Semantics 2018, Vienna

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

- Ontotext Introduction
- Technology and Portfolio
- Cognitive Analytics Meet Big Knowledge Graphs
- Big Company Data: Knowing, Matching and Cleaning
- Product Roadmap
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
“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
      [Link](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
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**
Presentation Outline

- Ontotext Introduction
- Technology and Portfolio
- Cognitive Analytics Meet Big Knowledge Graphs
- Big Company Data: Knowing, Matching and Cleaning
- Product Roadmap
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
Apple CEO Tim Cook was at a conference with the CEO of Samsung. Tim explained how smart phones are changing the consumer electronics market.
Sample Knowledge Graph with Metadata

Document

Annotation

mentions
relevance:56%
textpos:123,142

target

Annotation

about
relevance:68%
textpos:123,142

Annotation

about
relevance:87%
textpos:123,142

Organisation

Samsung

type

Organisation

Apple

type

Person

Tim Cook

type

USA

NASDAQ

Computer Hardware

location

exchange

sector

type

type

ceo

target

target

target
Linking News to Big Knowledge Graphs

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

Try it at http://now.onotext.com
Semantic Media Monitoring

For each entity:
- popularity trends
- relevant news
- related entities
- knowledge graph information

Try it at [http://now.ontotext.com](http://now.ontotext.com)
Visual Graph: Node details
GraphDB Workbench: Class Instances & Hierarchy
Presentation Outline

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

<table>
<thead>
<tr>
<th>Category</th>
<th>Representatives</th>
<th>Size (Orgs.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exhaustive Global Databases</td>
<td>Dun &amp; Bradstreet, BvD, Factset</td>
<td>&gt; 200M</td>
</tr>
<tr>
<td>Rich Company Databases</td>
<td>Capital IQ (S&amp;P), Thomson Reuters (various)</td>
<td>5-10M</td>
</tr>
<tr>
<td>Investment Databases</td>
<td>CrunchBase, PitchBook, CBI, DJ Venture Source</td>
<td>200-600K</td>
</tr>
<tr>
<td>Very Big Open Databases</td>
<td>OpenCorporates</td>
<td>130M</td>
</tr>
<tr>
<td>Global Official Open Databases</td>
<td>GLEI (Global Legal Identifier), EU BRIS</td>
<td>1-30M</td>
</tr>
<tr>
<td>Open Encyclopedic</td>
<td>DBPedia, Wikidata</td>
<td>0.3-1.2M</td>
</tr>
<tr>
<td>Open Leaks and Investigations</td>
<td>Panama Papers (Offshore Leaks), Trump World Data</td>
<td>3-300K</td>
</tr>
</tbody>
</table>
# Company Data Species (2/2)

<table>
<thead>
<tr>
<th>Category</th>
<th>Locations</th>
<th>Industry Classification</th>
<th>High Tech. Fields</th>
<th>Invest. Info</th>
<th>Org-Org Relations (e.g. Tree)</th>
<th>Org-Person Relations</th>
<th>Clean, Correct, Predictable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exhaustive Global Databases</td>
<td>++</td>
<td>+/-</td>
<td>-</td>
<td>-</td>
<td>++</td>
<td>+/-</td>
<td>6</td>
</tr>
<tr>
<td>Rich Company Databases</td>
<td>++</td>
<td>+</td>
<td>+/-</td>
<td>+/-</td>
<td>++</td>
<td>+/-</td>
<td>8</td>
</tr>
<tr>
<td>Investment Databases</td>
<td>+/-</td>
<td>+/-</td>
<td>+</td>
<td>+</td>
<td>+++/-</td>
<td>+/-</td>
<td>4-6</td>
</tr>
<tr>
<td>Very Big Open Databases</td>
<td>+</td>
<td>+/-</td>
<td>-</td>
<td>-</td>
<td>+/-</td>
<td>-</td>
<td>8</td>
</tr>
<tr>
<td>Global Official Open Databases</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+/-</td>
<td>-</td>
<td>8</td>
</tr>
<tr>
<td>Open Encyclopedic</td>
<td>+/-</td>
<td>+/-</td>
<td>+</td>
<td>-</td>
<td>+/-</td>
<td>+</td>
<td>3-5</td>
</tr>
<tr>
<td>Open Leaks and Investigations</td>
<td>+/-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+/-</td>
<td>+</td>
<td>4-6</td>
</tr>
</tbody>
</table>
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
Data Consolidation Across Data Sources (1/2)
**Data Consolidation Across Data Sources (2/2)**

