euBusinessGraph: Company and Economic Data for Innovative Products and Services

Vladimir Alexiev vladimir.alexiev@ontotext.com
Ontotext Corp

Atanas Kiryakov atanas.kiryakov@ontotext.com
Ontotext Corp

Plamen Tarkalanov plamen.tarkalanov@ontotext.com
Ontotext Corp

ACM Subject Categories
Resource Description Framework (RDF)
Ontologies
Enterprise applications

Keywords

euBusinessGraph firmographics company data public procurement linked data business graph economics

Abstract

Corporate information, including basic company firmographics (e.g., name(s), incorporation data, registered addresses, ownership and related entities), financials (e.g., balance sheets, ratings) and contextual data (e.g., addresses, economic activity classification, key officers, public tenders data, press mentions and events) are the foundation that many data value chains are built on. Furthermore, this type of information contributes to the transparency and accountability of enterprises, is instrumental when input to the process of marketing and sales, and plays a key role in many business interactions. Collecting and aggregating data about a business entity from several public sources (be it private/public, official or non-official ones), and especially across country borders and languages is a tedious, time consuming, error prone, and expensive operation which renders many potential business models non-feasible.

The euBusinessGraph project integrates European company and economic data from various data providers, including OpenCorporates (the largest open database of company info crawled from official registers), Norway's Bronnoysund Register Center (official register data), SpazioDati (rich IT data from official registers, additional databases, web crawl of company sites, tender info, etc.), EventRegistry events, GLEI, Panama Leaks, etc.

euBusinessGraph is intended to overcome these barriers and provision several important business cases, such as economic journalism (Deutsche Welle), publication of rich company data (BRC), tender information service (CERVED), business intelligence (EVRY), etc. It will also provide a marketplace of company data, with some free search and faceting, leading to information about richer Data Offerings by specific providers and their pricing.

We will present the work done on exploring relevant ontologies and vocabularies for describing companies, systems of identifiers, development of a unified data model, plans for data flows, data aggregation, matching and cross-linking, and the opportunities that lie ahead for the business cases and the data marketplace.

This session will demonstrate services like company popularity ranking, monitoring mentions of related entities in news and finding suspicious relation patterns in FactForge – a knowledge graph with more than 2 billion edges of POL (persons, organizations, locations) data interlinked with 1 million news articles.

An accompanying poster and presentation are available.

1. Introduction

euBusinessGraph is a 30-month research project funded by the EU H2020 programme on Big Data integration and experimentation that started in January 2017. Its purpose is to integrate a large number of economics-related datasets such as companies, public procurement tenders, company events, etc.; to provision 6 business cases using that data; and to establish a data marketplace for such data. The project partners comprise:

1. SINTEF , Norway's largest research organization;
2. OpenCorporates , the largest aggregation of company data from official registers;
3. Cerved , an Italian leader in business information;
4. SpazioDati , an innovative provider of data about Italian companies (Atoka);
5. EVRY , Norway's largest IT services provider;
6. Deutsche Welle (DW), Germany's international broadcaster;
7. Ontotext , a world leader in semantic technologies, databases and semantic cloud offerings;
8. Brønnøysund Register Centre (BRC), Norway's national business register agency (18 registers);
9. Institut "Jozef Stefan" (IJS), creator of EventRegistry (6.5M world events extracted from 190M articles);
10. University of Milano-Bicocca (UNIMIB), specializing in data management and service science.

The project concept including data value chain and customer segmentation, is shown below.

