Carlos Montoya
Rauf Agayev
Vuqar Agayev

PROBLEM DEFINITION

During the development of Web Semantic technologies huge volume of data has being published on the web as a Linked Open Data. In order to use LOD for specific purpose, we should check its quality in advance. The quality term means in Semantic Web is fitness of use. Then we should have into account that not all the data is meaningful and if the quality is poor it leads to Enterprise problems such as data standardization, multiple data with duplicates in the data sets, meaningless information and so on. To assure that the Enterprise is trustful, then in the design we should take into account that the quality measurement depends on which domain the data will be used. On the other hand, we should assure that our LOD the data is measure for the correct quality metrics, which means that the given data in one LOD be useful for one of the use cases but not favorable for other ones. Therefore, for measurement of LOD we are going to use metrics and we identify the fitness of it use.

USE CASE 1

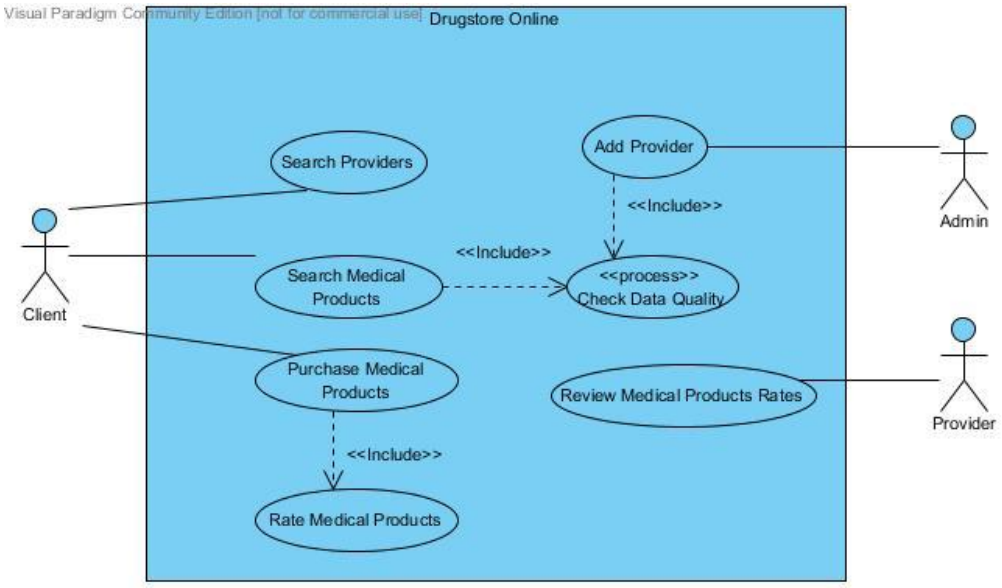
Based on the fact that there is some people who cannot go out because they suffer from some illness or they are recovering at home of some kind of medical treatment. Then they need a service that provide cheaper and trustful medicaments online that can be send easily to their home. For that reason the use case is related with the creation of a Drugstore Online which obtain information of different websites (providers) and then shows the best options to the user. Of course in order to offer a medicament or a treatment we must be able to check the quality of the data that we display on the screen.

We must be able to check in advance the data quality, for that reason we should be able to use the metrics (initial approximation):

1. Verifiability
2. Reputation
3. Believability
4. Consistency
5. Availability
6. Understandability
7. Conciseness
8. Volatility

The next image describes the basic uses cases that the enterprise system require.

Carlos Montoya
Rauf Agayev
Vuqar Agayev



USE CASE 2

During the improvement of information technologies, approximately every person uses Computer, Tablet, Telephone, etc. In addition, they prefer, buy such kind of devices online for getting lower prices. Our use case related with this situation. We want to create Online Computer Store, which gets information about Computers from different web sites and shows customers. In order to achieve this, we should check quality of data in advance. On the other hand, we apply some metrics which fit our requirements, to the given data and then we decide data is usable or not.
Metrics:

1. Completeness
2. Amount of Data
3. Relevancy
4. Verifiability
5. Reputation
6. Believability
7. Response time
8. Availability
9. Understandability

USE CASE 3

My use case is about intelligent hotel reservation system which aggregating data from several data sources, websites. It gets information about location of countries, cities and particular addresses from a spatial dataset and information about hotels from a hotel dataset. Additionally, all related information about hotels gathered from different booking services and represented as RDF. This integrated dataset allows user to find a suitable hotel between any arrival and departure time for his/her wonderful vacation.

Carlos Montoya **EIS Lab SS2014**
Rauf Agayev
Vuqar Agayev

The use case that we are going to develop is the related with Check Data Quality.

