ontotext

From the 24th to the 26th of March, the Data Science Society organized a Datathon – the first of its kind data analysis competition in Central and Eastern Europe. The event was held on the grounds of the Software University in Sofia with the support of partner companies and organizations such as Kaufland, Telenor, Experian, HyperScience, ReceiptBank, SAP, ShopUp, A4E, GemSeek, Ontotext, Helecloud, VMware, NSI and Open Government from Council of Ministers.

The Data Science Society team and the partner companies provided various business cases in the field of data science, offering challenges to the participants who set out to solve them in less than 48 hours. At the end of the event, there were 16 teams presenting their results of a weekend of work.

The winning team consisted of Iva Delcheva, Nikolay Petrov, Yasen Kiprov, and Viktor Senderov (who is the author of this blog-post). They worked on two cases: (1) “Hacking” the Bulgarian Commercial Register (a case provided by Ontotext), and (2) Analyzing and exploring data about public procurement in Bulgaria (a case provided by Open Data portal of Bulgarian government). The team decided to join the two datasets together thus generating a Linked Open Data dataset in RDF, which then to query and analyze inside GraphDB.

The Bulgarian Commercial Register is available online as a set of XML files and it covers deeds from 2008 onwards. A deed is a legal term describing the entering into the register of data pertaining to a company such as address, managers, legal status, etc. The data for one company is distributed amongst several deeds and needs to be aggregated. Ontotext recognized the need and offered both a data model for commercial register data and a Java program to RDF-ize the data (see Fig. 1).