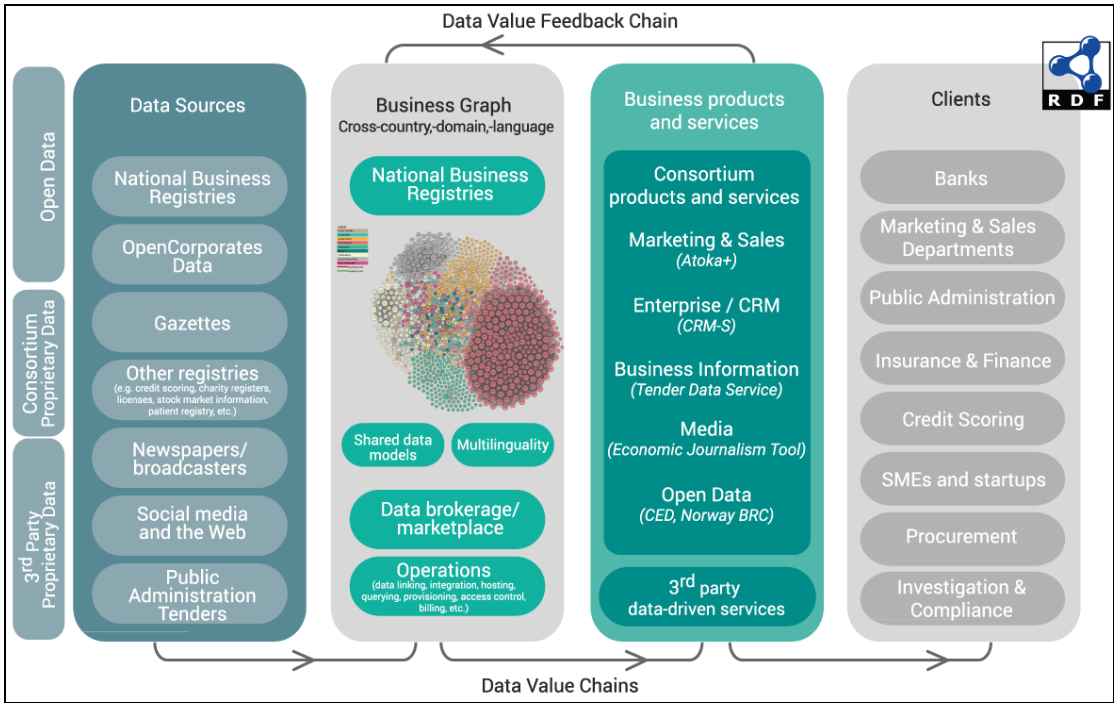

Figure 1. Project Concept

2. Business Cases and Lean Business Modeling

The business cases addressed by the project include the following:

- OpenCorporates will enlarge its offering with a Corporate Events Data Service. It will increase the

number of EU Government Gazettes ingested (see OpenGazettes) and add non-register sources of company events.

- Cerved will implement a Tender Discovery Service that will integrate data from various sources of public procurement info and implement a recommendation component that can suggest relevant tenders to companies.
- SpazioDati will increase the reach of its Atoka company information service (Atoka+) by aggregating data about more countries.
- Deutsche Welle will implement a new Economic Journalism tool to save cost in the editorial department and increase public reach through improved story telling.
- EVRY will leverage the business graph to be developed by the project in a set of Customer Relationship Management (CRM) systems, increasing data reach and quality.
- BRC will publish the full spectrum of official register data in RDF and describe the data in a semantic way, increasing the value of Norwegian LOD.

Business case development proceeds in an agile fashion, using Lean Business Modeling and Minimum Viable Product approaches to ensure the developed products meet market demand. We use tools such as Product Vision Statement, Lean Business Model Canvas, Value Proposition Canvas, GO product roadmap, Test card, and Learning card as adopted by SINTEF's innovation department. Below is an example of a lean Business Model Canvas [1] for Atoka+

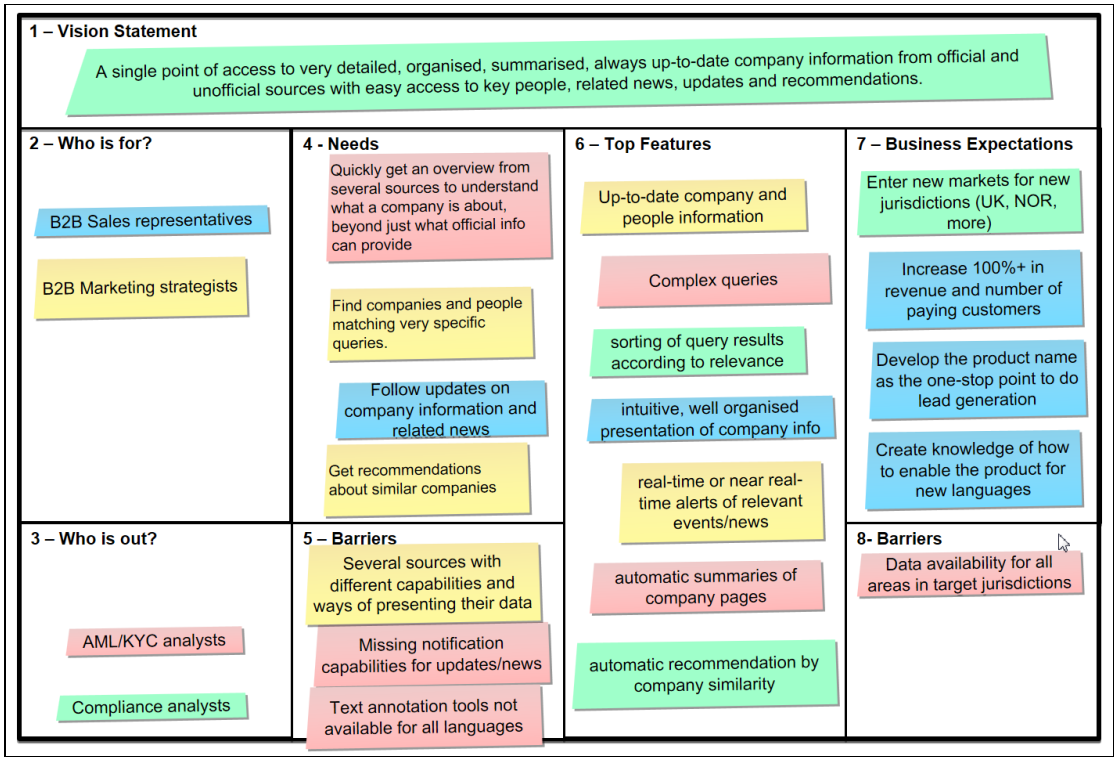

Figure 2. Example Business Model Canvas

Various monetization models are considered by the partners, including free, freemium and paid (for premium data). Describing such dataset offerings is an important function of the project.

