MPRA

Enhancing Security Exchange Commission Data Sets Querying by Using Ontology Web Language

necula, sabina-cristiana

Alexandru Ioan Cuza University of Iasi

September 2011

Enhancing Security Exchange Commission Data Sets Querying by Using Ontology Web Language

Sabina-Cristiana Necula
Department of Research, Faculty of Economics and Business Administration
Alexandru Ioan Cuza University of Iasi
Iasi, Romania
sabina.mihalache@gmail.com

Abstract. Offering financial information represents a constant preoccupation of governmental and regulatory bodies from all over the world. A large quantity of information to develop semantic web search application is often already available, even if not integrated. In this article, we report our experience in addressing practical computer-based issues from the standpoint of semantics by using Resource Description Framework and Ontology Web Language specifications. We discuss the positive aspects, future work and possible advances.

Keywords: OWL, semantic search, RDF/XML, financial data sets

1   Introduction

The Semantic Web is the extension of the World Wide Web that enables people to share content beyond the boundaries of applications and websites. It has been described in rather different ways: as a utopic vision, as a web of data, or merely as a natural paradigm shift in our daily use of the Web. Most of all, the Semantic Web has inspired and engaged many people to create innovative semantic technologies and applications [1]. The core technological building blocks are now in place and widely available: ontology languages, flexible storage and querying facilities, reasoning engines, etc. Standards and guidelines for best practice are being formulated and disseminated by the World Wide Web Consortium (W3C) [2].

We address in this paper the problem of semantic search for financial data sets. The field of economy and finance is a conceptually rich domain where information is complex, huge in volume and a highly valuable business product by itself [3].

It should be noted that a lot of information important for an organization/company resides in its databases. External information is also highly important. Although the intent to use Semantic Web technologies is useful there are privacy and security problems that determine using Semantic technologies in the presence of an organization portal/intranet. That poses a technology issue because the main problem consists in integrating private data with public data from the semantic point of view.

This paper has 4 Sections. Section 1 presents an introduction. Section 2 contains some aspects related to the current Semantic Web standards addressed by our paper. Section 3 presents with examples the uses of vocabularies, Resource Description Format (RDF), Ontology Web Language (OWL) and SPARQL Protocol and RDF Query Language (SPARQL) for querying financial data sets. Section 4 treats the main conclusions.

2   Semantic Web standards addressed by querying financial data sets

The field of economy and finance is a conceptually rich domain. In seeking to describe the basic categories and relationships of entities and concepts of any financial activity, one can think of creating a knowledgebase of the domain. By using an ontology, one can represent the financial information with all of its complexity and relationships.

Ontologies are shared models of some domain that encode a view which is common to a set of different parties [4]; Contexts are local (where local is intended here to imply not shared) models that encode a party's view of a domain [5,6,7].

TheW3C has defined two languages for the Semantic Web: RDF and OWL. The Resource Description Framework (RDF) [8] plays a basic role by allowing the expression of statements, in the form of subject-predicate-object triples. The Web Ontology Language (OWL) [9] allows the expression of ontologies, which define the meaning of terms used in RDF statements. Simple ontologies can already be expressed using the RDF Schema (RDFS) vocabulary [8].

Although the standard syntax for RDF and OWL uses XML, it should be noted that the meaning of RDF and OWL knowledge bases is independent of XML and abstracts from the XML serialization used. Here the notion of RDF graph [10] plays a role.

In the case of dealing with multiple ontologies, applications also require to integrate such ontologies. A Resource Description Foundation (RDF) vocabulary is a defined set of predicates that can be used in an application. One can define a vocabulary for an application by creating an ontology file, which is an RDF document that contains all possible predicates for an application. An ontology not only defines the predicates themselves, but defines the data type of each predicate and the relationship, if any, of one predicate to another.

RDF vocabularies can describe relationships between vocabulary items from multiple vocabularies that have been developed independently.

Some analysis has been done on the topic of RDF stores which can handle large datasets. (A large dataset in this context is usually considered one on the order of tens or hundreds of millions of triples). The W3C ESW wiki contains information on a variety of RDF stores which can scale to large numbers of triples, but does not speak specifically to the performance of SPARQL queries against these stores.

The performance of a SPARQL query against any particular dataset depends not only upon the size of the dataset but also on the nature of the dataset's storage (a relational store, a native triple store, LDAP, etc.), the complexity of the query itself, optimizations in use by the SPARQL engine, the distribution of the data, and other environmental factors. To date, little work has been done in analyzing SPARQL query performance in particular, and the field of SPARQL query optimization is relatively inchoate.

SQWRL (Semantic Query-Enhanced Web Rule Language) is a SWRL-based language for querying OWL ontologies. It provides SQL-like operations to retrieve knowledge from OWL.

3   Our example
We used Security Exchange Commission data sets available in n3 format at http://www.rdfabout.com/demo/sec/. Besides this data about companies we used some financial data available in Excel files at http://pages.stern.nyu.edu/~adamodar/New_Home_Page/data.html. We wanted to show how we can integrate financial data by making use of Semantic Web technologies.

We developed a vocabulary/ontology for merging data from the two sources. The vocabulary and its namespaces is available at http://sandbox.metadataregistry.org/schemaprop/list/schema_id/49.html and depicted in Figure 1.

