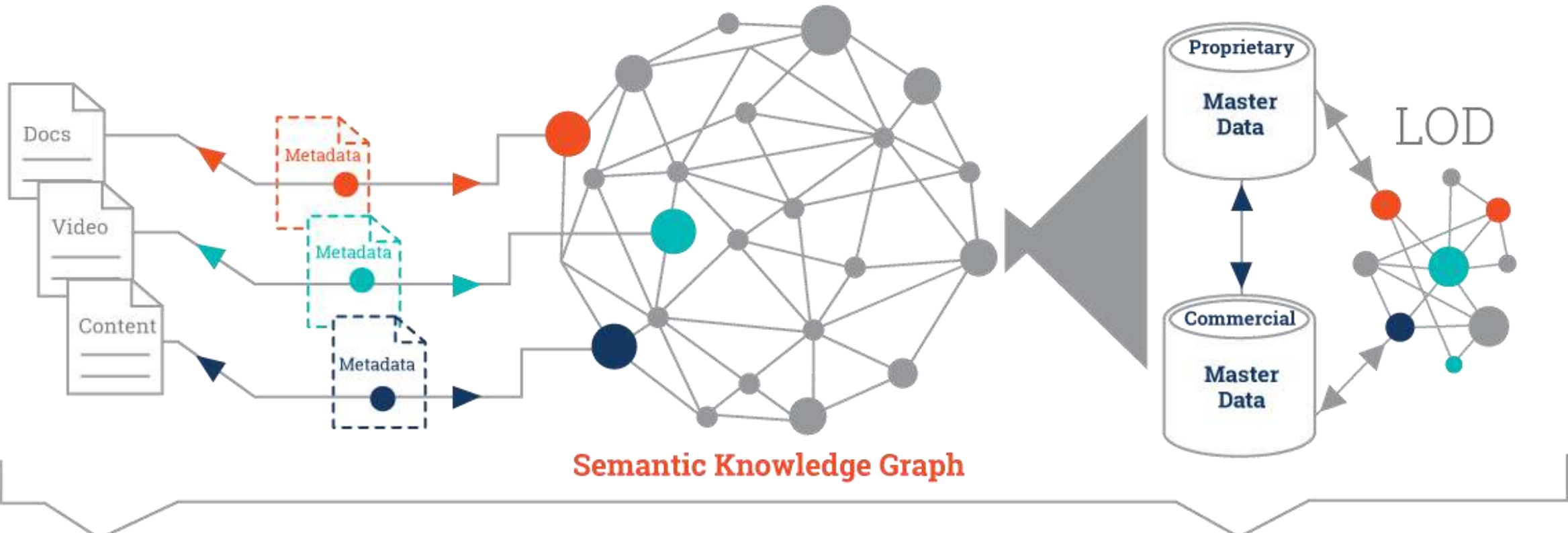
## ○ Show news relevant to these companies





## Content Analytics & Exploration Platform

## GraphDB



## Semantic Knowledge Graph

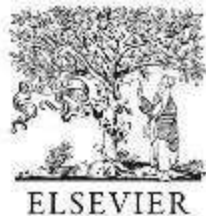
Automated Tagging  
Content Publishing  
Personalized Recommendation  
Regulatory Compliance

Professional Services  
Consultancy

Data Integration  
Master Data Management  
Information Discovery  
Open Data Publishing

# Technology Excellence Delivered

- **Unique technology mix:** GraphDB™ engine + Text mining
- **Robust technology:** powers BBC.CO.UK/SPORT and FT.COM
- **We serve the most knowledge intensive enterprises**



ontotext

# Presentation Outline

- Ontotext Introduction
- **Technology and Portfolio**
- Cognitive Analytics Meet Big Knowledge Graphs
- Big Company Data: Knowing, Matching and Cleaning
- Product Roadmap



ontotext

# Linking Text to Big Knowledge Graphs

## 1. Integrate relevant structured data

- ✓ Build a Big Knowledge Graph from proprietary databases and taxonomies combined with Linked Open Data

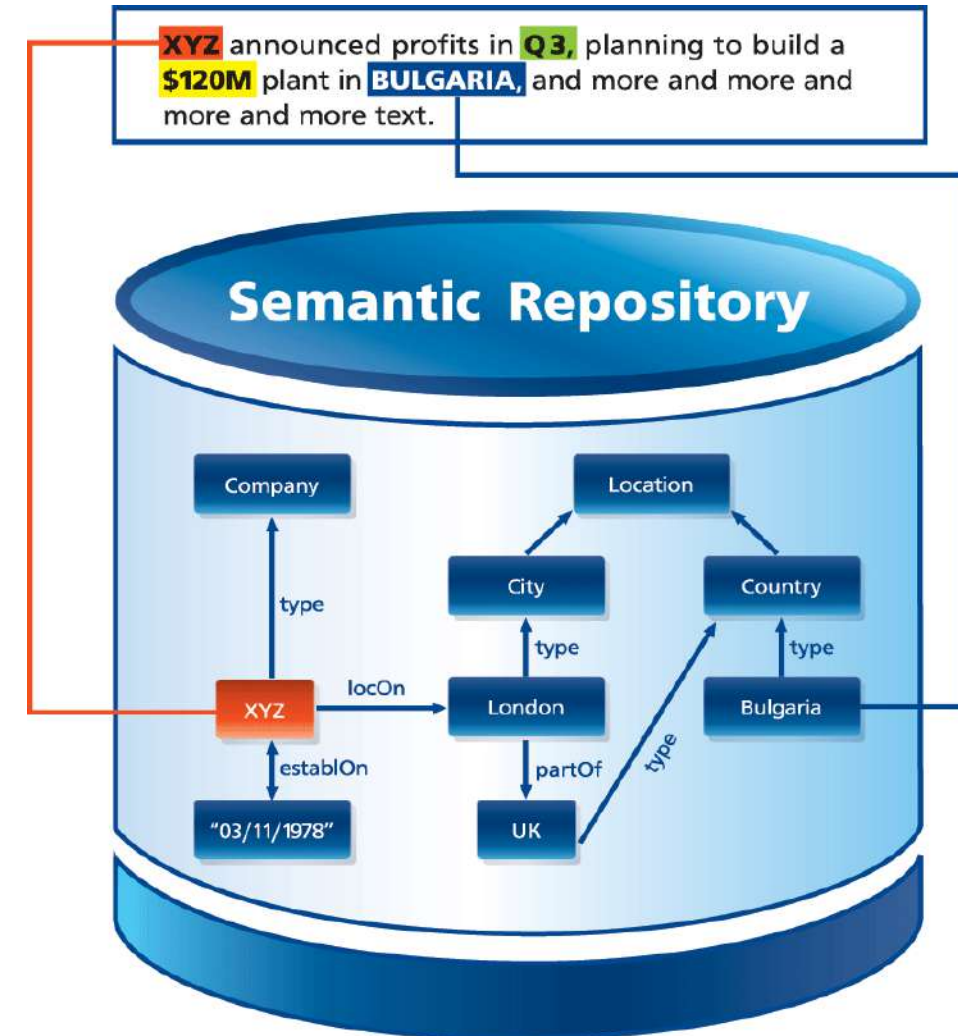
## 2. Infer new facts and unveil relationships