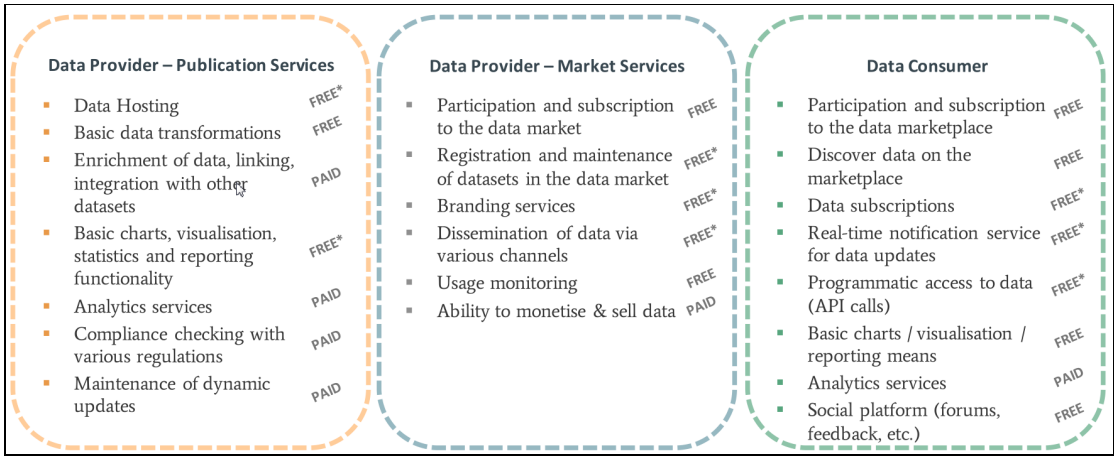

Figure 3. Example Monetization Models

3. Company Datasets and Ontologies

The project studied numerous ontologies and datasets that represent company and related economic data, and which were taken into account when creating the project's data model. Amongst them are:

1. Global Legal Entity Identifier (GLEI) [13] ;
2. Business Registers Interconnection System (BRIS) [9] [10] ;
3. Financial Industry Business Ontology (FIBO) [8] ;
4. OpenCorporates schema [19] ;
5. Bulgarian Trade Register schema [18] ;
6. W3C Organization ontology [5] ;
7. W3C Registered Organization ontology [20] ;
8. W3C Location ontology [2] ;
9. Dun & Bradstreet company data;
10. Panama Papers offshore company dataset [11] and its RDF representation Linked Leaks [17] ;
11. Wikidata properties for describing companies, especially company identifiers in various registers;
12. A number of related and subsidiary ontologies and code lists such as Schema.org, Dublin Core, IANA language tags, NUTS and LAU (EU administrative regions), NACE (EU economic activities), etc.

[23] presents some of these data artefacts, and relates each of them to the "5V" of big data (Variety, Volume, Velocity, Veracity, leading up to Value). Many of these data sources are not in semantic format, therefore mapping and conversions are required if they are to be integrated semantically. Below is an example of such mapping: the GLEI XML schema to the FIBO ontology and a custom GLEI Ontology (GLEIO) [16] .