1. Availability
2. Licensing
3. Interlinking
4. Security
5. Performance
6. Accuracy
7. Consistency
8. Conciseness
9. Reputation
10. Believability
11. Verifiability
12. Objectivity
13. Understandability
14. Versatility

METRICS DEFINITION

The metrics for the chosen dimensions are:

- Verifiability
  - a Authenticity of the dataset
  - b Usage of digital signatures
  - c Correctness of the dataset
- Reputation
  - a Reputation of the dataset
- Believability
  - a Meta-information about the information provider
  - b Indication of metadata about a dataset
  - c Computing the trustworthiness of RDF statements
  - d Computing the trust of an entity
  - e Accuracy of computing the trust between two entities
  - f Acquiring content trust from users
  - g Assigning trust values to data/sources/rules
  - h Determining trust value for data
  - i Computing personalized trust recommendations
  - j Detection of reliability and credibility of a data source
  - k Computing the trustworthiness of RDF statements
  - l Detect the reliability and credibility of the dataset publisher
- Consistency
  - a Re-use existing terms
  - b Re-use existing vocabularies
- Availability
  - a Accessibility of the SPARQL end point and the server
  - b Accessibility of the RDF dumps
  - c Dereferencability issues

- d No structure data available
- e No dereferenced back-links
- f No dereferenced forwards-links
- g Misreported content-types

Understandability
- a Human-readable labeling of classes, properties and entities
- b Dereferenced representations: providing human-readable metadata
- c Indication of one or more exemplary URIs
- d Indication of a regular expression that matches the URIs of a dataset
- e Indication of an exemplary SPARQL query
- f Indication of the vocabularies used in the dataset
- g Provision of message boards and mailing lists

Conciseness
- a Keeping URIs short
- b No use of prolix RDF features

Volatility
- a Frequency of change
- b Time validity interval

Completeness
- a Schema Completeness, Property Completeness, Interlinking Complteness
- b Number of Classes, Interlinkings, Properties, Values present in dataset compared to ideal

METRICS FOR THE NEW DIMENSIONS

Easy of Manipulation
- a Keeping URIs short
- b Mixed metrics between believable and reputable dimensions

Free of Error
- a Dereferencability issues
- b Erroneous annotation/representation erroneous
- c Inaccurate annotation, labeling, classification

Value-Added
- a Evaluation by the expert users

Tepmoral Relatability
- a Meaning and semantics changes over time
- b Look for loss of history data with no record of previous values

REQUREMENT SPECIFICATION

Requirement Name	Process Data Quality		
Main Users	System and Admin	**Priority**	High
Author	Carlos Montoya	**Creation Date**	14-05-204

Carlos Montoya
Rauf Agayev
Vuqar Agayev

Revision Date		**Modifications Date**	
Summary	The data that display the main system of the company, should be check in advance, this use case is in charge to measure the data sets quality, by measuring the dimensions of: Free of Error, Verifiability and XXXX After verify those metrics, the system should be able to display the values obtained and the user should be able to compare those values between each other.		
Main Track Events			
User		**System**	
1. The user access the program.		2. The system display the name of the dimensions to be evaluated, under the question "Please chose a Dimension", those are shown as a combo box, and the options are: 1st Verifiability, 2nd Free of Error and 3rd Measurability	
3. The user select one of the dimension displayed.		4. Depending of the option selected, the system should display a new combo box with the metrics to be evaluated. The system have three different cases (depending on the dimension selected).	
		4.1 When the user choose Verifiability, then the system show the metrics: 1st Authenticity of the dataset, 2nd Usage of digital signatures.	
		4.2 When the user choose Free of Error, then the system show the metrics: 1st Dereferencability issues, 2nd Erroneous annotation/representation erroneous.	
		4.3 When the user choose Measurability, then the system show the metrics:	
5. The user select one of the metrics displays in the point 4.		6. The system now show the different dataset available to be evaluated under those metrics. They show the options under a multiple select box. The datasets are: DATASET1, DATASET2, and DATASET3.	
		7. When the user select at least one dataset the system display the bottom "evaluate"	
8. The user press the button "evaluate".		9. The system should display a graph where is easy for the user to identify the metrics vs the datasets. And each dataset is easy to identify.	
Alternatives Tracks			
Exeption Tracks			

After the user in the point 8 select the information for evaluate and the info of the evaluation is not available, then the system should display a message "The information is not available right now, please choose another dataset or another metric."
Extension Points
Pre - Conditions
To display easily and fast the information of the results of the evaluation of the metrics, those values (metric values) should be evaluated in advance and those result should be saved into some media that should be defined by the development group.
The selected datasets to be evaluated should be free of access, then the development is not going to violate any copyright of the information used.
Post- Conditions
All the datasets are evaluated and these results are store in an accessible media to be used by another use cases.
Considerations