- ✓ Performing reasoning across data from different sources

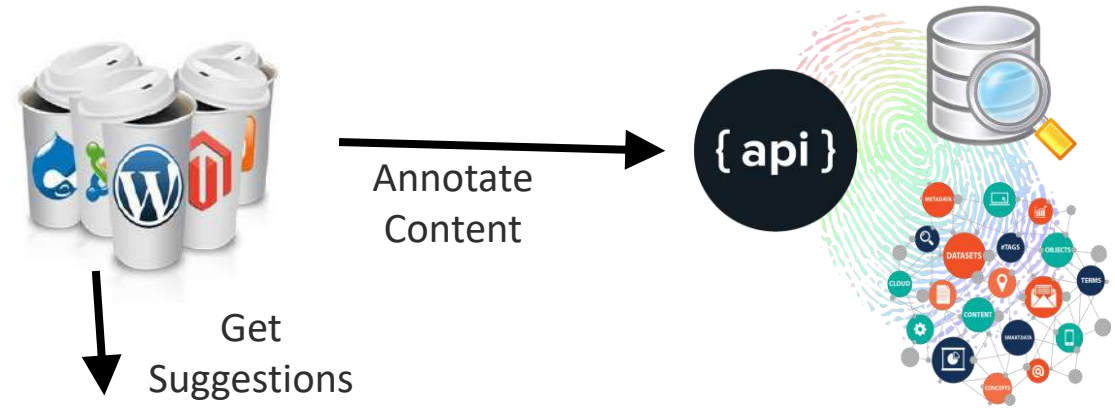
## 3. Link text mentions to the Knowledge Graph

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

## 4. Hybrid Queries and Search in GraphDB



# Text Analytics: Semantic Disambiguation



**Suggestions**

Apple CEO Tim Cook was at a conference with the CEO of Samsung. Tim explained how smart phones are changing the consumer electronics market.

**Entity Detection from Vocabulary**

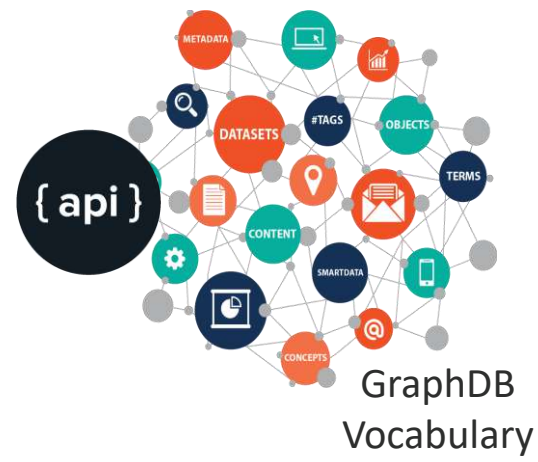
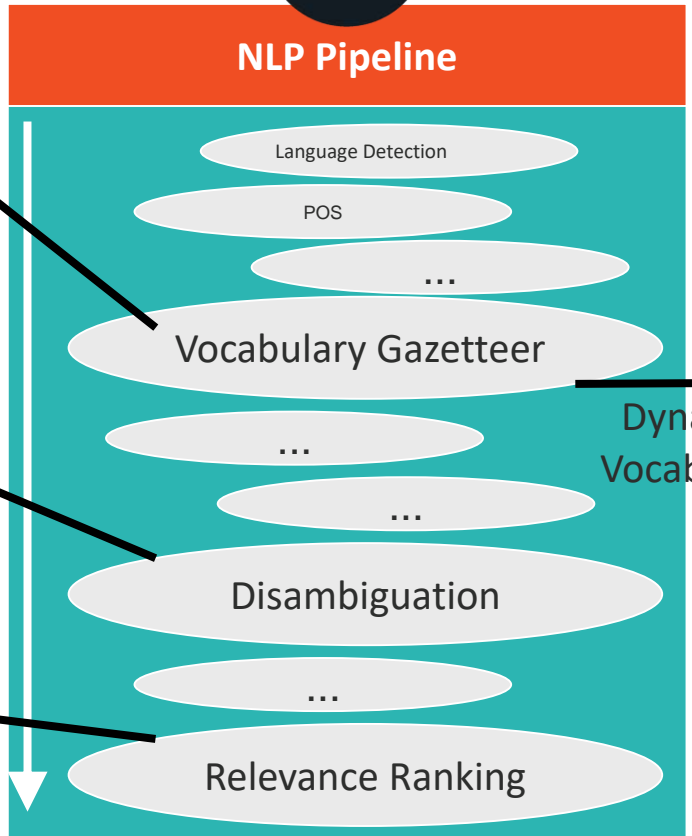
Apple : Organisation  
 Tim Cook : Person, CEO  
 Tim Cook : Person, Footballer  
 Samsung : Organisation

**Disambiguation**

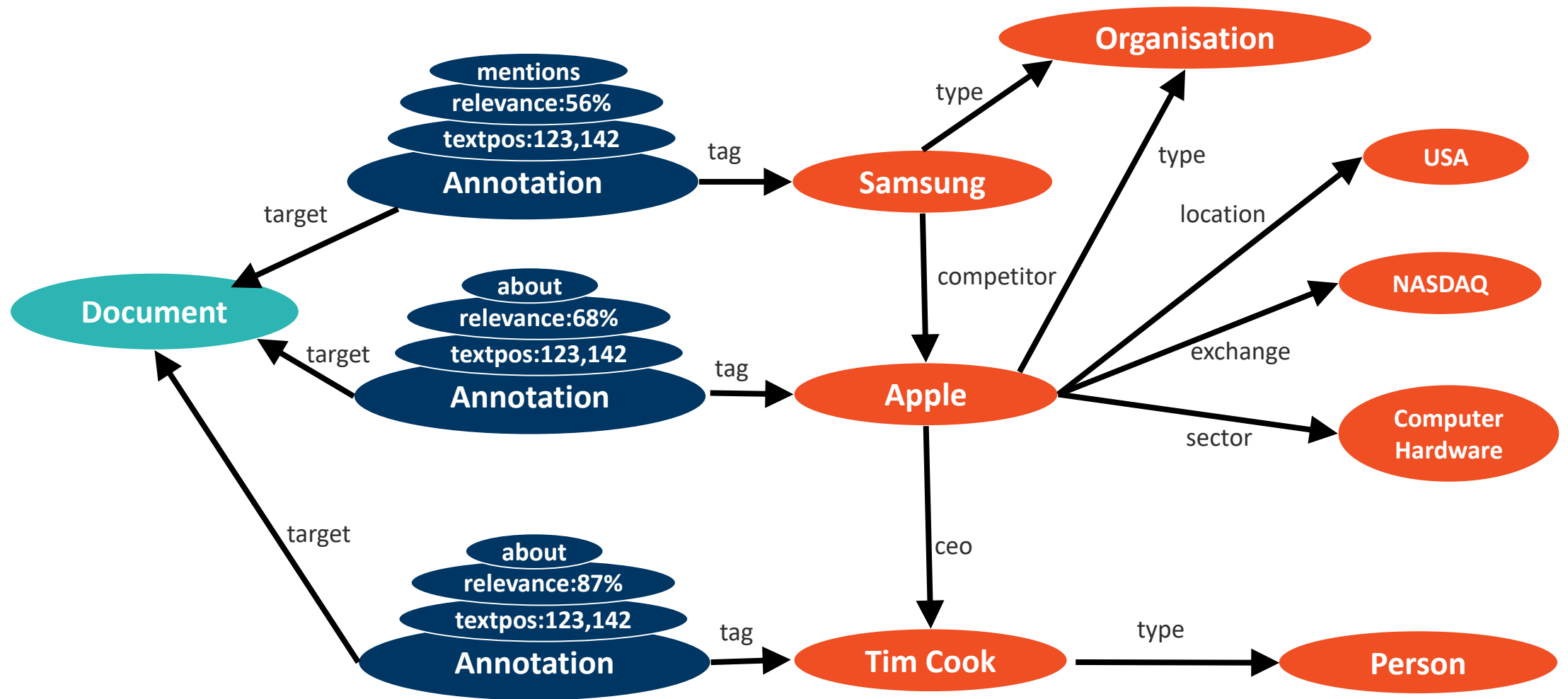
Apple : Organisation  
 Tim Cook : Person, CEO  
 Tim Cook : Person, Footballer  
 Samsung : Organisation

**Relevance**

87% - Tim Cook : Person, CEO  
 68% - Apple : Organisation  
 56% - Samsung : Organisation



# Sample Knowledge Graph with Metadata



# Linking News to Big Knowledge Graphs

- Link text to knowledge graphs
- Navigate from news to concepts and from there to other news

The screenshot shows a news article titled "EU migrants 'more likely to be working than Brits' - AOL News UK". The article text includes: "Migrants from the European Union are more likely to be in work than UK nationals, but also have a higher rate of claiming tax credits and child benefits, according to a think-tank. The study found 83% of members - including the eastern EU countries - were in work, as were 75% from EU countries, compared with 74% for UK migrants".