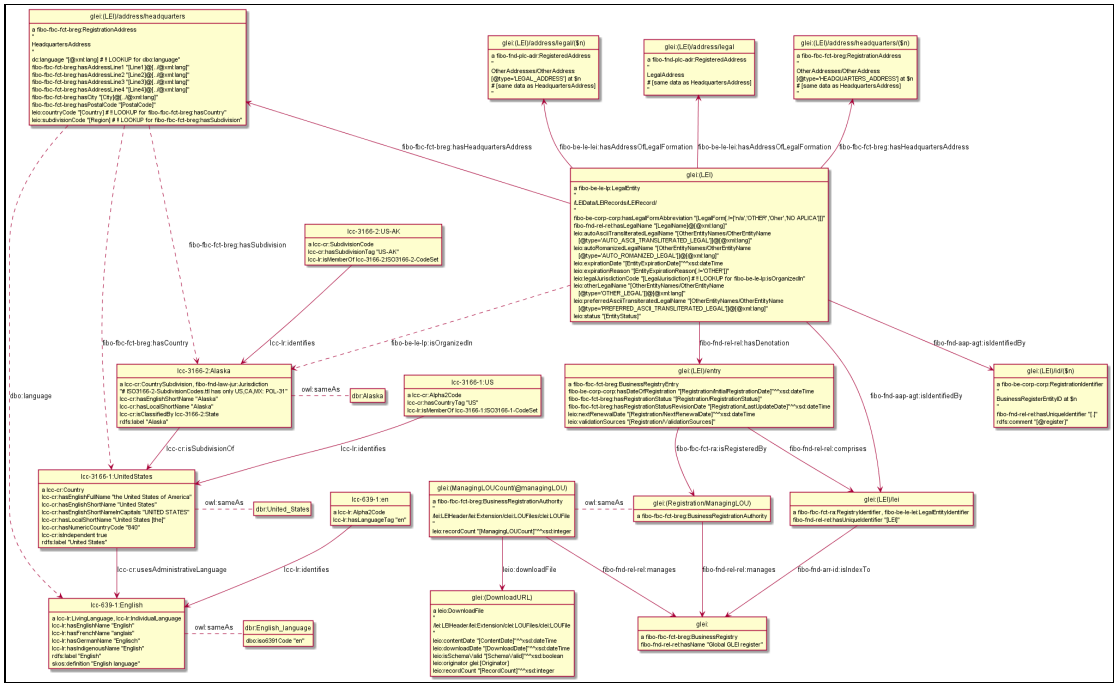

Figure 4. Mapping of GLEI to FIBO and GLEIO

4. Semantic Data Model

The euBusinessGraph semantic data model [24] is a synthesis of the various data artefacts that we studied, fulfilling the data needs of the project. It includes detailed treatment of classes, properties, values, scope notes, data provider rules, URL conventions, etc. It reuses ontologies such as W3C Org, RegOrg, Locn, ADMS; DC, DCT, schema.org, SKOS, SIOC. It reuses datasets such as NACE (economic classification), NUTS+LAU (EU regions), GLEI RAL (registration authority list).

This initial draft covers companies, company types, status, addresses, NUTS+LAU regions, NACE classification of economic activity and registrations (identifiers). We will shortly add detailed information about registers: dataset offerings, what level of detail they cover, per-company URL templates, MIME types. Future versions will add information about officers (directors, executives), provenance (who contributed which data). We used rdfpuml [25] to generate semantic model diagrams (see Figure 5) and Object-Role Modeling through the Norma tool [15] (see Figure 6) to generate an RDF representation.

