
Munich Personal RePEc Archive

Testing the Quality of a Semantic Web Database

necula, sabina-cristiana

Alexandru Ioan Cuza University of Iasi

November 2012

Online at https://mpra.ub.uni-muenchen.de/51556/
MPRA Paper No. 51556, posted 18 Nov 2013 21:13 UTC

Testing the Quality of a Semantic Web Database

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

Abstract

The present paper treats an important aspect which concerns semantic web databases. Our research motivation is given by the problems that semantic web databases raise for the quality characteristic. We try to identify which are the main criteria to take into consideration when building a semantic web database. The research methodology consists in approaching the current literature, discussing the main practical aspects resulted from tests and analysis and treating the problem of quality in respect with a number of characteristics related to the semantic model. The main findings and implications are in the area of testing this quality. We start the paper with an introduction, we continue by presenting the research design, we discuss the semantic model, the main characteristics of a semantic web database in terms of quality and testing.

Keywords: testing, quality, semantic model, semantic web, RDF

Introduction

Semantic web represents a particular research topic derived from Artificial Intelligence. But semantic web is not artificial intelligence. Tim-Berners Lee is the first one who noted this (Tim-Berners Lee, 2006). It is about providing links to different web sources respecting the semantic model represented by a standard representation format which is Resource Description Format (RDF). In the last years, semantic web concretizes in the form of Linked Open Data which is an entire movement in the practical semantic web field. Otherwise saying, given the fact that the semantic web community developed common standard formats for representing web sources it had to take a real form which had been proved to be Linked Data.

The semantic model relates to ontologies which are a standard conceptualization for knowledge. Knowledge representation consisted in many forms and represented a long debate for discussing both the concepts of knowledge and model.

In computer science and information science, an ontology formally represents knowledge as a set of concepts within a domain, and the relationships among those concepts. It can be used to reason about the entities within that domain and may be used to describe the domain.

In theory, an ontology is a "formal, explicit specification of a shared conceptualisation"(Gruber, T.R., 1993). An ontology renders shared vocabulary and taxonomy which models a domain with the definition of objects and/or concepts and their properties and relations (Arvidsson, F., Flycht-Eriksson, A., 2008).

Ontologies are the structural frameworks for organizing information and are used in artificial intelligence, the Semantic Web, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture as a form of knowledge representation about the world or some part of it.

The standard formalism to represent ontologies for Semantic Web is Resource Description Format (RDF) which represents web sources as triples in the form of subject-predicate-object.

The RDF data model (W3C) is similar to classic conceptual modeling approaches such as entity-relationship or class diagrams, as it is based upon the idea of making statements about resources (in particular Web resources) in the form of subject-predicate-object expressions. These expressions are known as triples in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object.

Literature review

The semantic web field of research is relatively new for the application field. Also, the main semantic web technologies are in their beginning phase of development. Although, there are a lot of studies, many of them not being necessary related to the Linked Open Data preoccupations.

We can classify the current studies related to our research motivation by:

- Studies related with the problem of establishing the relationships between web data sources, concepts or entities from the semantic model (Tran, T, et.al., 2010; Halpin, H, et.al., 2010; Dang, L., et. al., 2010)

- Studies related with the problem of querying web data sources (Yingjre, L., Heflin, J., 2010; Ladwing, G., Tran, T., 2010; Wylot, M., et.al., 2011; Le-Phuoc, D., et. al., 2011)

- Studies related to RDF data analysis (Yu, Y., Heflin, J., 2011; Kharlamov, E., Zheleznyakov, D., 2011)

- Studies concerning the problem of ontology matching and ontology mapping (Jimenez-Ruiz, Cuenca Grau, B., 2011; Cheng, G., et. al., 2011)

The main organism concerned with semantic web applications is World Wide Web Consortium. They develop standards, there is a big community of developers which created tools, test big data sets, create semantic web applications and try to promote this research area.