Fig. 1. Financial data sets vocabulary

We created a turtle file in order to represent data from the Excel file. Sample content is presented in Figure 2.

```
@prefix dc:         <http://purl.org/dc/elements/1.1/> .
@prefix vcard:      <http://www.w3.org/2001/vcard-rdf/3.0#> .
@prefix                                                 ns:
<http://sandbox.metadataregistry.org/uri/schema/fin> .

:company1
    ns:name     "China Infrastructure Invsmt" ;
    ns:tickersymbol  "CIIC" ;
      ns:industryname "Diversified Co" ;
      ns:SICCode "9913" ;
      ns:xschangeCode "NDQ" ;
      ns:SizeClass "4" ;
      ns:StockPrice "0.69" ;
      ns:TradingVolume "241738" ;
      ns:MarketCap "45.6" ;
      ns:TotalDebt "473" ;
      ns:FirmValue "519.1" ;
      ns:EnterpriseValue "517.8".

:company2
    ns:name     "AFP Imaging Corp" ;
    ns:tickersymbol  "IWKS" ;
      ns:industryname "Medical Services" ;
      ns:SICCode "8000" ;
      ns:xschangeCode "NDQ" ;
      ns:SizeClass "6" ;
      ns:StockPrice "18.5" ;
      ns:TradingVolume "364" ;
      ns:MarketCap "340.6" ;
      ns:TotalDebt "8.6" ;
      ns:FirmValue "349.2" ;
      ns:EnterpriseValue "348.4".
```

Fig. 2. Financial data sets available in Turtle format

We configured Joseki in order to query data. Services and datasets configuration are presented in Figure 3.

```
<#service2>
    rdf:type           joseki:Service ;
    rdfs:label         "SPARQL on the company model" ;
    joseki:serviceRef  "company" ;
    joseki:dataset     <#company> ;
    joseki:processor   joseki:ProcessorSPARQL_FixedDS ;
    .
<#service3>
    rdf:type           joseki:Service ;
    rdfs:label         "sec" ;
    joseki:serviceRef  "sec" ;
    joseki:dataset     <#sec> ;
    joseki:processor   joseki:ProcessorSPARQL_FixedDS ;
    .
## Datasets

<#company>   rdf:type ja:RDFDataset ;
    rdfs:label "company" ;
    ja:defaultGraph
      [ rdfs:label "company.ttl" ;
        a ja:MemoryModel ;
        ja:content                        [ja:externalContent
<file:Data/company.ttl> ] ;
      ] ;
    .
<#sec>   rdf:type ja:RDFDataset ;
    rdfs:label "sec" ;
    ja:defaultGraph
      [ rdfs:label "sec.n3" ;
        a ja:MemoryModel ;
        ja:content [ja:externalContent <file:Data/sec.n3>
] ;
      ] ;
    .
```

Fig. 3. Services and datasets configuration in Joseki

Therefore we were able to address queried on financial data sets. An example of SPARQL query is presented in Figure 4.

```
prefix dc:          <http://purl.org/dc/elements/1.1/>
prefix vcard:       <http://www.w3.org/2001/vcard-rdf/3.0#>
prefix :            <http://example.org/company/>
prefix                                               ns:
<http://sandbox.metadataregistry.org/uri/schema/fin>
PREFIX dc:          <http://purl.org/dc/elements/1.1/>
SELECT ?name
WHERE
  { ?company ns:name ?name }
```

Fig. 4. SPARQL query that returns the name of the companies

In order to observe some comparisons we used Protégé for developing the ontology. We defined from scratch an ontology entitled company and queried instances by using SPARQL query tab available with Protégé. We will not present in detail because the scope of the paper is not to present some differences. Basically the two W3C recommendations base on RDF so we consider that presenting RDF files is enough. We used Protégé for some other important reasons: making rules on data.

Working with financial data often means making some calculus, obtaining the value for some financial rates. We addressed this issue and we found a proposal for SQWRL because with current SPARQL we weren't able to address this problem. An example of SWRL query is presented in Figure 5.

```
Clients(?c) ∧ hasBills(?c, ?b) ∧ hasProducts(?b,
?p) ∧ hasTotalValue(?b, ?v) °
sqwrl:makeSet(?s, ?v) ∧ sqwrl:groupBy(?s, ?p) °
sqwrl:avg(?avg, ?s) →
sqwrl:select(?p, ?avg)
```

Fig. 5. SQWRL query that returns the average value for each sold product

In the narrow sense, the accounting information necessary to make decisions is a metamodel of financial accounting information that provides organized and structured optimally adequate information to management.

Business models and rules of organization and organizing processes contain control statements for decision making.

4   Conclusions

This paper presents with examples querying financial data sets by using Semantic Web technologies. Although not sufficiently treated by semantic web developers there are a lot of opportunities for those who intend to query financial public data sets like governments or companies or financial regulatory bodies.

We present in this paper the necessary configurations that are to be made in order to access a SPARQL endpoint, essentially in building semantic web mash-ups. Our future work will refer to scalability in order that web applications may work on big data sets. We mention that we used for our examples 89523 triples stored in our triple store and that working with bigger data sets represents a problem in accessing data for Joseki server.

5   References

1. Wikipedia, search terms "semantic web"
2. Bishr, Y.: Overcoming the semantic and other barriers to GIS interoperability. International Journal of Geographical Information Science, 12(4):229–314, (1998)
3. Securities Exchange Commission RDF Data, http://www.rdfabout.com/demo/sec/
4. Rapoza, J.: "SPARQL Will Make the Web Shine", eWeek. http://www.eweek.com/article2/0,1759,1965980,00.asp. Retrieved 2007-01-17. (2006)
4. "W3C Semantic Web Activity News - SPARQL is a Recommendation". W3.org. 2008-01-15. http://www.w3.org/blog/SW/2008/01/15/sparql_is_a_recommendation. Retrieved 2009-10-01