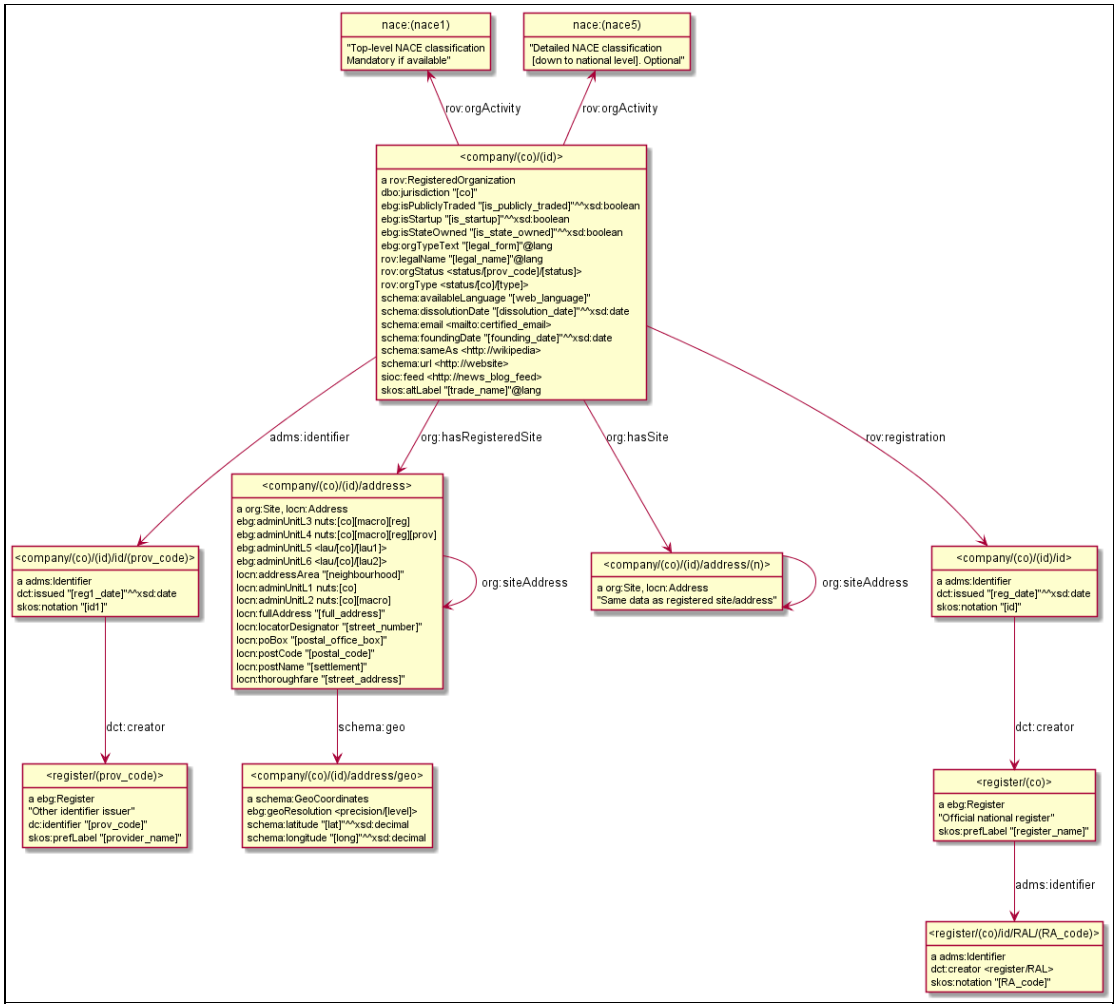

Figure 5. euBusiness Graph Semantic Model Diagram

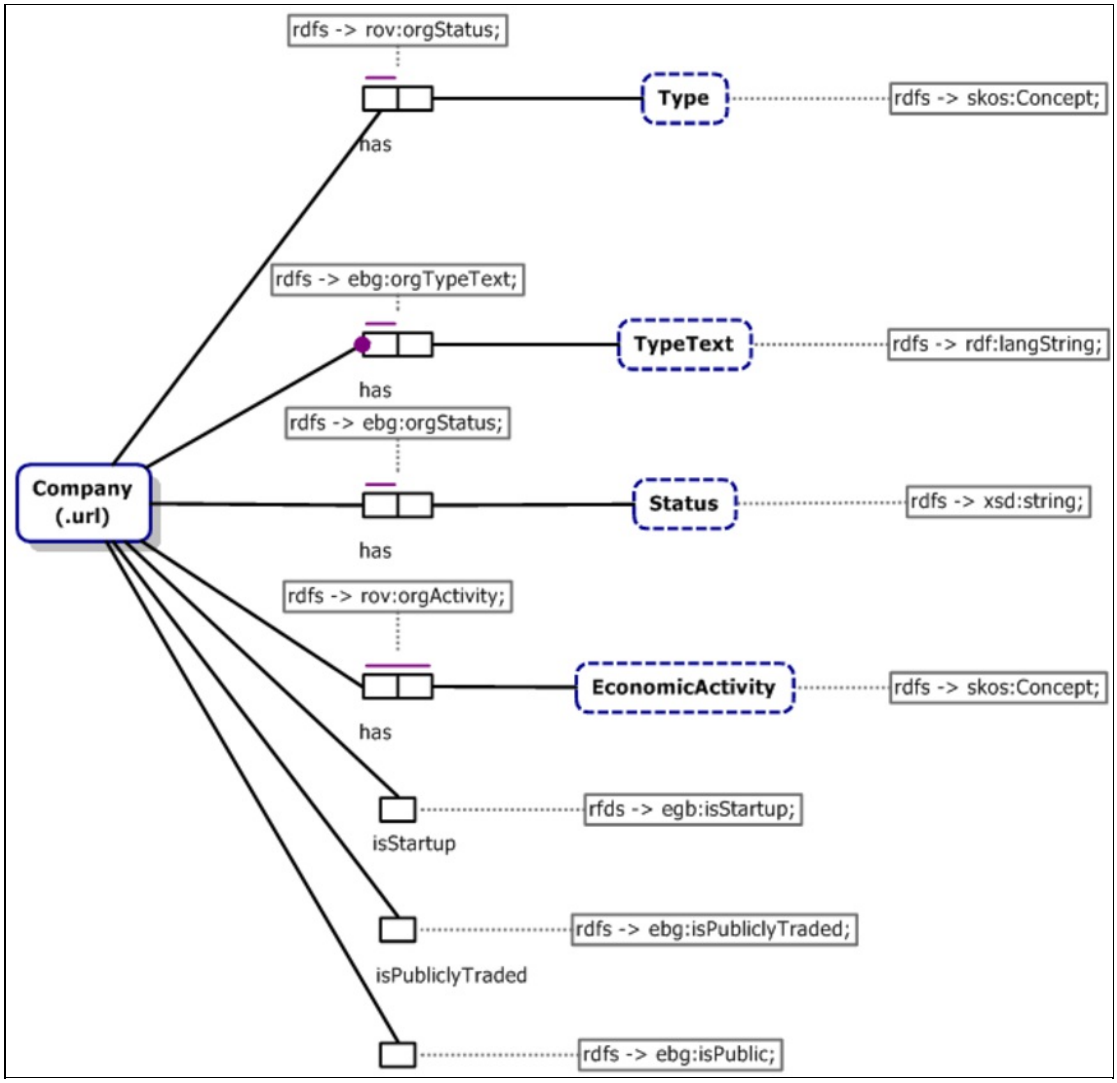

Figure 6. Object-Role Diagram of Part of the Semantic Model

5. Technologies Used

In addition to partner datasets, the project may use some of the following technologies and tools:

- The Ontotext GraphDB (formerly OWLIM) semantic database [3]
- Ontotext Cognitive Cloud (successor of the Self-service semantic system) provides on-demand access to text analytics, semantic graph databases and Linked Data technology in the cloud (Amazon Web Services). A user can start building Smart Data prototypes without the need for licensing, provisioning, installation and maintenance. In the project, the Cognitive Cloud will be used as platform for developing, running and hosting the euBusinessGraph Marketplace and Services.

