A Framework to Include and Exploit Probabilistic Information in SHACL Validation Reports

Rémi Felin, Catherine Faron and Andrea G. B. Tettamanzi

ESWC 2023 - 2023/06/01
Introduction

- Evaluation of RDF graphs against domain constraints
- SHACL, the SHApes Constraints Language
- Real-world RDF graphs are incomplete and contain errors
SHACL Shapes

- An instance of `sh:NodeShape` or `sh:PropertyShape`

- **targets a specific set** of nodes in RDF graph
  - `sh:targetClass`
  - `sh:targetNode`
  - `sh:targetSubjectsOf`
  - ...

- evaluates these nodes against a set of constraints
  - value type (`sh:datatype`)
  - cardinality (`sh:minCount` and `sh:maxCount`)
  - ...

*Inspired by the SHACL shapes examples: [https://www.w3.org/TR/shacl/](https://www.w3.org/TR/shacl/)*
SHACL Validation Report

Validate targeted nodes against the shape :PersonShape

```
ex:Benjamin a ex:Person ; ex:age "21"^^xsd:integer .
ex:Christopher a ex:Person ; ex:age "twenty-one" .
```

```
<1> a sh:ValidationResult ;
   sh:focusNode ex:Christopher ;
   [...] 
   sh:sourceConstraintComponent sh:DatatypeConstraintComponent .

[ a sh:ValidationReport ;
   sh:conforms false ;
   sh:result <1> ] .
```
Research Question

How to design a validation process considering physiological errors in real-life data?

Physiological errors

In a real-world context, RDF graphs can be imperfect and incomplete

● Collaborative building of large RDF graphs (e.g. Wikidata)
● Automatically constructed RDF graphs (e.g. DBpedia)
A Probabilistic Model for SHACL Validation

Let a shape $S$ and an RDF graph $\mathcal{G}$, we note:

- $\mathcal{V}_S$ the set of **triples tested** during the validation
- $\mathcal{V}^-_S$ the set of **violations**
- $\mathcal{V}^+_S$ the set of **confirmations**

\[
\mathcal{V}_S = \mathcal{V}^+_S \cup \mathcal{V}^-_S
\]
A Probabilistic Model for SHACL Validation

- **Assumption**: the validation process of a shape follows a binomial distribution considering a rate of physiological errors $p$

  When a triple violates a shape we consider it is a success (1)

  Otherwise, it is a failure (0).

- **Likelihood** of observing $\|v_S^-\|$ violations in $v_S$

  $$L_{\|v_S^-\|} = P(X = \|v_S^-\|) = \left(\|v_S^-\|\right) \cdot p^{\|v_S^-\|} \cdot (1 - p)^{\|v_S^+\|}$$
A Probabilistic Model for SHACL Validation

Generality measure:

\[ G(S) = \frac{\|v_S\|}{\|v\|} \]

representativeness of a shape \( S \) considering \( v \)
Extended SHACL Validation Report