Knowledge graph overlays are present:

- Person:** Barack Obama, David Cameron, Hillary Rodham Clinton, Donald Trump, Presidency of Barack Obama, Jeb Bush, Ted Cruz.
- Organization:** Associated Press, Reuters, Twitter, Republican Party, European Union, Facebook, Apple Inc.
- Location:** United States of America, world, Washington, D.C., New York City, Europe, London, United Kingdom.
- THIS ARTICLE IS ABOUT:** EUROPEAN UNION (with EU flag icon).
- OTHER MENTIONS:** United Kingdom (with UK flag icon), Eastern Europe, British people, Marley Morris, News UK.
- RELATED:** (empty section).

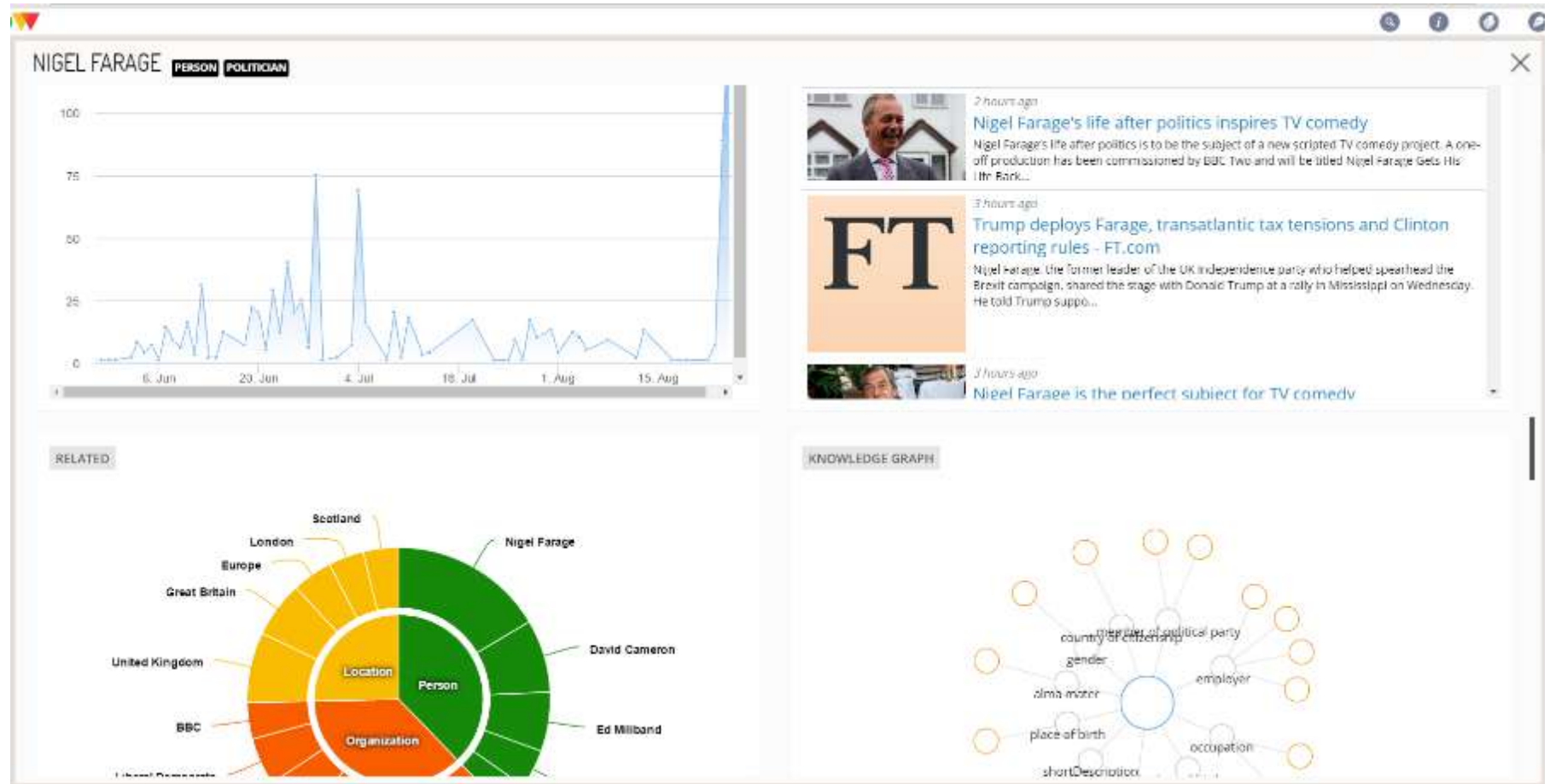
A tooltip for the "Location" overlay shows: relevance: 53.1%, confidence: 85.7%, and "known concept".

Footer: © 2013 - 2016 Ontotext AD | DSP Platform | Publishing Solutions

# Semantic Media Monitoring

For each entity:

- popularity trends
- relevant news
- related entities
- knowledge graph information