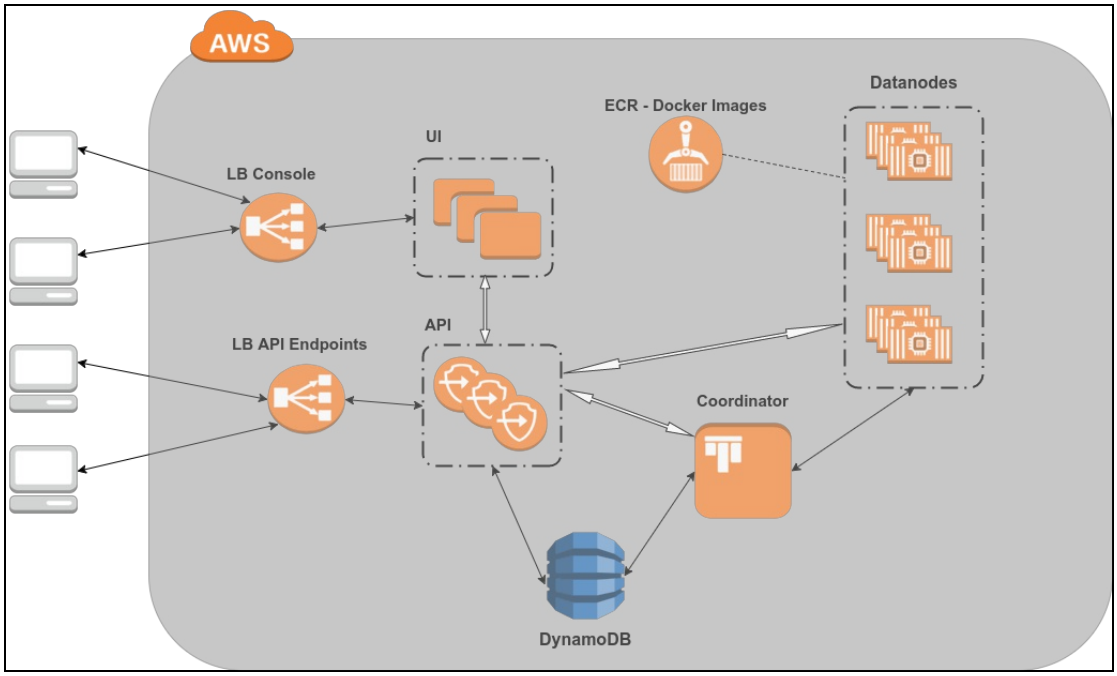

Figure 7. Ontotext Cognitive Cloud Architecture

- DataGraft [6] [7] "One stop shop for hosted data management" is a tool to interactively build, modify and share data transformations; reuse transformations to repeatably clean and transform spreadsheet data; and finally host and share datasets in a semantic way.
- RDF by Example [25] is an Ontotext tool for semantic modeling and R2RML Generation
- Dandelion API is a semantic text analysis tool used by SpazioDati.
- Wikifier is a semantic annotation tool for 100s of languages used by JSI for EventRegistry.
- ABSTAT is a tool for Linked Data Summarization with ABstraction and STATistics by UNIMIB.

Other semantic conversion tools that may potentially be used include:

- TARQL is a command-line tool for converting CSV files to RDF using SPARQL 1.1 syntax.
- XSPARQL [4] is W3C submission for a query language that melds XQuery [12] and SPARQL [21] , and the respective tool that implements it. It transforms XML, JSON and relational data (RDB2RDF) to RDF (lifting), SPARQL querying and conversions to other formats (lowering), the full power of XQuery for control flow and sequence processing, and scripting for Web data integration in general.

6. Demonstration with FactForge

As part of our presentation, we demonstrate services such as company popularity ranking, monitoring mentions of related entities in news, and finding suspicious relation patterns. We use data from FactForge [14] , a knowledge graph of about 2 billion triples. It represents a hub of open data and news about persons, organizations and locations (POL data):

1. Popular LOD datasets: DBPedia, Geonames, WorldFacts and Wordnet.
2. POL data: Panama Papers, GLEI, Trump World. People and organizations from these datasets are mapped to DBPedia. The schemata of these datasets and DBPedia are partially mapped to FIBO classes and relationships in order to allow for unified querying and analytics across the different datasets.
3. A live stream of news metadata, linking the articles to entities and concepts: about 2000 news per day tagged with the News On the Web (NOW) semantic news demonstrator.

FactForge allows one to play with the data through GraphDB's Visual Graph Explorer – a customizable interactive exploration tool (see below). See [22] for a detailed report on visualization capabilities.

