A Short Introduction to SHACL for Logicians

Magdalena Ortiz magdalena.ortiz@umu.se



UMEÅ UNIVERSITY

July 13, 2023

What is SHACL and why do we need it?

- 2 SHACL as a Logic
- **3** SHACL vs. Description Logics
- 4 Semantics of Recursive SHACL
- **5** Explanations for non-validation
- 6 Conclusions and Outlook

On the web, data is stored as RDF graphs

triples (subject, predicate, object)

@prefix : <http://example.com/> .
:cervantes :authorOf :elQuixote .



On the web, data is stored as **RDF graphs**

triples (subject, predicate, object)

@prefix : <http://example.com/> .
:cervantes :authorOf :elQuixote .



DBPedia hundreds of millions Wikipedia facts https://dbpedia.org/

Yago high-quality knowledge graph (people, countries, organizations, ...) > 2 billion facts, 50 million nodes https://yago-knowledge.org/

edge-labeled graphs set of node names N set of property names P



Querying Knowledge graphs

```
Dedicated query language called SPARQL, access via 
endpoints
```

PREFIX : <http://example.com/>

```
SELECT ?author
WHERE {
    ?author :authorOf ?book .
    ?book :hasTitle "El Quixote" .
}
```

Querying Knowledge graphs

```
Dedicated query language called SPARQL, access via 
endpoints
```

PREFIX : <http://example.com/>

```
SELECT ?author
WHERE {
    ?author :authorOf ?book .
    ?book :hasTitle "El Quixote" .
}
```

But what can I query for?

Formalisms for Describing RDF Graphs

OWL, Web Ontology Languages since 2009 Sharable knowledge to infer implicit facts

SHACL, Shapes Constraint Language since 2017 Validate constraints on the graph

Formalisms for Describing RDF Graphs

OWL, Web Ontology Languages since 2009 Sharable knowledge to infer implicit facts

SHACL, Shapes Constraint Language since 2017 Validate constraints on the graph

Ann Person

 $Person \sqsubseteq \exists hasNumber.PersNumber$

In OWL, we infer that Ann has a PersonNumber

Formalisms for Describing RDF Graphs

OWL, Web Ontology Languages since 2009 Sharable knowledge to infer implicit facts

SHACL, Shapes Constraint Language since 2017 Validate constraints on the graph

Ann Person

 $\mathsf{Person} \sqsubseteq \exists \mathsf{hasNumber}.\mathsf{PersNumber}$

In OWL, we infer that Ann has a PersonNumber

 $PersonShape \equiv \exists hasNumber.PersNumber$

In SHACL, Ann does not validate the *PersonShape*

The logics behind

OWL is based on Description Logics

What about SHACL? https://www.w3.org/TR/shacl/

For example, the <u>component sh:MinCountConstraintComponent</u> declares the <u>parameter</u> sh:minCount to represent the restriction that a node has at least a minimum number of values for a particular property.

For a constraint component C with mandatory parameters p1, ..., pn, a shape s in a shapes graph SG declares a **constraint** that has **kind** C with mandatory parameter values <p1, v1>, ..., <pn, vn> in SG when s has vi as a value for pi in SG. For constraint components with optional parameters, the constraint declaration consists of the values that the shape has for all mandatory and optional parameters of that component.

- Corman, Reuter & Slavkovic (2018) converted the 'core' of the standard into an abstract, logic-like syntax
- Others have built on it

Recall: set of **property names** P, set of **node names** N. We also have a set of **shape names** S.

We write shape constraints of the form

 $s \equiv \varphi$

using shape expressions φ and path expressions E

$$\varphi ::= s \mid \top \mid \{a\} \mid \neg \varphi \mid \varphi \land \varphi \mid \ge_n E.\varphi \mid E = E$$
$$E ::= p \mid p^- \mid E \cup E \mid E \circ E \mid E^*$$

where $s \in \mathbf{S}$, $a \in \mathbf{N}$, $p \in \mathbf{P}$, and $n \ge 0$.

