Universitá degli Studi di Milano Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di Laurea in Informatica

Publishing Linked Data in e-Commerce: Designing and Testing

(Summary)

Supervisor:
Dott. Matteo Palmonari

Co-supervisor:
Dott. Riccardo Porrini

Thesis presented by:
Davide Brando Preda
Matricola: 735850
Phone: 3498938931
Mail: b.preda91@gmail.com

2012-2013

31/10/2013

The purpose of the Semantic Web is to introduce structured data beside the traditional web composed by documents, in order to support the machines interpreting the available information. The idea of Tim Berners-Lee, the inventor of the Semantic Web, is about an "intelligent" web. For this reason, parallel to the web composed by documents, the Semantic Web has been developed. The Semantic Web allowed the application to provide features based on structured data. For example, search engines process queries in the form of keywords sets to show documents found on the web related to those queries; while with the Semantic Web it is possible for the users to pose factual queries on structured data which produce precise results. For example, keywords that the user consider relevant like "Woman sportive shoes Mizuno" are submitted in a search engine it could be possible to obtain only partial results. Instead, factual complex queries can be posed to a structured knowledge base specifying e.g. brand, type, color, texture, etc..

The Semantic Web is based on the presence of structured data and the existence of shared vocabularies to be interpreted. However, one of the biggest problem encountered so far in the context of Semantic Web has been the difficulty of finding publicly available data. This has happened because the great part of data on the web is unstructured. According to this, the invitation of Tim Berners-Lee made to the users of the web[1], is precisely that to put data online. According to that, data can be used to create application based on integration and analysis of information. From that call on there has been an exponential increase in the number of published data on the web.

One of the difference between the Web and the Semantic Web is the presence of structured data. The data itself have a certain relevance, but once integrated, can be combined and used to create more interesting application scenarios than the individual parts. In order to integrate web data a common standard of data structuring is required. In this direction moved the Semantic Web community, defining some standard languages for publishing, structuring and querying data (RDF[2], RDFS[3], OWL[4], SPARQL[5]). The structured data following those standards are called *Linked Data*. The innovative characteristics of Lined Data come not only from the amount of data, but also from the ability to connect data with each other in relation to the links binding them. When data are connected to each other it is possible to enable fruition mechanisms and data analysis which can not be achieved in the traditional web composed by documents. Furthermore, structured data can be used to provide better answer to queries and to return accurate information. Moreover obtained any information would be provided as Linked Data.

[1] http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html
[2] http://www.w3.org/TR/2004/REC-rdf-primer-20040210/
[3] http://www.w3.org/TR/rdf-schema/
[4] http://www.w3.org/TR/owl2-primer/
[5] http://www.w3.org/TR/rdf-sparql-query/

Within this project a model for the semantic representation of consumer products under the E-Commerce has been defined and it has been tested publishing a product catalog, existing as relational database. An ontology to represent the structure of the catalog has been designed. A mapping between the database elements and the ontology concepts has been defined to publish the catalog as Linked Data. Finally, a testing has been conducted to prove that the SPARQL language queries made through the Linked Data have reasonable response times in relation to the corresponding SQL queries, and that are syntactically easier to write.

It has been modelled in the form of Linked Data a catalog of products represented in a real relational database. It has been made possible to query the catalog using the SPARQL language that allows to pose articulated queries. Also, it has been made possible to browse through the catalog by following the links and the semantic relations that bind the different products. SPARQL allows to pose articulated and meaningful queries. Using Linked Data and the appropriate RDF language it was possible to create structured contents that can be used to enrich product catalog with semantic descriptions. A contribution for application that need to perform semantic disambiguation or *Named Entity Recognition* is provided, in order to permit these applications to use structured information and not only text.

Linked Data in E-Commerce can be used to support disambiguation in searching consumer product. In fact, when a user examines a catalog and pose a query targeting specific products with specific characteristics and thus results should be less ambiguous as possible. In this project it has been worked within the E-Commerce and the database used in the project as a source is a real database belonging to the company 7Pixel[6], which owns price comparison engine sites Shoppydoo[7] and TrovaPrezzi[8]. The database is private and not accessible from the web and it contains a product catalog consisting of several tables, inside of which there are more than 217000 products. Each product has more than one offer of several vendors, each with its own selling price. The advantage provided by the presence of a rich product catalog is that, once an offer is associated to the corresponding product, it automatically inherits the technical characteristics.

Some E-Commerce companies began to add semantic annotations to their catalogs (e.g. by using GoodRelations[9] and Schema.org[10] vocabularies). However, in the context of this project, these vocabularies were not considered comprehensive and expressive enough, so a special one has been defined. We are not aware of other projects in which it has been defined a dedicated vocabulary for semantic

[6] www.7pixel.it
[7] www.shoppydoo.com
[8] www.trovaprezzi.it
[9] http://www.heppnetz.de/projects/goodrelations/
[10] http://schema.org/

description of the products. Moreover we are not aware of other projects that have treated and published an amount of data like the one in 7Pixel's catalog.