### Classifications

<table>
<thead>
<tr>
<th>Type</th>
<th>Private Company[^CIO]</th>
<th>Private[^DJVS], company[^CB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status</td>
<td>operating[^CB], Operating Subsidiary[^CIO]</td>
<td></td>
</tr>
<tr>
<td>Industry sector</td>
<td>Technology, Media &amp; Telecommunications[^CIO], Application Software[^CB], Business Services[^DJVS], Data Processing &amp; Outsourced Services[^CB], Software[^DJVS], IT Services[^CIO], software[^CB], Internet Software &amp; Services[^CB], Application Software[^DJVS], data integration[^CB], Technology, Media &amp; Telecommunications[^CB]</td>
<td></td>
</tr>
<tr>
<td>Industry GICS</td>
<td>Application Software[^CB], Data Processing &amp; Outsourced Services[^CB], Internet Software &amp; Services[^CB], Application Software[^DJVS]</td>
<td></td>
</tr>
<tr>
<td>Hitech area</td>
<td>Big Data and Analytics[^CIO], Big Data and Analytics[^CB]</td>
<td></td>
</tr>
</tbody>
</table>

### Geography

<table>
<thead>
<tr>
<th>Country</th>
<th>Bulgaria[^DJVS], Bulgaria[^CIO], BGR[^CB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region</td>
<td>Sofia[^CB]</td>
</tr>
<tr>
<td>City</td>
<td>Sofia[^CIO], Sofia[^DJVS], Sofia[^CB]</td>
</tr>
<tr>
<td>Street address</td>
<td>135 Tsarigradsko Chaussee[^CIO], Polygraphia Office Center fl.4, 47A Tsarigradsko Shosse[^CB]</td>
</tr>
<tr>
<td>Postal code</td>
<td>1784[^CIO], 1124[^CB]</td>
</tr>
</tbody>
</table>
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
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

<table>
<thead>
<tr>
<th>ID</th>
<th>City</th>
<th>City2</th>
<th>Country</th>
<th>Code</th>
<th>Code2</th>
<th>Lon</th>
<th>Lat</th>
<th>Alt</th>
<th>Timezone offset</th>
<th>DST</th>
<th>Timezone name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Goroka</td>
<td>Goroka</td>
<td>Papua New Guinea</td>
<td>GKA</td>
<td>AYDA</td>
<td>-6.081689</td>
<td>145.391881</td>
<td>5382</td>
<td>10</td>
<td>U</td>
<td>Pacific/Port_Moresby</td>
</tr>
<tr>
<td>2</td>
<td>Madang</td>
<td>Madang</td>
<td>Papua New Guinea</td>
<td>MAG</td>
<td>AYMO</td>
<td>-5.207083</td>
<td>145.7887</td>
<td>20</td>
<td>10</td>
<td>U</td>
<td>Pacific/Port_Moresby</td>
</tr>
<tr>
<td>3</td>
<td>Hornafjörður</td>
<td>Hofn</td>
<td>Iceland</td>
<td>HFN</td>
<td>BJHN</td>
<td>64.295866</td>
<td>-15.272222</td>
<td>24</td>
<td>0</td>
<td>N</td>
<td>Atlantic/Raykjavik</td>
</tr>
<tr>
<td>4</td>
<td>Shearwater</td>
<td>Halifax</td>
<td>Canada</td>
<td>YAW</td>
<td>CYAW</td>
<td>44.639721</td>
<td>-63.499444</td>
<td>167</td>
<td>-4</td>
<td>A</td>
<td>America/Halifax</td>
</tr>
</tbody>
</table>
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 RDF resources or enter term keywords

**search options**

**Search type:** Term | Entity

**Result type:** Term | Entity

Semantic Vectors search parameters:

- numsearchresults 20
- searchtype

See the full list of supported parameters.


<table>
<thead>
<tr>
<th>entity</th>
<th>score</th>
</tr>
</thead>
</table>
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


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

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