We can express

$$\varphi_1 \lor \varphi_2 \quad \exists E. \varphi \quad \forall E. \varphi \quad \leq_{n-1} E. \varphi$$

```
a sh:NodeShape ;
sh:property [
    sh:path pizza:hasTopping ;
    sh:minCount 2 ].
```

 $Pizza \equiv \geq_2$ hasTopping. \top , $VeggiePizza \equiv Pizza \land \forall$ hasTopping.VeggieTopping, $VeggieTopping \equiv \{$ mozzarella $\} \lor \{$ tomato $\} \lor \{$ basil $\} \lor \{$ artichoke $\}$

A shapes specification is a pair

(C,T)

- C is a set of shape constraints (usually, one constraint s ≡ φ for each s)
- T is a set of target atoms s(a) with $s \in \mathbf{S}$ and $a \in \mathbf{N}$

$$\begin{split} \mathcal{C} &= \{ & Pizza \equiv \geq_2 \texttt{hasTopping.T}, \\ & VeggiePizza \equiv Pizza \land \forall \texttt{hasTopping.} VeggieTopping, \\ & VeggieTopping \equiv \{\texttt{mozzarella}\} \lor \{\texttt{tomato}\} \lor \{\texttt{basil}\} \lor \{\texttt{artichoke}\} \; \} \end{split}$$

 $T = \{Pizza(apricciosa), VeggiePizza(margherita)\}$

Interpretations

We consider $I = (\Delta, \cdot^{I})$ with a non-empty **domain** $\Delta \subseteq \mathbb{N}$ and an **interpretation function** \cdot^{I} that assigns to node names $a \in \mathbb{N}$, an element $a^{I} \in \Delta$, to shape names $s \in \mathbb{S}$, a set $s^{I} \subseteq \Delta$, to property names $p \in \mathbb{P}$, a set of pairs $p^{I} \subseteq \Delta \times \Delta$

Interpretations

We consider $I = (\Delta, \cdot^{I})$ with a non-empty domain $\Delta \subseteq \mathbb{N}$ and an interpretation function \cdot^{I} that assigns to node names $a \in \mathbb{N}$, an element $a^{I} \in \Delta$, to shape names $s \in \mathbb{S}$, a set $s^{I} \subseteq \Delta$, to property names $p \in \mathbb{P}$, a set of pairs $p^{I} \subseteq \Delta \times \Delta$

 \cdot^{I} is extended to complex expressions as expected:

• path expressions E are binary relations E^{I} over Δ

Interpretations

We consider $I = (\Delta, \cdot^{I})$ with a non-empty domain $\Delta \subseteq \mathbb{N}$ and an interpretation function \cdot^{I} that assigns to node names $a \in \mathbb{N}$, an element $a^{I} \in \Delta$, to shape names $s \in \mathbb{S}$, a set $s^{I} \subseteq \Delta$, to property names $p \in \mathbb{P}$, a set of pairs $p^{I} \subseteq \Delta \times \Delta$

 \cdot^{I} is extended to complex expressions as expected:

- path expressions E are binary relations E^{I} over Δ
- shape expressions φ interpreted as sets $\varphi^{\mathcal{I}} \subseteq \Delta$

$$\begin{split} \{a\}^{I} &= \{a\} \quad \forall^{I} = \Delta \quad (\neg \varphi)^{I} = \Delta \setminus \varphi^{I} \\ (\varphi_{1} \land \varphi_{2})^{I} &= \varphi_{1}^{I} \cap \varphi_{2}^{I} \\ (\geq_{n} E.\varphi)^{I} &= \{d \in \Delta \mid \text{there exist distinct } d_{1}, \dots, d_{n} \\ & \text{with } (d, d_{i}) \in E^{I} \text{ land } d_{i} \in \varphi^{I} \text{ for each } 1 \leq i \leq n\} \\ (E = E')^{I} &= \{d \in \Delta \mid \text{ for all } c \in \Delta : (c, d) \in E^{I} \text{ iff } (c, d) \in E'^{I} \} \end{split}$$

In SHACL, the graph itself is an interpretation that we can adorn with the shape names

Let $G = (N, E, \ell)$ be a data graph

An shape adorment of of *G* is an *I* with $p^{I} = \{(a, b) \in E \mid p \in \ell(a, b)\}$ for each property *p*.

I satisfies a constraint $s \equiv \varphi$ if $s^I = \varphi^I$

G validates (C,T) if there is an adorment satisfies all constraints in *C* and has $a \in s^{I}$ for every $s(a) \in T$.

Definition (SHACL validation)

The **SHACL validation problem** consists on deciding, for a given G and (C,T), whether G validates (C,T).