Therefore, the goals of this project are:

- to define an ontology to represent the product catalog structure;
- to publish a product catalog, starting from its relational representation, as a Linked Data knowledge base ;
- to enable the execution of detailed and significant queries on the basis of knowledge, using the SPARQL syntax;
- to give the possibility to browse through the catalog following semantic relationships;
- to make sure that the published knowledge base will be both dynamic and automatically adaptive according to the changes made in the catalog.

It was necessary to define an ontology with the aim to define in a formal way the vocabulary used. In order to structure the ontology and to comply with the standards for the publication of Linked Data the RDFS[11] language it has been used. The goal was to ensure maximum expressiveness in the product description. For this reason, the main concepts (products, categories, brands, etc...) and the main relations (features, sub-categories, etc. ..) between them are represented in the ontology.

In order to expose the catalog as Linked Data, the D2R-Server tool has been employed in order to map the relational database elements to the ontology concepts. Before deciding which tool to use, other existing tools that map automatically or semi-automatically relational schemes to RDF were evaluated, in order to choose the most suitable and mature one. In particular, it has been chosen a tool that would allow the mapping of the data in a read-only way, because it was not relevant, in the context of the project, to write directly on the database. The goal was to ensure maximum expressiveness in the descriptions, and the read-only tools ensured a higher expressiveness, so it was decided to use one belonging to the *Read-only general-purpose* category. Specifically, the D2R-Server tool was used, which supports the D2RQ language, since at the time the project was the most mature and comprehensive tool. Although, the W3C[12] adopted as a standard for mapping from RDB to RDF the R2RML[13] language. However, when the project was started it was not yet available a tool that supported the R2RML language.

[11] http://www.w3.org/TR/rdf-schema/
[12] http://www.w3.org/
[13] http://www.w3.org/TR/r2rml/

Thus has been opted for the D2RQ language with a fully supporting tool. In addition, the D2R-Server development team is currently working on the tool to make the language D2RQ compliant with W3C standards.

A SPARQL endpoint integrated into the tool D2R has been used to make possible posing SPARQL queries to the knowledge base exposed in the form of Linked Data. The SPARQL endpoint gets requests through the HTTP protocol and returns structured information. D2R translate SPARQL queries in SQL queries before submitting them to the relational database. The translation is done by exploiting the mapping previously defined that specifically map the rules of the relational database to ontological concepts. In this way, it is possible to take advantage of the simplicity of the SPARQL language that allows to perform detailed queries.

The D2R tool allows to automatically generate HTML pages through which it is possible to browse the catalog following the semantic links that connect the various products. For example, while viewing a product, all its characteristics are browsable. Moreover, following the link to one of those additional information of the characteristic can be obtained or it is possible to view the other products that have that characteristic.

The catalog is dynamically updated with respect to the data in the relational database. This is why has been chosen to use the virtual publication instead of migrating the data. In this way all changes that are made to the database will automatically reflect on what is published. Virtual publication allowed to avoid keeping multiple copies of the same data on the machine.

Finally, on 7Pixel database, an analysis of the queries was conducted in terms of time and expressiveness. The testing was conducted by posing SPARQL queries on the knowledge base, and SQL queries on the database used as source. Different type and complexity queries were posed, and the data obtained are the result of the average of several tests performed on the individual queries. The testing phase demonstrated compliance with the following constraints:

- response times of SPARQL queries must be reasonable compared to the corresponding SQL queries;
- the complexity and length of the SPARQL queries must be reasonable compared to the corresponding SQL queries.

Experimental results showed that, although the SPARQL queries require more time compared to the corresponding SQL queries, this discrepancy is not high and can be regarded as reasonable. In fact, it was found that SPARQL queries take an average of 9% more time than the same queries formulated in the SQL

language. The overhead is inevitable because the SPARQL queries must first be translated into SQL, in order to be submitted to the database. However it is a very reasonable overhead whereas the average response times are in the order of seconds.

In addition, SPARQL queries, not only have a reasonable complexity compared to SQL queries, but also that they can be written with a syntax more simple in terms of characters. The results obtained through SPARQL queries performed on the database exposed as Linked Data, can also be made directly on the relational database using the SQL language, but conducting a comparative analysis between the two types of queries, it has been proved that it is more simple to write SPARQL queries. In fact, the evaluation showed that the length of the number of characters decreases, on average, of 86%. Furthermore, it is not needed to make any joins between the tables because they are all implicitly contained within the mapping.

Finally, the contribution given by this project consisted in designing an ontology for the publication of a product catalog. Starting from a relational schemes representation of the database, a knowledge base structured as Linked Data has been defined. A testing has been done on processable queries posed to the knowledge base, from which interesting conclusions about response times and semantic complexity have been taken.