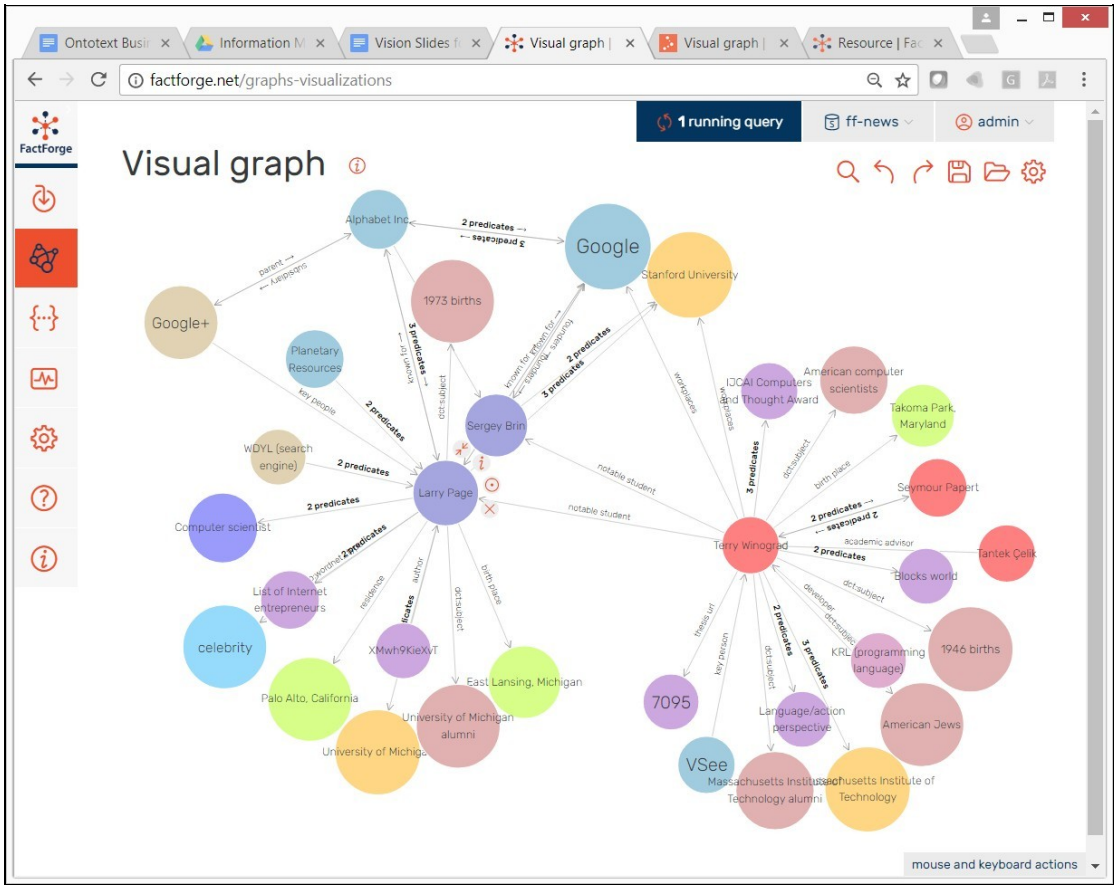

Figure 8. GraphDB Workbench Visual Graph

Sample queries in FactForge demonstrate media monitoring, tracing company control, industry trends and statistics. One of them is a SPARQL query that detects the following pattern: a company that controls another company in the same country, through a company registered in an offshore zone.

Finally, Rank is a demonstration service for news popularity ranking of companies. It supports ranking with consolidation of mentions of subsidiaries and is based on FactForge.