Definition (SHACL validation)

The **SHACL validation problem** consists on deciding, for a given G and (C,T), whether G validates (C,T).

$$C = \{ Pizza \equiv \geq_2 \text{hasTopping.} \top, \\ VeggiePizza \equiv Pizza \land \forall \text{hasTopping.} VeggieTopping, \\ VeggieTopping \equiv \{\text{mozzarella}\} \lor \{\text{tomato}\} \lor \{\text{basil}\} \lor \{\text{artichoke}\} \}$$

 $T = \{Pizza(pizza_capricciosa)\}$ $T' = \{VeggiePizza(pizza_capricciosa)\}$



If we call S concept names, then the same syntax defines concept expressions in the description logic $\mathcal{RLCOIQ}_{reg}^{=}$.

OWL and Description Logics

If we call S concept names, then the same syntax defines concept expressions in the description logic $\mathcal{RLCOIQ}_{reg}^{=}$. An $\mathcal{RLCOIQ}_{reg}^{=}$ ontology is a set of concept inclusions

 $\varphi_1 \sqsubseteq \varphi_2$

In the special case of terminologies we have definitions

 $s \equiv \varphi$

If we call S concept names, then the same syntax defines concept expressions in the description logic $\mathcal{ALCOIQ}_{reg}^{=}$. An $\mathcal{ALCOIQ}_{reg}^{=}$ ontology is a set of concept inclusions

 $\varphi_1 \sqsubseteq \varphi_2$

In the special case of terminologies we have definitions

 $s \equiv \varphi$

Semantics of concept expressions in *I* as for SHACL

- *I* satisfies an inclusion $\varphi_1 \sqsubseteq \varphi_2$ if $\varphi_2^I = \varphi_2^I$
- *I* is a **model** of an ontology *O* if it satisfies all inclusions.

If we call S concept names, then the same syntax defines concept expressions in the description logic $\mathcal{ALCOIQ}_{reg}^{=}$. An $\mathcal{ALCOIQ}_{reg}^{=}$ ontology is a set of concept inclusions

 $\varphi_1 \sqsubseteq \varphi_2$

In the special case of terminologies we have definitions

 $s \equiv \varphi$

Semantics of concept expressions in *I* as for SHACL

- *I* satisfies an inclusion $\varphi_1 \sqsubseteq \varphi_2$ if $\varphi_2^I = \varphi_2^I$
- *I* is a model of an ontology *O* if it satisfies all inclusions.

Relaxed definition of model: I is a model of G if it contains G

- G is consistent with O if there exists a model of G and O
- G and O entail a fact α if it's true in every model of G and O

SHACL validation vs. DL entailment



 $C'_{\text{Pizza}} = \{ VeggiePizza \equiv \forall \text{hasTopping}. VeggieTopping, \\ VeggieTopping \equiv \{ \text{mozzarella} \} \lor \{ \text{tomato} \} \lor \{ \text{basil} \} \lor \{ \text{artichoke} \} \}$

G_{Pizza} validates *VeggiePizza*(pizza_margherita)

SHACL validation vs. DL entailment



 $\begin{aligned} C'_{\text{Pizza}} &= \{ \textit{VeggiePizza} \equiv \forall \texttt{hasTopping}. \textit{VeggieTopping}, \\ \textit{VeggieTopping} &\equiv \{\texttt{mozzarella}\} \lor \{\texttt{tomato}\} \lor \{\texttt{basil}\} \lor \{\texttt{artichoke}\} \} \end{aligned}$

G_{Pizza} validates *VeggiePizza*(pizza_margherita)

Assuming that *VeggieTopping* and *VeggiePizza* are concept names, consider the ontology:

 $\begin{aligned} O_{\text{Pizza}} &= \{ \textit{VeggiePizza} \equiv \forall \texttt{hasTopping}. \textit{VeggieTopping}, \\ \textit{VeggieTopping} &\equiv \{\texttt{mozzarella}\} \lor \{\texttt{tomato}\} \lor \{\texttt{basil}\} \lor \{\texttt{artichoke}\} \} \end{aligned}$

 $O_{\text{Pizza}}, G_{\text{Pizza}}$ does not entail $VeggiePizza(\text{pizza}_{margherita})$

DL resoning and SHACL satisfiability

In DLs, complexity of consistency and entailment known

DL resoning and SHACL satisfiability

In DLs, complexity of consistency and entailment known

$$\varphi ::= s \mid \top \mid \{a\} \mid \neg \varphi \mid \varphi \land \varphi \mid \leq_n E.\varphi \mid E = E$$
$$E ::= p \mid p^- \mid E \cup E \mid E \circ E \mid E^*$$