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

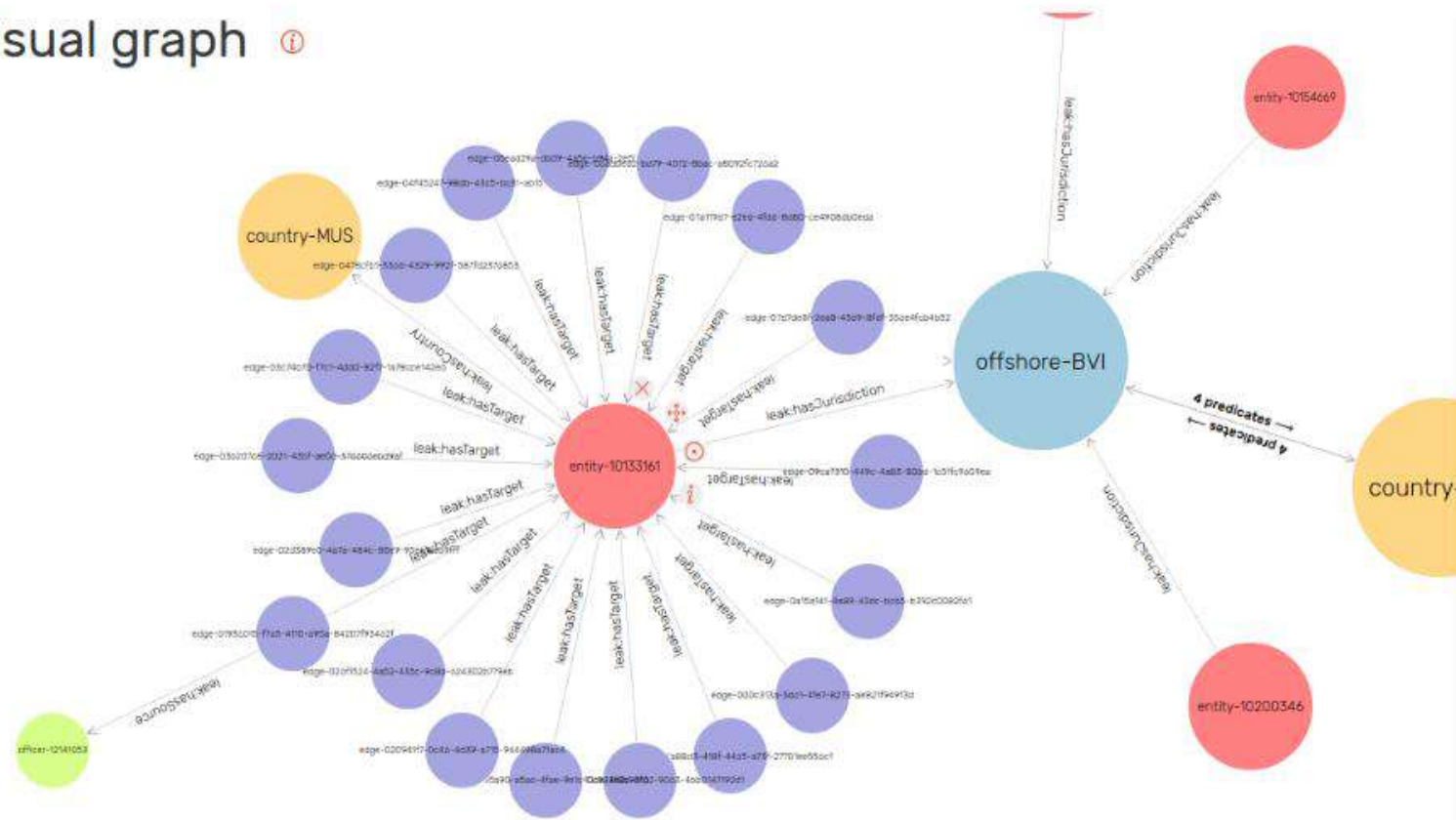


# Visual Graph: Node details

**GraphDB** STANDARD

- Import
- Explore
- Graphs overview
- Class hierarchy
- Class relationships
- Visual graph**
- SPARQL
- Monitor
- Setup
- Help

Visual graph ⓘ



leaks2

entity-10133161

Types:  
**leak.Entity**

RDF rank:  
0.33

Search instance properties

leak:address  
FORTENBERRY CORPORATE SERVICES LTD 519 ST JAMES C  
COURT; ST DENIS STREET; PORT LOUIS MAURITIUS

leak:countries  
Mauritius

leak:country\_codes  
MUS

leak:former\_name  
Trinity Financial Group Limited

leak:incorporation\_date  
2007-11-07

leak:internal\_id  
590021

leak:jurisdiction  
BVI

leak:jurisdiction\_description  
British Virgin Islands

leak:name  
Dale Capital Group Limited

leak:node\_id  
10133161

# GraphDB Workbench: Class Instances & Hierarchy

Class hierarchy ⓘ



dbo:Agent

Domain-Range Graph

Analogous to a `foaf:Agent`, an `agent` is an entity that acts. This is intended to be the super class of ...

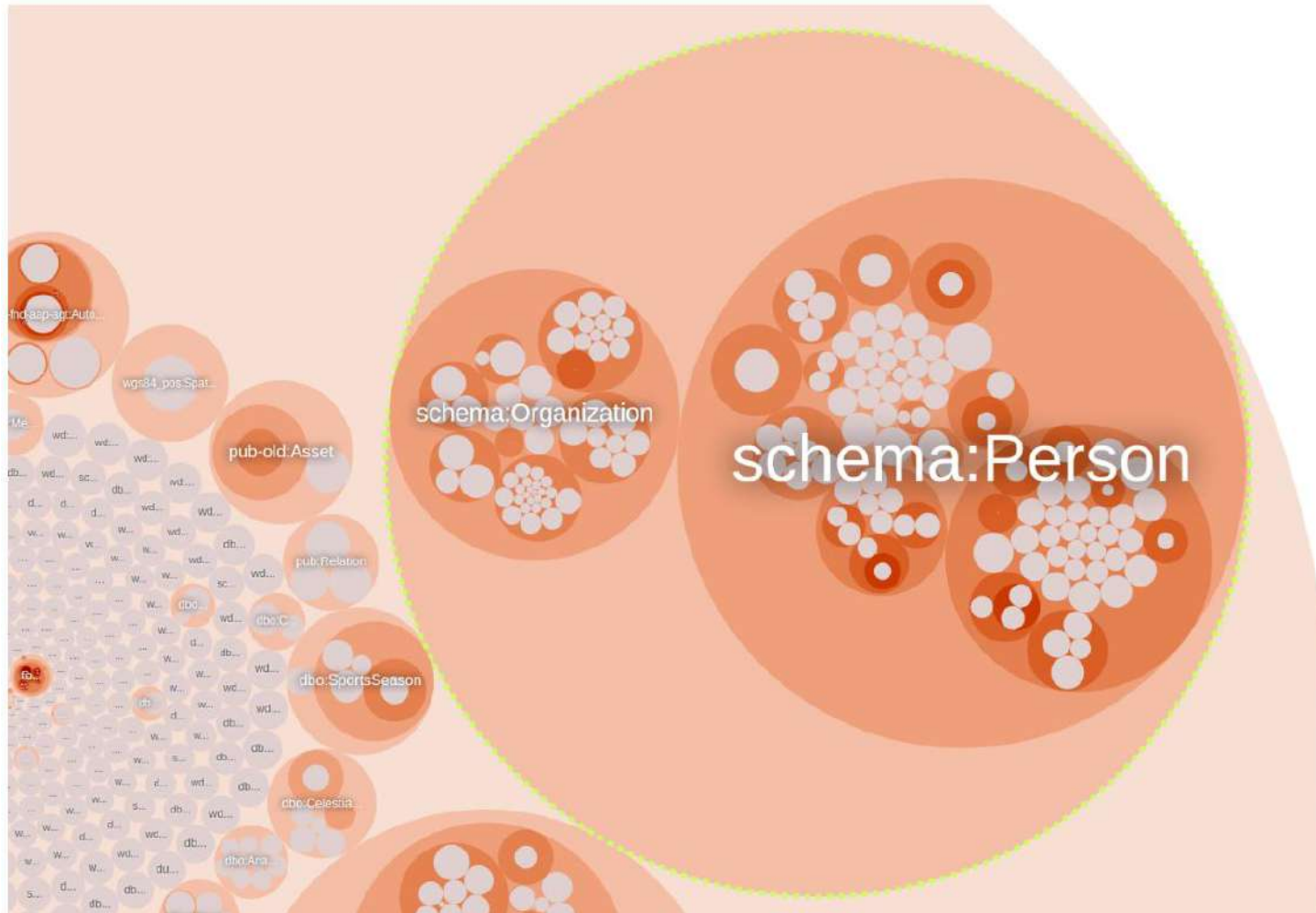
Show full comment

agent

2,959,286 instances

Search first 1000 class instances

- dbr:Alice\_Lloyd\_College
- dbr:Archbishop\_Shaw\_High\_School
- dbr:Anacostia\_High\_School
- dbr:Bishop\_Ireton\_High\_School
- dbr:Aoyama\_Gakuin\_University
- dbr:Bethesda-Chevy\_Chase\_High\_School
- dbr:Belvidere\_High\_School\_(Belvidere\_Illinois)
- dbr:Andover\_Elementary\_School
- dbr:Alchesay\_High\_School
- dbr:Alleghany\_High\_School\_(Virginia)
- dbr:Albert\_Spencer\_Wilcox
- dbr:Blackfeet\_Community\_College
- dbr:Bishop\_Maginn\_High\_School
- dbr:Bethel\_College\_(Indian



ontotext

# GraphDB Workbench: Class Relations

**FactForge**

- Explore
- Graphs overview
- Class hierarchy
- Class relationships**
- Visual graph
- SPARQL
- Help
- About