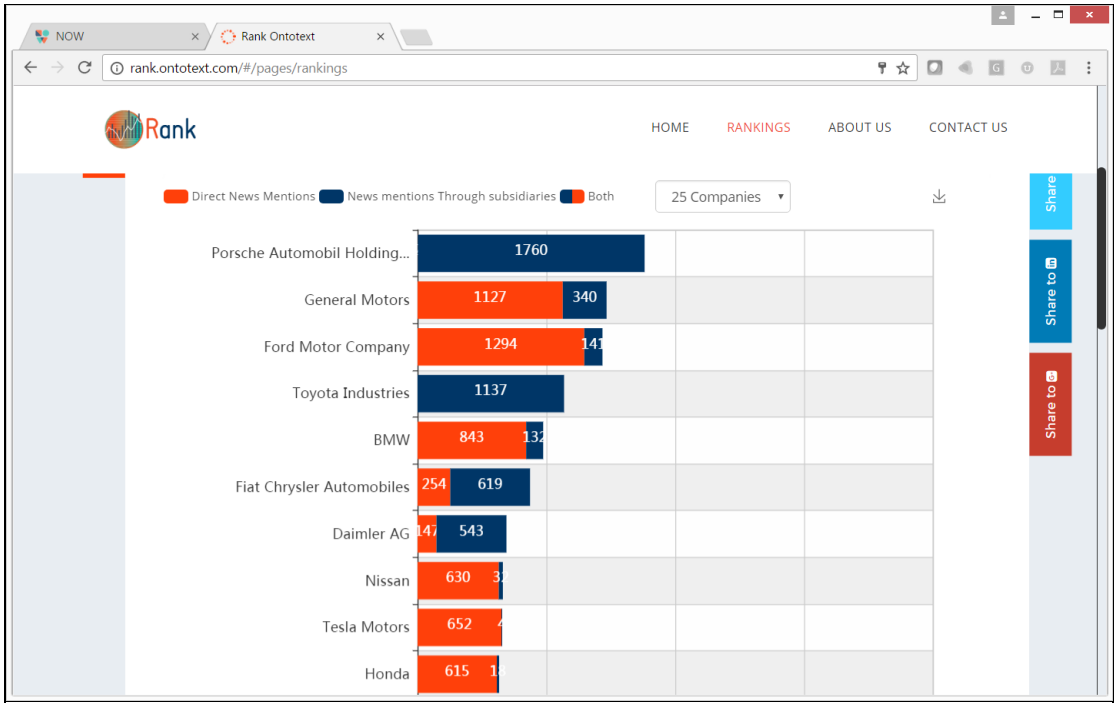

Figure 9. Global Ranking of Automotive Companies

Acknowledgements

euBusinessGraph has received funding from the EU H2020 research and innovation programme "Big Data PPP: cross-sectorial and cross-lingual data integration and experimentation" under grant agreement No 732003.

References

1. Alexander Osterwalder, Yves Pigneur. Business Model Generation , 2010.
2. Andrea Perego, Michael Lutz. ISA Programme Location Core Vocabulary . Second version, 2015-03-23
3. Barry Bishop, Atanas Kiryakov, Damyan Ognyanoff, Ivan Peikov, Zdravko Tashev, Ruslan Velkov. OWLIM: A Family of Scalable Semantic Repositories. Semantic Web Journal, Volume 2, Number 1. May 2011.
4. Daniele Dell'Aglio, Axel Polleres, Nuno Lopes, and Stefan Bischof. Querying the Web of Data with XSPARQL 1.1 . ISWC Developers Workshop, 2014
5. Dave Reynolds. The Organization Ontology , W3C Recommendation, 16 January 2014
6. Dumitru Roman, Marin Dimitrov, Nikolay Nikolov, Antoine Putlier, Brian Elvesæter, Alex Simov, Yavor Petkov. DataGraft: a Platform for Open Data Publishing . Semantic Development, 2016
7. Dumitru Roman, Nikolay Nikolov, Antoine Pultier, Dina Sukhobok, Brian Elvesæter, Arne Berre, Xianglin Ye, Marin Dimitrov, Alex Simov, Momchill Zarev, Rick Moynihan, Bill Roberts, Ivan Berlocher, Seon-Ho Kim, Tony Lee, Amanda Smith, Tom Heath. DataGraft: One-Stop-Shop for Open Data Management . Semantic Web Journal, 2016
8. Enterprise Data Management Council. Financial Industry Business Ontology . 2017
9. EU CEF Digital. Business Registers Interconnection System . 2017
10. EU DG Justice. Interconnection of EU Business Registers , Jun 2017.
11. International Consortium of Investigative Journalists. The Panama Papers : Politicians, Criminals and

the Rogue Industry That Hides Their Cash, 2016.

12. Jonathan Robie, Michael Dyck, Josh Spiegel. XQuery 3.1: An XML Query Language . W3C Recommendation, 21 March 2017
13. LEI Foundation. Introducing the Legal Entity Identifier (LEI) . 2016
14. Mariana Damova, Kiril Simov, Zdravko Tashev, Atanas Kiryakov. FactForge: Data Service or Diversity through Inferred Knowledge over LOD . In Proceedings of AIMSA'2012 Varna, Bulgaria. DOI: 10.1007/978-3-642-33185-5_16
15. NORMA - The Software! The ORM Foundation, accessed 6 August 2017
16. Ontotext Corp. GLEI Mapping to RDF . LEIO Ontology . Jan 2017, last updated Jun 2017.
17. Ontotext Corp. Offshore Leaks as LOD , Data model , RDF download , SPARQL queries , Ontology , Ontology documentation , May 2016.
18. Open Data Bulgaria. Bulgarian Trade Register open data . Created Apr 2016, updated Jul 2017.
19. OpenCorporates. API Reference: version 0.4.6 . 2017
20. Phil Archer, Marios Meimaris, Agisilaos Papantoniou. Registered Organization Vocabulary . W3C Working Group Note, 01 August 2013
21. Steve Harris, Andy Seaborne. SPARQL 1.1 Query Language . W3C Recommendation, 21 March 2013
22. Vladimir Alexiev, Data Visualization with GraphDB and Workbench . Technical report, Ontotext Corp, June 2017. Slides , Video
23. Vladimir Alexiev, Organization Datasets and Ontologies , Presentation at euBusinessGraph Project Kickoff, Oslo, Norway, Jan 2017.
24. Vladimir Alexiev, Tatiana Tarasova, Fredrik Seehusen, David Norheim. euBusinessGraph Semantic Data Model . Working draft 0.9, 7 Aug 2017.
25. Vladimir Alexiev. RDF by Example: rdfpuml for True RDF Diagrams, rdf2rml for R2RML Generation . In Semantic Web in Libraries 2016 (SWIB 16), Bonn, Germany, November 2016. HTML , PDF , Video