Research design

We relate or discussion by approaching a number of important aspects concerning the quality of a semantic web database.

In this respect we consider necessary to discuss the size of the database, the number of established relationships, the number of namespaces and the number of URIs.

We will relate the abve identified variables with the current developed semantic web applications and with the characteristics which concern the semantic model in order to sustain our idea. In this way we consider that we can improve the current studies by introducing some test variables.

Our research design model is presented in Figure 1.

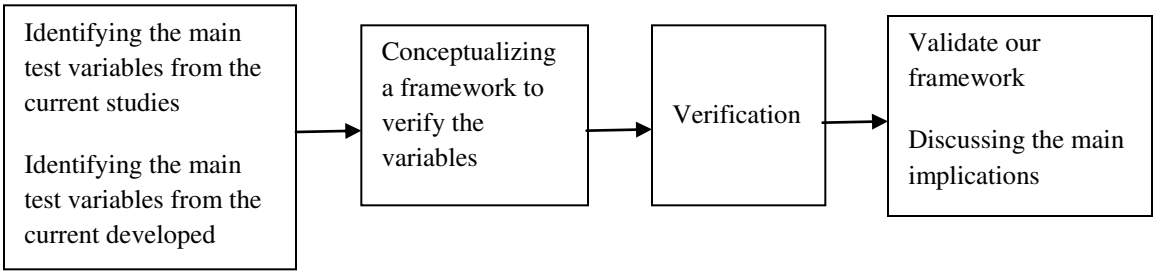

Figure 1. The research model

Starting from this research model we will section our article in some important parts which respect the research design model.

Current test variables related to semantic web databases

From the current studies we found out that testing semantic web databases is very often treated by addressing the relational databases. In database testing, the following issues need to be considered: 1) Atomicity; 2) Consistency; 3) Isolation; 4) Durability; 5) Integrity; 6) Execution of triggers and 7) Recovery.

The semantic web data model is very directly connected with the model of relational databases. A relational database consists of tables, which consists of rows, or records. Each record consists of a set of fields. The record is nothing but the content of its fields, just as an RDF node is nothing but the connections: the property values. The mapping is very direct:

- a record is an RDF node;
- the field (column) name is RDF propertyType;
- the record field (table cell) is a value.

RDB systems have datatypes at the atomic (unstructured) level, as RDF and XML will/do. Combination rules tend in RDBs to be loosely enforced, in that a query can join tables by any columns which match by datatype - without any check on the semantics.

Tim Berners-Lee outlined four principles of linked data in his Design Issues: Linked Data note, paraphrased along the following lines: (Berners-Lee, T., 2006)

1. Use URIs to identify things.

2. Use HTTP URIs so that these things can be referred to and looked up ("dereferenced") by people and user agents.

3. Provide useful information about the thing when its URI is dereferenced, using standard formats such as RDF/XML.

4. Include links to other, related URIs in the exposed data to improve discovery of other related information on the Web.

Defining a framework to verify the variables

Given the fact that in semantic web every node can be related to multiple nodes from the entire www it is normal to consider that those issues related with database testing are more intriguing for semantic databases.

The entire set of issues related to testing databases for the semantic web databases is addressed when querying data. The query language for semantic web is SPARQL.

It becomes clear that in order to analyze data those data must be queried before.

Some issues specific to the semantic model arises: which is the size of an RDF file that semantic web base applications can support? How much do scale the semantic web based applications? How can be controlled the number of URIs? How can be controlled the number of namespaces?

Semantic web needs high processing speed and for this not only the processing speed of the CPU is useful but the speed to access data is also very important.

Therefore there is a need that the semantic database be available on memory storages which offer fast data access.

For this, from the existent memory storages the most suitable to assure fast data access is the Random Access Memory (RAM). The most common RAM capacity existent on the market is of 8 GB. This means that the models which can be loaded by this dispositive cannot be bigger than 8 GB. We observe a limitation on the performance of semantic database.

Given this limitation any semantic web application developer will look to assure the horizontal scalability.