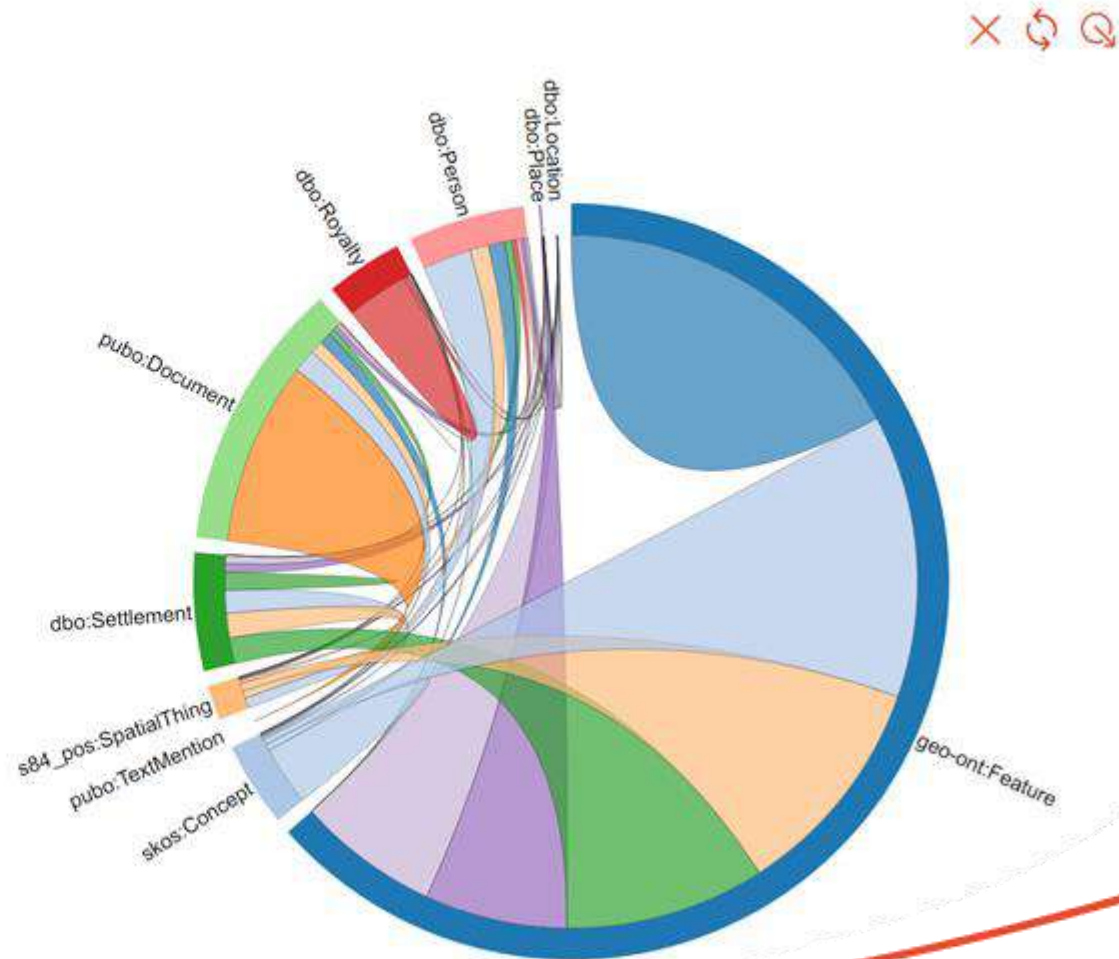
## Class relationships

Showing the dependencies between 10 classes

Filter classes

All Incoming Outgoing

Class	Links	
geo-ont:Feature	612.99M	⊖
skos:Concept	196.74M	⊖
pubo:TextMention	147.96M	⊖
wgs84_pos:SpatialThing	112.94M	⊖
dbo:Settlement	108.66M	⊖
pubo:Document	99.26M	⊖
dbo:Royalty	65.28M	⊖
dbo:Person	49.16M	⊖
dbo:Place	46.75M	⊖
dbo:Location	45.89M	⊖
dbo:PopulatedPlace	45.85M	+
foaf:Person	49.18M	+
pub:Thing	40.96M	+
dbo:Country	35.32M	+



# Presentation Outline

- Ontotext Introduction
- Technology and Portfolio
- **Cognitive Analytics Meet Big Knowledge Graphs**
- **Big Company Data: Knowing, Matching and Cleaning**
- Product Roadmap



ontotext

# Context and Awareness

- **Context allows concepts to be identified, the way people do**
- **Big knowledge graph can provide context for the entities in it**
  - ✓ *Differentiating features and similar nodes*
  - ✓ *How important and how popular it is*
  - ✓ *Related entities and concepts*
  - ✓ *Entities it is typically mentioned together with (co-occurrence)*
- **This is awareness!**
- **The kind of knowledge that people mean saying "*I am aware of X*" or "*She is cognizant of Y*"**

# The Critical Mass

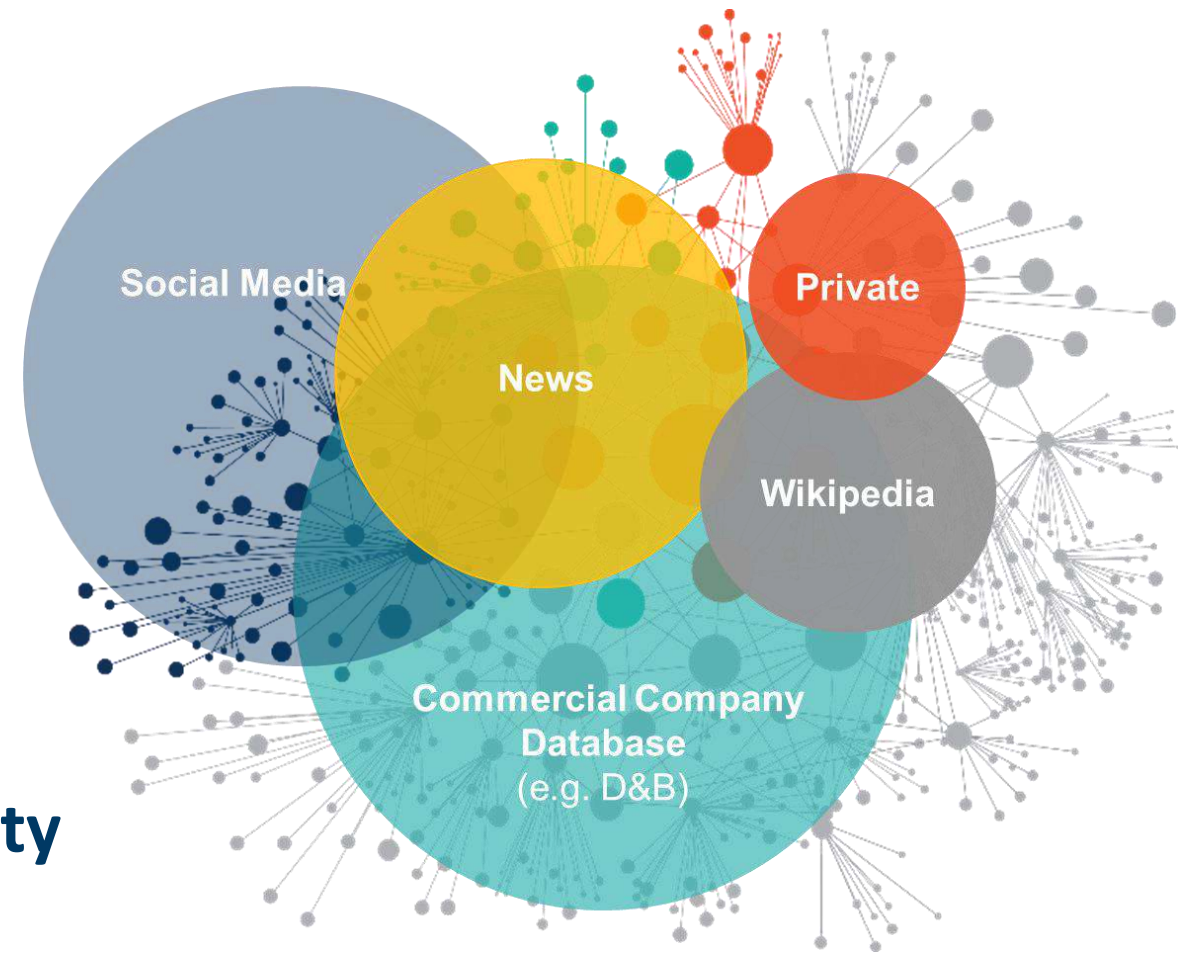
Malcolm Gladwell claims that one needs to devote 10 000 hours to become an expert in something, e.g. violin or hokey