**Dereferencing:**
https://ns.inria.fr/probabilistic-shacl/

**OWL documentation:**
https://ns.inria.fr/probabilistic-shacl/psh.html

**Probabilistic SHACL Validation**

**IRI:**
http://ns.inria.fr/probabilistic-shacl/

**Current version:**
1.0

**Authors:**
Rémi Felin

**Contributors:**
Andrea G. B. Tettamanzi
Catherine Faron
Rémi Felin

**Publisher:**
Inria

**Other visualisation:**
Ontology source - WebVowl

**LOV:**
https://lov.linkeddata.es/dataset/lov/vocabs/psh
Extended SHACL Validation Report

```
[ a sh:ValidationReport ;
  sh:conforms boolean ;
  sh:result r ;
  # Probabilistic SHACL extension
  psh:summary [  
    a psh:ValidationSummary ;
    psh:focusShape S
    psh:referenceCardinality ||v_S||
    psh:numConfirmation ||v_S^+||
    psh:numViolation ||v_S^-||
    psh:likelihood L ||v_S^-||
    psh:generality G(S) ;
  ] ;
] .
```

```
[ a sh:ValidationReport ;
  sh:conforms false ;
  sh:result :v1 ;
  sh:result :v2 ;
  [...]  
  # SHACL Extension
  # shape s1
  psh:summary [  
    a psh:ValidationSummary ;
    psh:focusShape :s1
    psh:referenceCardinality 200 ;
    psh:numConfirmation 178 ;
    psh:numViolation 22 ;
    psh:likelihood "0.0806"^~xsd:decimal ;
    psh:generality "0.2"^~xsd:decimal ;
  ] ;

  with ||v|| = 1000 and \ p = 0.1
```
Hypothesis Testing for Shape Acceptance

\[ P(X = k) \]

\[ p \ast ||u_{s_1}|| \]

Critical value given by \( \chi^2_{1; \alpha = 0.05} = 3.84 \)
Hypothesis Testing for Shape Acceptance

$$P(X = k)$$

$$p \ast ||v_{s_1}||$$

Critical value given by $$\chi^2_{1; \alpha = 0.05} = 3.84$$

$$||v_{s_1}^-|| = 22$$
Hypothesis Testing for Shape Acceptance

\[ p \times \|\nu_{s_1}\| \]

Critical value given by \( \chi^2_{1; \alpha=0.05} = 3.84 \)

\[ X_{s_1}^2 \approx 0.222 \implies X_{s_1}^2 < \chi^2_{1; \alpha=0.05} \implies \nu \models s_1 \]

\[ \|\nu_{s_1}\| = 22 \]
Experiments


  - *CovidOnTheWeb*: scientific articles annotated with *Wikidata* NE

    | #RDF triples   | 226,647 |
    | #distinct articles | 20,912 |
    | #distinct named entities | 6,331 |
    | avg. #named entities per article | 10.52 |

  - shapes represent association rules [Cadorel & al, WI-IAT, 2020]

- Estimation of the theoretical error proportion of the RDF graph

  Evaluations performed with multiple rates of physiological errors $p$
Experiments

Representing association rules as SHACL shapes

```
:1 a sh:NodeShape ;
  sh:targetClass entity:Q10295810 ;
  sh:property [ sh:path rdf:type ;
    sh:hasValue entity:Q43656 ; ] .
```
Results: Determining a Theoretical Error Proportion

Hypothesis tests performed with a significance level $\alpha = 5\%$

$$\text{avg}(L_{||v_S^-||}) = 0.0362\%$$

(a) $L_{||v_S^-||}$ average

(b) $X_S^2$ average
Results: Shapes acceptance as a function of the theoretical error proportion $p$
Experiments

Results: Shapes acceptance as a function of the theoretical error proportion $p$

- 245/377 shapes required hypothesis testing:
  - 63 are accepted (25.7%)
  - 182 are rejected (74.3%)
Experiments

**Results:** Shapes acceptance as a function of the theoretical error proportion $p$

- **245/377** shapes required hypothesis testing:
  - 63 are accepted (25.7%)
  - 182 are rejected (74.3%)

- **187/377** shapes have been accepted:
  - 33.7% with a hypothesis test
  - 66.3% without a hypothesis test
Experiments

Results on Scalability

Computation time for the evaluation of CovidOnTheWeb against the 377 shapes:

- with standard validation: 1 minute 29
- with a probabilistic validation: 1 minute 35

Linear and small increase of the computation time (6.31%)
Conclusion

● A probabilistic framework relying on \textit{likelihood} and \textit{generality} measures

● A reliable \textbf{automatic acceptance model} based on these measures

● A model for \textit{estimating the theoretical error proportion} from the evaluation of RDF data against a comprehensive set of SHACL shapes

● A \textbf{scalable framework} that can be applied to large RDF graphs

● Perspective: \textbf{shape mining} from RDF graphs using this probabilistic framework
Thank you !