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

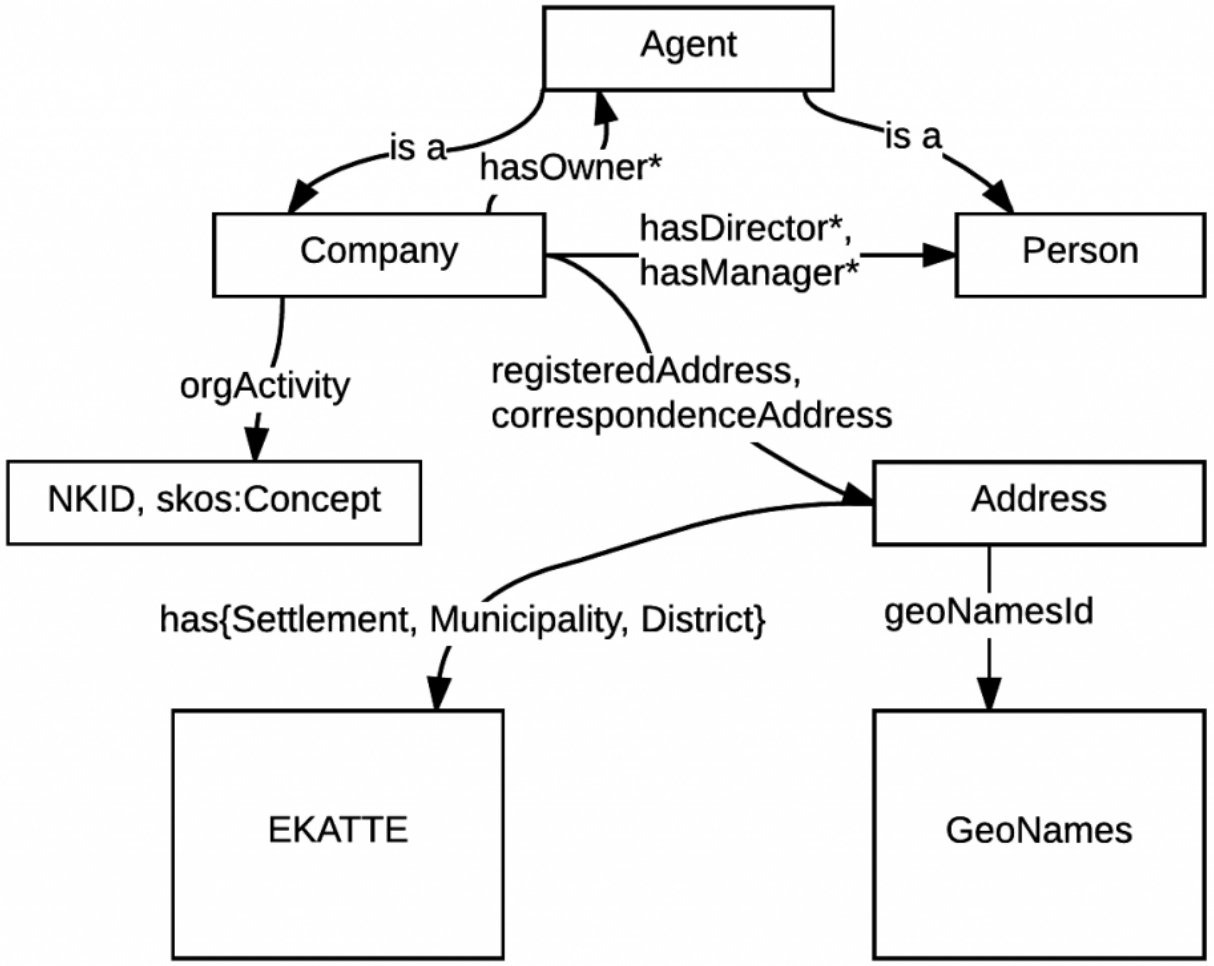

Fig. 1 Simplified data model for the Bulgarian Commercial Register.

In addition to the model, the mentors suggested a scheme for issuing URI's to companies allowing for the easy merging of the data. The identifier each company gets is `:Company_UIC`, where UIC is the Unified Identification Code of the company.

The data for the Bulgarian public procurement from the Public Procurement Agency was made available by Dr. Anton Gerunov. It contains information from 2007 to the middle of 2016. The data is in a CSV format and has columns for the principle and the contractors of

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

the procurement, procurement objective, value in a currency, etc. These entities were modeled by the team as in Fig. 2.

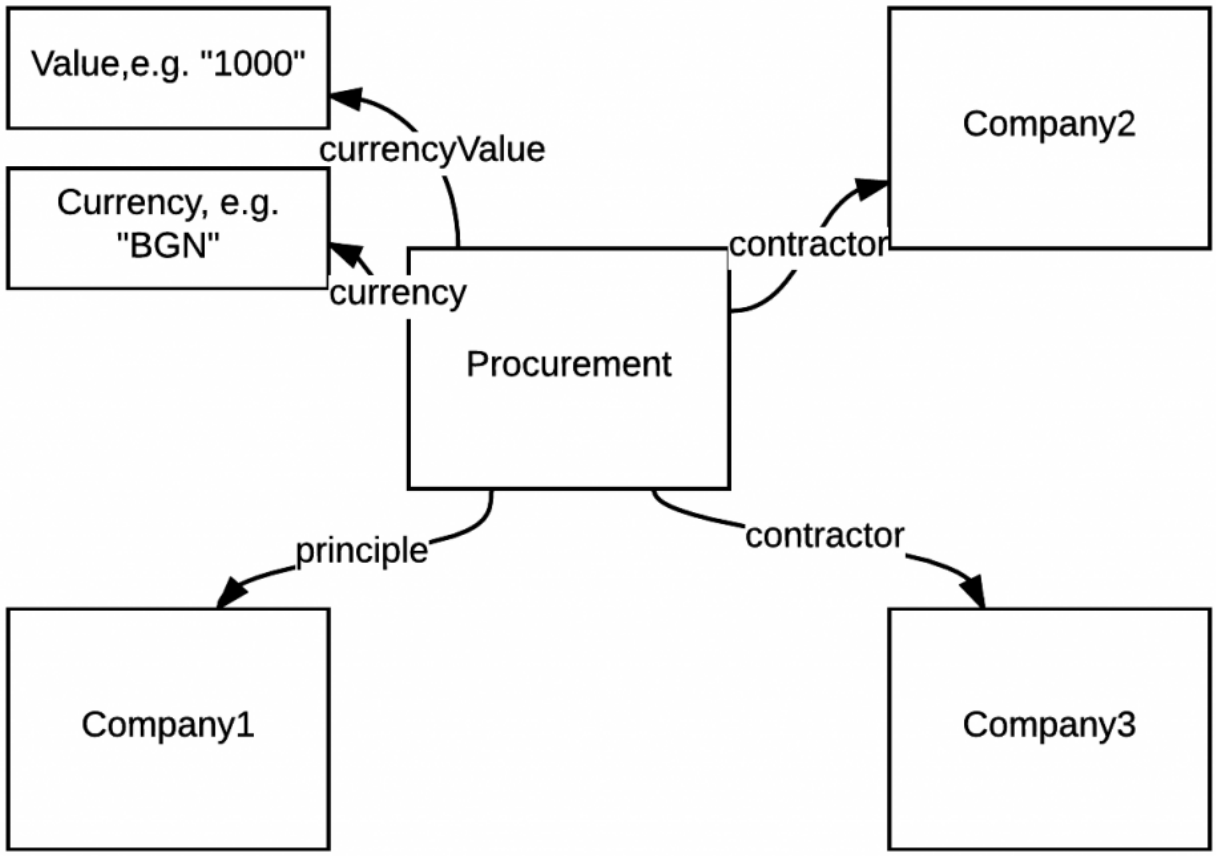

Fig. 2. Simplified data model for the public procurements.