*(Outliers)*



# The Critical Mass

- **A cognitive system needs:**
  - ✓ To know 1B facts
  - ✓ About 100M concepts and entities
  - ✓ Read 1M news articles
- **In order to reach concept and entity awareness in a specific domain**
  - ✓ The level of awareness that people mean saying  
*“My background is X”*



# Let's play an Awareness game!

- **Important airports near London?**
- **The most popular banks in UK?**
- **Companies similar to Google?**
- **People mentioned together with IBM in news?**





# We are getting closer!

- **Our Business Knowledge Model can already answer many of these questions better than you**
- **Most of this intelligence is available in the Ontotext Platform**
- **Knowledge model = KG + text mining + analytics**
- **We already offer two such knowledge models:**
  - ✓ **Business and general news:** one for processing general business master data (like people, organizations, locations and their mentions in the news)
  - ✓ **Life sciences and healthcare**



# Customized Cognitive Marketing Intelligence

- **Developing from scratch cognitive system with global knowledge is infeasible**
- **We can provide and “onboard” one for you:**
  - ✓ Suggest open and commercial data sources
  - ✓ Integrate them with your proprietary data sources
  - ✓ Tune text analytics
  - ✓ Develop specific analytics, reports, dashboards, etc.
- **We can also maintain it for you:**
  - ✓ Various support and maintenance options, including ...
  - ✓ Managed data service: updates, monitoring, data quality



# Presentation Outline

- Ontotext Introduction
- Technology and Portfolio
- Cognitive Analytics Meet Big Knowledge Graphs
- **Big Company Data: Knowing, Matching and Cleaning**
- Product Roadmap



ontotext

# Person, Organization, Location (POL) Data

- **POL data is the most common type of master/reference data**
  - ✓ Considering business applications and news
- **Open POL data is available in vast quantities**
  - ✓ Geonames covers locations exhaustively; DBpedia covers well popular POL entities; Wikidata, ...
  - ✓ Open company data grows: OpenCorporates, GLEI, open national registers, various “data leaks”
- **Within 3 years exhaustive global POL data will be commodity!**
  - ✓ And it will be widely used for BI and decision making
- **Ontotext delivers Global POL data solutions today.**
  - ✓ We make them more affordable with more cognitive analytics



# Company Data Species (1/2)

Category	Representatives	Size (Orgs.)
Exhaustive Global Databases	Dun & Bradstreet, BvD, Factset	> 200M
Rich Company Databases	Capital IQ (S&P), Thomson Reuters (various)	5-10M
Investment Databases	CrunchBase, PitchBook, CBI, DJ Venture Source	200-600K
Very Big Open Databases	OpenCorporates	130M
Global Official Open Databases	GLEI (Global Legal Identifier), EU BRIS	1-30M
Open Encyclopedic	DBpedia, Wikidata	0.3-1.2M
Open Leaks and Investigations	Panama Papers (Offshore Leaks), Trump World Data	3-300K



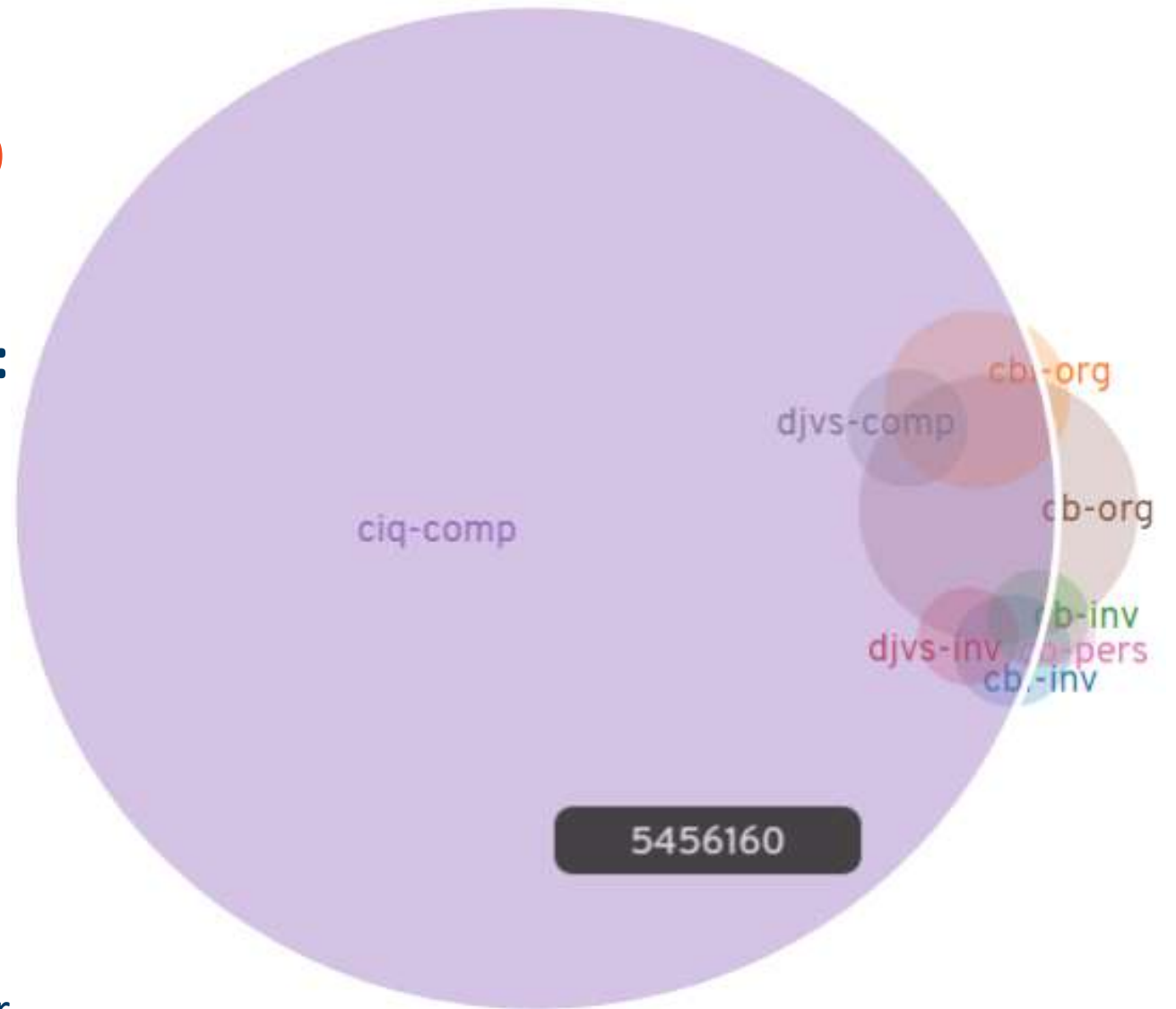
# Company Data Species (2/2)