The horizontal scalability raises the problem of partitioning a big data set on multiple systems which can be accessed in parallel. In order to realize this there is a need of a data management system which can assure the sharing of files so that any request on data be routed directly to the system which has the most relevant data to answer to the queries.

But the horizontal scaling not only requires a semantic data management system. It is also a problem of partitioning namespaces because every URI is described by different namespaces. We will discuss the semantic model in the next subsection to understand better the problem of semantic relations.

In this way we define the following test variables:

- The number of URIs;
- The number of namespaces;
- The size of the RDF data file.

We consider that these test variables are definitely important when developing a semantic web based application.

Validate our framework

We must relate our discussions with the main semantic web applications currently developed and to semantic web use cases. A list of these is available at http://www.w3.org/2001/sw/sweo/public/UseCases/ . For the moment there are 33 Case studies and 13 use cases.

In terms of the number of URIs every use case is an example of using relatevly big numbers of URIs. Every use case relates to a small number of namespaces given the discussed problems of scalability.

Conclusions

As semantic web base applications evolve one can expect that common standards for testing semantic web based applications appear. Of course, the principal test is represented by their very use. Up to that moment, every developer must consider useful a framework of set conditions to verify in order to develop appropriate semantic web based applications. In this moment, it seems that there are a lot of semantic web based applications that are not used because they do not fit to user expectations or need.

We consider that our framework is important in assuring semantic web databases quality.

Or framework can present implications for the theory field and for the practical field.

Acknowledgment

This work was supported by CNCSIS-UEFISCSU, project number PN II-RU code 188/2010.

References

Gruber, Thomas R. (June 1993). "A translation approach to portable ontology specifications" (PDF). *Knowledge Acquisition* **5** (2): 199–220.

Arvidsson, F.; Flycht-Eriksson, A.. "Ontologies I" (PDF). Retrieved 26 November 2008.

http://www.w3.org/TR/PR-rdf-syntax/ "Resource Description Framework (RDF) Model and Syntax Specification"

Thanh Tran, *Lei Zhang*, *Rudi Studer*, **Summary Models for Routing Keywords to Linked Data Sources**, 2010, ISWC Conference

Li Ding, *Joshua Shinavier*, *Zhenning Shangguan*, *Deborah L. Mcguinness*, **SameAs Networks and Beyond: Analyzing Deployment Status and Implications of owl:sameAs in Linked Data**, 2010, ISWC Conference

Harry Halpin, *Patrick J. Hayes*, *James P. McCusker*, *Deborah L. Mcguinness*, *Henry S. Thompson*, **When owl:sameAs isn't the Same: An Analysis of Identity in Linked Data**, 2010, ISWC Conference

Yingjie Li, *Jeff Heflin* , **Using Reformulation Trees to Optimize Queries over Distributed Heterogeneous Sources**, 2010, ISWC Conference

Marcin Wylot, Jigé Pont, Mariusz Wisniewski and Philippe Cudré-Mauroux, dipLODocus[RDF]--Short and Long-Tail RDF Analytics for Massive WebsofData, 2011, ISWC Conference

Danh Le-Phuoc, Minh Dao-Tran, Josiane Xavier Parreira and Manfred Hauswirth, A Native and Adaptive Approach for Unified Processing of Linked Streams and Linked Data, 2011, ISWC Conference

Yang Yu and Jeff Heflin, Extending Functional Dependency to Detect Abnormal Data in RDF Graphs, 2011, ISWC Conference

Evgeny Kharlamov and Dmitriy Zheleznyakov, Capturing Instance Level Ontology Evolution for DL-Lite, 2011, ISWC Conference

Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau, LogMap: Logic-based and Scalable Ontology Matching, 2011, ISWC Conference

Gong Cheng, Saisai Gong and Yuzhong Qu, An Empirical Study of Vocabulary Relatedness and Its Application to Recommender Systems, 2011, ISWC Conference