DL resoning and SHACL satisfiability

In DLs, complexity of consistency and entailment known

$$\varphi ::= s \mid \top \mid \{a\} \mid \neg \varphi \mid \varphi \land \varphi \mid \leq_n E.\varphi \mid E = E$$
$$E ::= p \mid p^- \mid E \cup E \mid E \circ E \mid E^*$$

• In *ALCOIQ* they are NExpTime complete

 $\varphi ::= s \mid \top \mid \{a\} \mid \neg \varphi \mid \varphi \land \varphi \mid \geq_n p.\varphi$

OWL is an extension of this DL called SHOIQ

In ALCOIQ_{reg} they are long-standing open problems

 $\varphi ::= s \mid \top \mid \{a\} \mid \neg \varphi \mid \varphi \land \varphi \mid \leq_n p.\varphi \mid \exists E.\varphi$

Extensions with $\geq_1 p.\varphi$ and $\mu \mathcal{RLCOIQ}$ are undecidable

SHACL satisfiability and containment are undecidable

 $Director \equiv \exists creatorOf.Movie$ $Movie \equiv \exists creatorOf^-.Director$

 $Director \equiv \exists creatorOf.Movie$ $Movie \equiv \exists creatorOf^{-}.Director$



 $Director \equiv \exists creatorOf.Movie$ $Movie \equiv \exists creatorOf^{-}.Director$



We can validate Director(Shakespaere) !?!

$certifiedNode \equiv \exists hasCertificate.Certificate \lor \exists approvedBy.certifiedNode$

$certifiedNode \equiv \exists hasCertificate.Certificate \lor \exists approvedBy.certifiedNode$



$certifiedNode \equiv \exists hasCertificate.Certificate \lor \exists approvedBy.certifiedNode$



We can validate *certifiedNode*(node_1)

"The validation with recursive shapes is not defined in SHACL and is left to SHACL processor implementations. For example, SHACL processors may support recursion scenarios or produce a failure when they detect recursion."

SHACL Recommendation, §3.4.3

Two robust semantics that avoid dubious validations **Stable model semantics** minimal models of the Gelfond-Lifschitz reduct Can be defined using **levels** NP-complete Well-founded semantics 3-valued

approximation of stable models P-complete

Two robust semantics that avoid dubious validations **Stable model semantics** minimal models of the Gelfond-Lifschitz reduct Can be defined using levels NP-complete

Well-founded semantics 3-valued approximation of stable models P-complete

Not only intuitive, also computationally more manageable

Two robust semantics that avoid dubious validations **Stable model semantics** minimal models of the Gelfond-Lifschitz reduct Can be defined using **levels** NP-complete

Well-founded semantics 3-valued approximation of stable models P-complete

Not only intuitive, also computationally more manageable

For stratified constraints (only positive recursion cycles)

- Perfect model = stable model = well-founded model
- P-complete validation

Two robust semantics that avoid dubious validations **Stable model semantics** minimal models of the Gelfond-Lifschitz reduct Can be defined using **levels** NP-complete

Well-founded semantics 3-valued approximation of stable models P-complete

Not only intuitive, also computationally more manageable

For stratified constraints (only positive recursion cycles)

- Perfect model = stable model = well-founded model
- P-complete validation
- In contrast, supported validation is NP-complete

Ideas from Logic (Programming) also for:

- validating large graphs
 - Magic sets
 - combines advantages of bottom-up and top-down validation
 - only 'relevant' neighbourhood of potentially huge graphs
- inconsistency tolerant validation
 - disregard inconsistencies that are relevant to my targets

Explanations

SHACL specification describes validation reports

- explain the outcome of validating an RDF graph against a shapes graph
- explaining why the input graph does not satisfy the constraints is challenging.



Explanations

SHACL specification describes validation reports

- explain the outcome of validating an RDF graph against a shapes graph
- explaining why the input graph does not satisfy the constraints is challenging.



The standard leaves it open how to provide such explanations to the users!

22/30 Magdalena Ortiz A Short Introduction to SHACL for Logicians

We propose a notion of explanations based on logic-based abduction and database repairs!

Definition

Let *G* be a graph, (C,T) a SHACL specification, and the set of hypotheses *H* a graph disjoint from *G*. Then $\Psi = (G, C, T, H)$ is a SHACL Explanation Problem (SEP). An **explanation** for Ψ is a pair (A, D), such that $D \subseteq G, A \subseteq H$, and $(G \setminus D) \cup A$ validates (C,T).