Category	Locations	Industry Classification	High Tech. Fields	Invest. Info	Org-Org Relations (e.g.Tree)	Org-Person Relations	Clean, Correct, Predictable
Exhaustive Global Databases	++	+/-	-	-	++	+/-	6
Rich Company Databases	++	+	+/-	+/-	++	+/-	8
Investment Databases	+/-	+/-	+	+	++/-	+/-	4-6
Very Big Open Databases	+	+/-	-	-	+/-	-	8
Global Official Open Databases	+	-	-	-	+/-	-	8
Open Encyclopedic	+/-	+/-	+	-	+/-	+	3-5
Open Leaks and Investigations	+/-	-	-	-	+/-	+	4-6



# Matching and Overlap

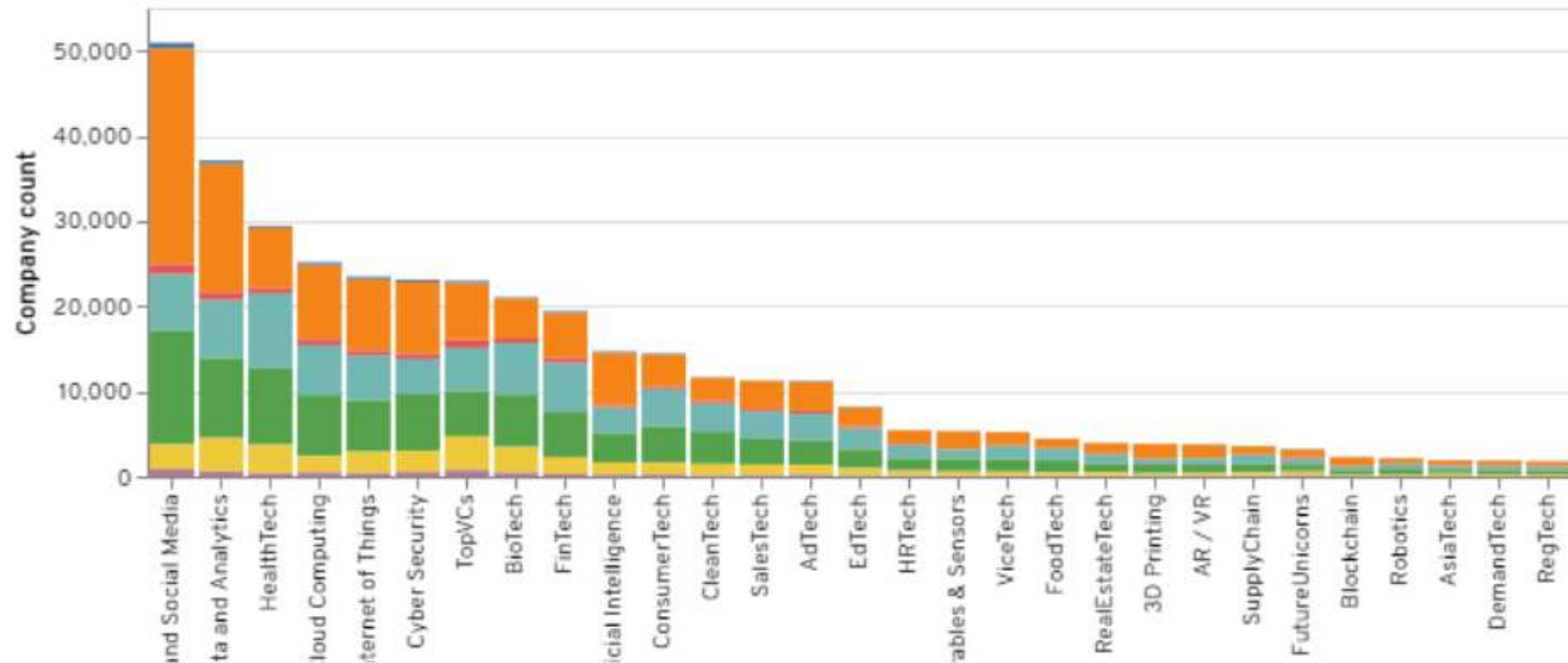
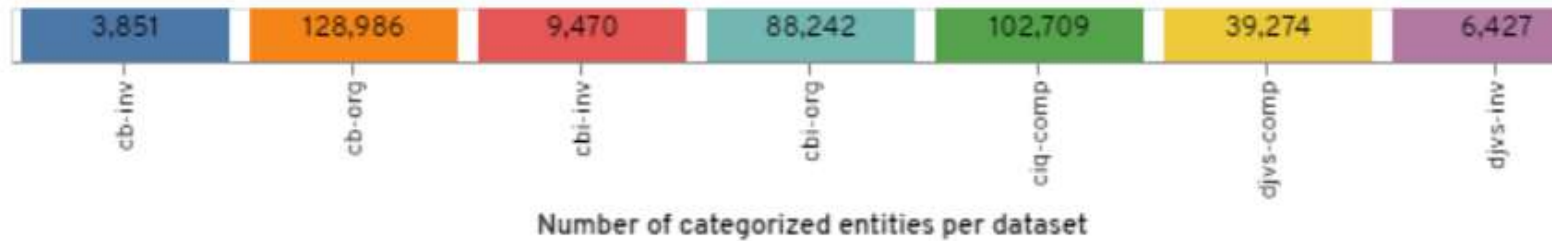
- **Organizations matched across:** CrunchBase (CB), CB Insights (CBI), Capital IQ (CIQ), DJ Venture Source, ...
- **The Venn diagram presents the overlap between sources**
  - ✓ The size of the circle indicates number of entities per source
  - ✓ The level of overlap indicates number of entities matched between the two sources



# Slice and Dice by Data Source, Industry and High-tech Field

Matching industries High-tech areas Transactions Geography

This is the distribution of entities across the high-tech areas according to the harmonized taxonomy. Click on the dataset band to select different data sets. Shift click to compare several datasets. Click on the bars of the bar chart to select individual high-tech areas. Shift-click to select multiple high-tech areas.





# Data Consolidation Across Data Sources (1/2)



## Ontotext AD

Provider of core semantic technology, text mining, and web mining solutions.

[See More](#)

Sofia-Capital, Sofia, Bulgaria

[Info](#) [Fundings](#) [Investments](#) [Ecosystem](#) [News](#)

### Basic Info

Identifier	45244983 <sup>CIQ</sup> , 85d338b8-c1d2-96b8-4c6e-57c960854878 <sup>CB</sup>
Name	Ontotext AD <sup>CIQ</sup> , Ontotext AD <sup>DJVS</sup> , Ontotext <sup>CB</sup>
Founded on	2000-11-01 <sup>CB</sup> , 2000 <sup>CIQ</sup>

### Links

Semantic Data	<a href="https://data.embryoai.com/resource/em/agent/ciq_company_45244983">https://data.embryoai.com/resource/em/agent/ciq_company_45244983</a>
Facebook	<a href="http://www.facebook.com/ontotext&lt;sup&gt;CB&lt;/sup&gt;">http://www.facebook.com/ontotext<sup>CB</sup></a>
LinkedIn	<a href="http://www.linkedin.com/company/ontotext-ad&lt;sup&gt;CB&lt;/sup&gt;">http://www.linkedin.com/company/ontotext-ad<sup>CB</sup></a>
Twitter	<a href="https://www.twitter.com/ontotext&lt;sup&gt;CB&lt;/sup&gt;">https://www.twitter.com/ontotext<sup>CB</sup></a>
CrunchBase	<a href="https://www.crunchbase.com/organization/ontotext&lt;sup&gt;CB&lt;/sup&gt;">https://www.crunchbase.com/organization/ontotext<sup>CB</sup></a>