RDF-ization of this CSV set can be done with OpenRefine, which is included in GraphDB, but in this case it was done with a custom-made R script written for a bioinformatics project of Viktor Senderov. Using the same identification scheme for the companies participating in public procurement as we used for the Commercial Register, it is possible to link the two datasets.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

ontotext

In addition to these two major datasets, companies were interlinked to their geo-coordinates by utilizing the Google API.

The resulting RDF dataset, a set of Turtle files, was uploaded to a GraphDB 8 installation running on the Amazon cloud. The size of the uploaded data is approximately 12.5 million triples (more than 2 GB of uncompressed data). The data was not only aggregated and formatted for easy querying, but also connected to previously disconnected information.

One interesting question that can be explored with this linked dataset is the question about conflicts of interest. A conflict of interest may arise if a person A managing a government entity is also a related party (such as, for example, an owner) of a private contractor of the government entity. The SPARQL query answering this question is:

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
prefix : <http://datathon.com#>
SELECT ?p_name ?p ?proc ?c_name ?c ?c2_name ?c2 ?val ?cur
WHERE {
   ?proc a :GovernmentProcurement ;
         :principle ?c ;
         :contractor ?c2 ;
         :currencyValue ?val ;
         :currency ?cur .
    ?c  :hasInfluencingPerson ?p ;
       skos:prefLabel ?c_name .
   ?c2 :hasInfluencingPerson ?p ;
       skos:prefLabel ?c2_name .
   ?p skos:prefLabel ?p_name .
   FILTER (?c != ?c2)
   FILTER NOT EXISTS {
            ?p skos:prefLabel ?p_name .
       FILTER  ( regex( lcase(?p_name),
“(.*община.*)|(.*държавата.*)|(.*министерство.*)” ))
   }
} ORDER BY DESC (?val)
```

The query above returns at least 455 potential issues for a total of 348,468,109 Bulgarian Leva in sectors such as energy and forestry (Fig. 3).

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

ontotext

Filter query results

Showing results from 1 to 455 of at least 455. Query took 15.135 s.

	p	proc	c	c2	val	cur
1	:Person_EGN_B1BB04C464A81590CAA7F75EBB51C0E29AD1D96BA7658E729B20C121001785B7	:Procurement_177029	:Company_106513772	:Company_831577794	99950	BGN
2	:Person_EGN_B1BB04C464A81590CAA7F75EBB51C0E29AD1D96BA7658E729B20C121001785B7	:Procurement_244973	:Company_106513772	:Company_831577794	98900	BGN
3	:Person_EGN_9BD4DEC075549BAD599DD9E64B9653B8FAEF71F84235A0B7976595982596840C	:Procurement_78176	:Company_833017552	:Company_815123066	98635	BGN
4	:Person_EGN_A31AF845AA10AA5419E32B997B8639188647CDDFC377F66635518F2E051B6A4A	:Procurement_192242	:Company_000740495	:Company_201064812	9682	BGN
5	:Person_EGN_B1BB04C464A81590CAA7F75EBB51C0E29AD1D96BA7658E729B20C121001785B7	:Procurement_291996	:Company_106513772	:Company_831577794	96356,79	BGN

Fig. 3. Query results.

In another example, based on the Commercial Register data, one can do a “board walk” (i.e., jump from company to company that share board members) and discover cliques of companies. The result is finding that there are certain individuals on the boards of dozens of companies. Could this be done for tax evasion? Yet another idea is to see what persons are most successful in receiving EU funds. For this last task, one would have to include other publicly available information into the dataset about EU procurements.

We, as data scientists, do not have the legal or commercial expertise to interpret this vast amount of issues. In fact, it may be the case that none of these potential conflict of interest is illegal or even unethical. We do, believe, however, that someone with expertise in legal and commercial matters may benefit from using this linked dataset. Such a person could be an investigative journalist, a public representative or simply a concerned citizen.

To conclude, we would like to note that in all our efforts we have only used previously publicly available data. However, by putting the data into a database we have vastly increased its searchability and usefulness. In our opinion, this is in the interest of society. The effort required for going from loosely-structured public data posted online to Linked Open Data stored in a database is worth it, given the public service.

Want to start revealing relationships and uncovering hidden facts from your data?

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

Datathon Case Overview: Revealing Hidden Links Through Open Data

Author: Viktor Senderov

ontotext

Ontotext's GraphDB
Give it a try today!
Download Now

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com