Metadata quality: Generating SHACL rules from UML class diagrams

Presentation

Emidio Stani
PwC, Belgium
emidio.stani@pwc.com

Keywords: model; metadata; RDF; validation; SHACL; quality; rules; Eclipse; UML

Abstract

Metadata plays a fundamental role beyond classified data, as data needs to be transformed, integrated, and transmitted. Like data, metadata needs to be harvested, standardized and validated. Metadata management processes require resources. The challenge for organizations is to make the processes more efficient, while maintaining and even increasing confidence in their data.

While RDF harvesting has already become an important step, implemented at large scale[1], there is now a need to introduce a RDF validation mechanism. However such a mechanism will depend upon the definition of RDF standards. When a standard is set, the provision of a validation service is necessary to determine if metadata complies, as for example with the HTML validation service. For example, DCAT-AP is used to describe public sector datasets in Europe; an online DCAT-AP validator[2] provides a way to validate DCAT-AP datasets.

When an organization wants to provide an RDF validation service, there are key considerations to take into account, notably the possibility for the user:

- to provide metadata to be validated in any RDF serialization, as metadata can be generated from different sources;
- to obtain the list of violations according to their severity/quality scores, allowing the user to address the most important in priority when fixing the metadata validated;
- to receive a message describing the violated rule as users might not be familiar with SPARQL or SHACL;
- and to provide/see the validation rules.

In addition, organizations have to continuously review these rules which in turn depend on the model. Thus organizations need to synchronize the rules with the current model.

Such requirements would be easily met by generating the rules automatically in order to make this process less error prone and more efficient.

A Model Driven mechanism which generates rules out of a model, is therefore a good practice since changes can be applied directly in the model and rules can automatically be generated. This approach is already a well-used technique, especially for Object Oriented Applications for models serialized, such as XML schema.

The proposed presentation will show one method to use model driven mechanism to generate automatically violations rules. Using tools for model design and model to text functions like Papyrus and Acceleo based on Eclipse, it is possible to generate SHACL constraints. A UML class diagram with stereotypes is used to describe the original metadata. Thanks to the UML

[1] **European Data Portal**, European Union. https://www.europeandataportal.eu/ Accessed on 27/08/2018
[2] **DCAT-AP validator**, Open Data Support. http://dcat-ap.semic.eu/dcat-ap_validator.html Accessed on 27/08/2018

https://doi.org/10.23106/dcmi.952139134

stereotypes, one can then generate automatically SHACL constraints that could be then used by a SHACL validator. The Flemish Government has implemented a similar method by using other tools[3] and publishing an online validator called OSLO2 validator[4].

[3] **OSLO-EA-to-RDF**, Flemish Government. https://github.com/Informatievlaanderen/OSLO-EA-to-RDF Accessed on 27/08/2018

[4] **OSLO2 validator**, Flemish Government. https://data.vlaanderen.be/shacl-validator/ Accessed on 27/08/2018

https://doi.org/10.23106/dcmi.952139134