# Data Consolidation Across Data Sources (2/2)

## Classifications

Type	Private Company <sup>CIQ</sup> , Private <sup>DJVS</sup> , company <sup>CB</sup>
Status	operating <sup>CB</sup> , Operating Subsidiary <sup>CIQ</sup>
Industry sector	Technology, Media & Telecommunications <sup>CIQ</sup> , Application Software <sup>CB</sup> , Business Services <sup>DJVS</sup> , Data Processing & Outsourced Services <sup>CB</sup> , Software <sup>DJVS</sup> , IT Services <sup>CIQ</sup> , software <sup>CB</sup> , Internet Software & Services <sup>CB</sup> , Application Software <sup>DJVS</sup> , data integration <sup>CB</sup> , Technology, Media & Telecommunications <sup>CB</sup>
Industry GICS	Application Software <sup>CB</sup> , Data Processing & Outsourced Services <sup>CB</sup> , Internet Software & Services <sup>CB</sup> , Application Software <sup>DJVS</sup>
Hitech area	Big Data and Analytics <sup>CIQ</sup> , Big Data and Analytics <sup>CB</sup>

## Geography

Country	Bulgaria <sup>DJVS</sup> , Bulgaria <sup>CIQ</sup> , BGR <sup>CB</sup>
Region	Sofia <sup>CB</sup>
City	Sofia <sup>CIQ</sup> , Sofia <sup>DJVS</sup> , Sofia <sup>CB</sup>
Street address	135 Tsarigradsko Chaussee <sup>CIQ</sup> , Polygraphia Office Center fl.4, 47A Tsarigradsko Shosse <sup>CB</sup>
Postal code	1784 <sup>CIQ</sup> , 1124 <sup>CB</sup>

# Entity Matching Across Datasets

- **Match IDs of one of the same real entity across different databases**
- **Data Challenges**
  - ✓ Different schemata
  - ✓ Name variations
  - ✓ Different classifications and codes
  - ✓ Lack of unique identifiers (even ticker symbols are not unique)
- **Technology challenges**
  - ✓ Pre-selection is needed; brute-force matching is not good for 1M against 5M companies
  - ✓ It is not trivial to come up with good pre-selection mechanism



# Company Matching Sample Project

- We matched **5+ big datasets within couple of months**
- **Fully automated procedure**, which takes few hours to execute
  - ✓ 90% SPARQL and GraphDB's FTS connectors
- **Location normalization through matching to Geonames**
  - ✓ Also industry classification alignment across the sources
- **About 85% F-Score with simple structural matching rules**
- **To get higher accuracy, you need:**
  - ✓ Massive amount of manual work and fine-tuning of weights ... or
  - ✓ Cognitive analytics (importance, similarity, highly accurate named entity recognition, etc.)

# Presentation Outline

- Ontotext Introduction
- Technology and Portfolio
- Cognitive Analytics Meet Big Knowledge Graphs
- Big Company Data: Knowing, Matching and Cleaning
- **Product Roadmap**



ontotext

# Product Roadmap

## ○ **Ontotext platform**

- ✓ Next version of our Manual Annotation Tool
- ✓ Streamlined ETL and entity matching based on SPARK
- ✓ Configurable Semantic Search front end

## ○ **GraphDB**

- ✓ Reconciliation
- ✓ Faster transactions on big knowledge graphs – 2x speed up of small transactions
- ✓ Faster SPARQL federation between local repositories
- ✓ Similarity based on Semantic Vectors



# Reconciliation

81 rows Permalink

Show as: **rows** records    Show: 5 10 25 **50** rows « first < previous 1 - 50 next > last »

All	ID	City	City2	Country	Code	Code2	Lon	Lat	Alt	Timezone offset	DST	Timezone name	
	1.	1	Goroka <small>Choose new match</small>	Goroka	Papua New Guinea	GKA	AYGA	-6.081689	145.391881	5282	10	U	Pacific/Port_Moresby
	2.	2	Madang Create new item <small>Search for match</small>	Madang	Papua New Guinea	MAG	AYMD	-5.207083	145.7887	20	10	U	Pacific/Port_Moresby
	3.	13	Hornafjörður Höfn (73) Create new item <small>Search for match</small>	Hofn	Iceland	HFN	BIHN	64.295556	-15.227222	24	0	N	Atlantic/Reykjavik
	4.	23	Shearwater Shearwater (86) Create new item <small>Search for match</small>	Halifax	Canada	YAW	CYAW	44.639721	-63.499444	167	-4	A	America/Halifax

# GraphDB Semantic Similarity Plugin

- Statistics similarity on knowledge graphs using Semantic vectors
- Creates statistical semantic models from your RDF data and search for similar terms and documents
- **Sample:**
  - Create index from the news from FactForge
  - Find similar news, find relevant terms for a news, etc..





# Similar News

## Search in content

search options 



Search type:  Term  Entity

Result type:  Term  Entity

Semantic Vectors search  
parameters:

[See the full list of supported parameters](#)

Showing results for <https://www.cnbc.com/2018/07/17/us-senate-republican-leader-warns-russia-not-to-meddle-in-2018-electio.html> [View SPARQL Query](#)

	entity 	score 
1	<a href="https://www.cnbc.com/2018/07/17/us-senate-republican-leader-warns-russia-not-to-meddle-in-2018-electio.html">https://www.cnbc.com/2018/07/17/us-senate-republican-leader-warns-russia-not-to-meddle-in-2018-electio.html</a>	"1.0"^^xsd:double
2	<a href="https://www.reuters.com/article/us-usa-russia-ryan-mccconnell/top-senate-republican-warns-russia-on-election-meddling-idUSKBN1KE2N5">https://www.reuters.com/article/us-usa-russia-ryan-mccconnell/top-senate-republican-warns-russia-on-election-meddling-idUSKBN1KE2N5</a>	"0.9297294882366121"^^xsd:double
3	<a href="https://www.politico.com/story/2018/08/06/rand-paul-russia-meeting-764589">https://www.politico.com/story/2018/08/06/rand-paul-russia-meeting-764589</a>	"0.9083294545242895"^^xsd:double
4	<a href="https://www.sfgate.com/news/article/A-letter-from-Trump-to-Putin-is-the-latest-flash-13141621.php">https://www.sfgate.com/news/article/A-letter-from-Trump-to-Putin-is-the-latest-flash-13141621.php</a>	"0.9032596414329951"^^xsd:double
5	<a href="https://www.cnbc.com/2018/07/18/trump-believes-next-meeting-with-putin-should-happen-after-the-russia.html">https://www.cnbc.com/2018/07/18/trump-believes-next-meeting-with-putin-should-happen-after-the-russia.html</a>	"0.9021681352123165"^^xsd:double
6	<a href="http://thehill.com/policy/defense/policy-strategy/399203-mattis-denies-policy-changes-made-at-trumps-meeting-with-putin">http://thehill.com/policy/defense/policy-strategy/399203-mattis-denies-policy-changes-made-at-trumps-meeting-with-putin</a>	"0.8998150046751944"^^xsd:double
7	<a href="https://nypost.com/2018/07/18/majority-of-americans-disapprove-of-trumps-dealings-with-russia/">https://nypost.com/2018/07/18/majority-of-americans-disapprove-of-trumps-dealings-with-russia/</a>	"0.8997648702930882"^^xsd:double

# Take home

- Business needs **global company data for market intelligence**
- **This is rocket science**
  - ✓ Mainstream tech cannot deal with such diversity
  - ✓ Semantic data integration and cognitive analytics needed
- **Ontotext is ready to help**
  - ✓ **Consulting:** help you build the concept for your next generation MI system
  - ✓ **Develop:** build one for you or support you developing your platform
  - ✓ **Support and operations:** from Level 3 support to Managed services



# Thank you!

Experience the technology with our demonstrators

**NOW:** Semantic News Portal <http://now.ontotext.com>

**RANK:** News popularity ranking for companies <http://rank.ontotext.com>

**FactForge:** Hub for open data and news about People and Organizations  
<http://factforge.net>