We consider 2 typical preference orders over explanations:

- Subset-minimal explanations (⊆)
- Cardinality-minimal explanations (≤)

Consider a SEP $\Psi = (G, C, T, H)$, where:

 $G = \{enrolledIn(Ben, C_1)\}$

 $C = \{ \mathsf{Student} \leftrightarrow \exists enrolledIn.Course \}$

 $T = \{\mathsf{Student}(Ben)\}$

 $H = \{Course(C_1), Course(C_2)\}$

Consider a SEP $\Psi = (G, C, T, H)$, where:

 $G = \{enrolledIn(Ben, C_1)\}$ $C = \{Student \leftrightarrow \exists enrolledIn.Course\}$ $T = \{Student(Ben)\}$ $H = \{Course(C_1), Course(C_2)\}$

G does not validate (C,T).

Consider a SEP $\Psi = (G, C, T, H)$, where:

 $G = \{enrolledIn(Ben, C_1)\}$

 $C = \{ \mathsf{Student} \leftrightarrow \exists enrolledIn. Course \}$

 $T = \{\mathsf{Student}(Ben)\}$

 $H = \{Course(C_1), Course(C_2)\}$

G does not validate (C,T).

Explanation (A, D), where $A = \{Course(C_1)\}$ and $D = \emptyset$.

Consider a SEP $\Psi = (G, C, T, H)$, where:

 $C = \{\mathsf{Student} \leftrightarrow \exists enrolledIn. Course \land =_1 hasID\}$

 $T = \{\mathsf{Student}(Ben)\}$

 $G = \{enrolledIn(Ben, C_1), hasID(Ben, id1), hasID(Ben, id2)\}$ $H = \{Course(C_1), Course(C_2), enrolledIn(Ben, C_2)\}$

There are 2 \leq -explanations for Ψ :

 $\begin{aligned} A_1 = \{Course(C_1)\}, D_1 = \{hasID(Ben, id1)\}. \\ A_2 = \{Course(C_1)\}, D_2 = \{hasID(Ben, id2)\}. \end{aligned}$

Let $\Psi = (G, C, T, H)$ be a SEP, let A, D be data graphs, let α be an atom in $G \cup H$, and let \leq be a (possibly empty) preorder.

- **1** \leq -ISEXPL: is (A, D) a \leq -explanation for Ψ ?
- **2** \leq -EXIST: does there exist a \leq -explanation for Ψ ?
- **3** \leq -NECADD: is α a \leq -necessary addition for Ψ , that is does α occur in A in every \leq -explanation (A, D) for Ψ ?
- ≤ -NECDEL: is α a ≤-necessary deletion for Ψ, that is does α occur in *D* in every ≤-explanation (*A*, *D*) for Ψ?
- **5** \leq -RELADD: is α a \leq -relevant addition for Ψ , that is does α occur in A in some \leq -explanation (A, D) for Ψ ?
- **6** \leq -RELDEL: is α a \leq -relevant deletion for Ψ , that is does α occur in D in some \leq -explanation (A, D) for Ψ ?

Complexity Results

	Ø	⊆	\leq
ISEXPL	NP-c	DP-c	DP-c
Exist	NP-c	NP-c	NP-c
NECADD	coNP-c	coNP-c	P ^{∥NP} -c
NECDEL	coNP-c	coNP-c	P ^{∥NP} -c
RelAdd	NP-c	Σ_2^P -C	P ^{∥NP} -c
RelDel	NP-c	Σ_2^P -c	P ^{∥NP} -c

All these results hold in data and combined complexity

We have also studied

- the non-recursive case same complexity
- a more general setting with explanation signatures
 - we can specify predicates that are read-only add-only delete-only
 - fine-grianed filtering of explanations
 - all the results hold also for this setting

Using ASP to explain non-validation

Simple prototype for providing explanations using Answer Set Programming (ASP)

Some of our ongoing efforts:

- validation in the presence of ontologies
- validation when the graph is subjected to updates
- better and more useful explanations for validation reports

Many open problems for logicians to tackle!

SHACL is a relatively young field that can use insights from well-established areas of logic

Some of our ongoing efforts:

- validation in the presence of ontologies
- validation when the graph is subjected to updates
- better and more useful explanations for validation reports

Many open problems for logicians to tackle!

SHACL is a relatively young field that can use insights from well-established areas of logic

Thanks to all collaborators!

In particular: Mantas Šimkus, Shqiponja Ahmetaj, Medina Andresel, Bianca Löhnert, Anouk Oudshoorn, Juan Reutter, Julien Corman.

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation and the Austrian Science Fund (FWF) projects P30360 and P30873.