Metadata quality

Generating SHACL rules from UML class diagrams

Emidio Stani

11/09/2018
Agenda

1. Introduction (5 min)
2. Approach (15 min)
3. Conclusions (5 min)
Introduction
Introduction
The cost of data quality

The challenge for organizations is to make their processes more efficient, while maintaining and even increasing confidence in their data.

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.

Poor quality data is simply costly to any organization. Estimates differ, but experts think organizations spend between 10-30% of revenue handling data quality issues. IBM estimated the cost of poor quality data in the US in 2016 was $3.1 trillions.

[DAMA-DMBOK]
Introduction
The need for an RDF validation service

RDF harvesting has already become an important step implemented at large scale.

This is possible only by defining a standard on which a validation service can be built as violation rules will depend on the model.

For example DCAT-AP is used to describe public sector datasets in Europe published by several open data portals.

The European Data Portal provides access to 77 harvested catalogues for a total of 861038 datasets.

Catalogues and datasets are described accordingly to the DCAT-AP standard and their metadata is monitored.
Approach
Approach

-Design and implementation of a method to automatically generate constraints from UML class diagram.

-Constraints are defined in SHACL, the W3C standard to define constraints in linked data world.

-Such constraints can be reused by a validator to test against data.

The INSPIRE European directive and the data specifications are based on a UML model of more than 1225 classes (just the themes).

The OSLO2 (Flemish Government) data model is represented as UML model of 748 classes counting more than 500 SHACL shapes.
**Approach**

**In practice**

- **Definition of UML stereotypes** subclassing the concept of class and attributes using Papyrus UML (Eclipse based, OMG compliant);

- Application of stereotypes in a UML diagram on explicit **classes and attributes** so that the method is not intrusive;

- **Creating an application** to perform the generation of SHACL constraints, using Acceleo language (Eclipse based);

- **Validation** of the SHACL constraints

On line validation could be performed using e.g. SHACL playground [http://shacl.org/playground/](http://shacl.org/playground/) or the SHACL (OSLO2) validation service of the Flemish Government [https://data.vlaanderen.be/shacl-validator/](https://data.vlaanderen.be/shacl-validator/).
Approach
Defining UML stereotypes
Approach
Implementation

Papyrus UML
- UML metamodel
  - UML model

Acceleo
- Application
  - Acceleo library

Eclipse

Config.properties
- Get Config.java
- util.mtl
- generateShapes.mtl

shacl_stereotype_shape = isa:shaclShape
shacl_stereotype_constraint = isa:shaclConstraint
... shacl_default_severity = @sh:Violation

output

```xml
forms:FrmAbsenceOfCriminalRecordShape
def: type sh:NodeShape;
sh:name "FrmAbsenceOfCriminalRecord Shape" fn;
sh:targetClass forms:FrmAbsenceOfCriminalRecord;
sh:property [
  sh:path forms:FormHeader;
  sh:name "Header";
  sh:mincount 1;
  sh:severity sh:Info;
];
sh:property [
  sh:path forms:AbsenceOfCriminalRecord;
  sh:name "Body";
  sh:minCount 1;
  sh:maxCount 1;
  sh:severity sh:Violation.
];
```
Approach

Implementation

UML model

SHACL rules

Violation File 1

Violation File 2

Violation File 3

JSON file

http://localhost:80/80/violation1.ttl

http://localhost:80/80/violation2.ttl

http://localhost:80/80/violation3.ttl

SHACL validator

OSLO² applicatie profiel

Data

Valideer

Klik hier om te valideren
Conclusions
Conclusions

Benefits

1) Reusability of the UML model

The creator of a model and data architect share the same model at base

2) Data quality

Data quality process is less error prone

3) Enhancing trust of data coming from third parties

Third parties might take advantage of the constraints when integrating their system
Conclusions

Questions ?