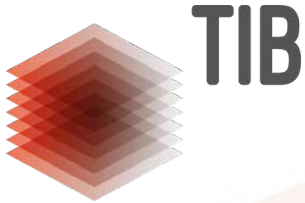


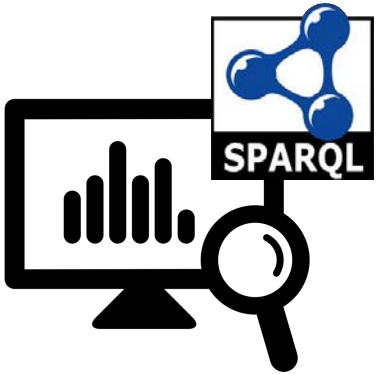
SHACL Constraint Validation during SPARQL Query Processing

Philipp D. Rohde

VLDB 2021 PhD Workshop
August 16th, 2021

Introduction

Knowledge Graphs (KGs) gain Momentum



Data Retrieval: SPARQL Queries

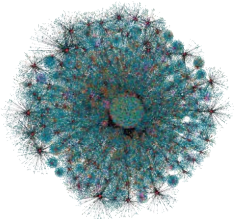
Integrity Constraints (ICs) on RDF KGs

SHACL is the standard to specify **ICs** on **RDF KGs**

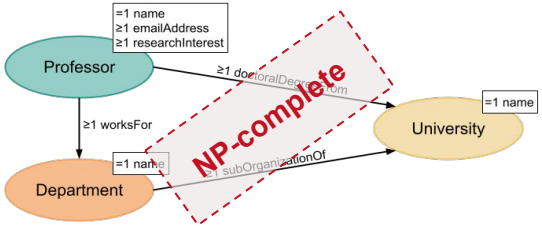
Motivating Example

INPUT

KG of a University System with 37,419 entities (~1M triples)



Shape Schema: Integrity Constraints on the KG



SPARQL Query: Retrieving Data from the KG

```
SELECT ?name ?ri ?uni WHERE {
  ?prof  rdf:type  ub:FullProfessor ;
    ub:name  ?name ;
    ub:worksFor  ?dept0 ;
    ub:doctoralDegreeFrom  ?uni ;
    ub:emailAddress  ?email ;
    ub:researchInterest  ?ri .
}
```

NP-complete

- find physical plan
- execute SHACL validation
- produce results efficiently
- new optimization techniques based on the context are needed

OUTPUT

SPARQL Query Result with SHACL Validation Result Annotation

name	ri	uni	__meta__
FullProfessor0	Research6	http://www.Univeristy6.edu	all requirements met
FullProfessor3	Research10	http://www.University888.edu	University888 violates name constraint
		...	

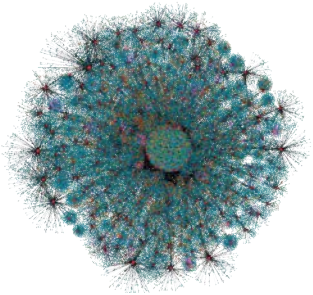
Query result annotation requires
- SPARQL query execution
- SHACL shape schema validation

and is **computationally complex**

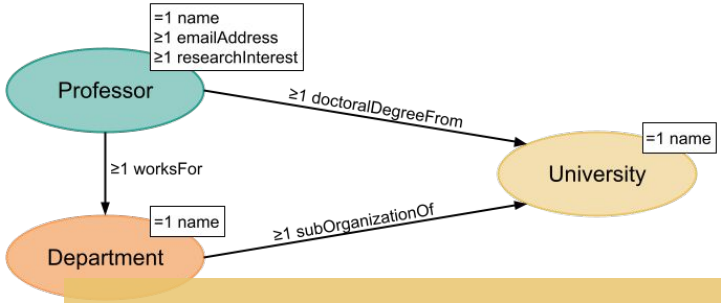
Motivating Example

INPUT

KG of a University System with 37,419 entities (~1M triples)



Shape Schema: Integrity Constraints on the KG



SPARQL Query: Retrieving Data from the KG

```
SELECT ?name ?ri ?uni WHERE {
  ?prof  rdf:type  ub:FullProfessor ;
    ub:name ?name ;
    ub:worksFor  :Dept0 ;
    ub:doctoralDegreeFrom  ?uni ;
    ub:emailAddress  ?email ;
    ub:researchInterest  ?ri .
}
```

Query result annotation requires
- SPARQL query execution
- SHACL shape schema validation

and is **computationally complex**

OUTPUT

SPARQL Query Result Annotation

name	ri	uni	__meta__
FullProfessor0	Research6	http://www.Univeristy6.edu	all requirements met
FullProfessor3	Research10	http://www.University888.edu	University888 violates name constraint
		...	

Presentation Outline

1. Related Work
2. Proposed Approach
   - 2.1. Problem Statement
   - 2.2. Proposed Solution
   - 2.3. Online vs Offline Validation
3. Preliminary Results
   - 3.1. Trav-SHACL
   - 3.2. SPARQL Query Result Annotation
4. Lessons Learned
5. Conclusions

Related Work

SHACL Validation

- Complexity analysis of SHACL
  - NP-complete
- Identification of tractable fragments
  - no negation, but disjunction
  - no negation in recursion
  - no recursion
- SHACL validation using SPARQL endpoints
  - before: in-memory knowledge graphs
  - Datalog-like rules
- Recursive SHACL
  - left open in the specification

Integrity Constraints in Query Processing

Abbas et al.

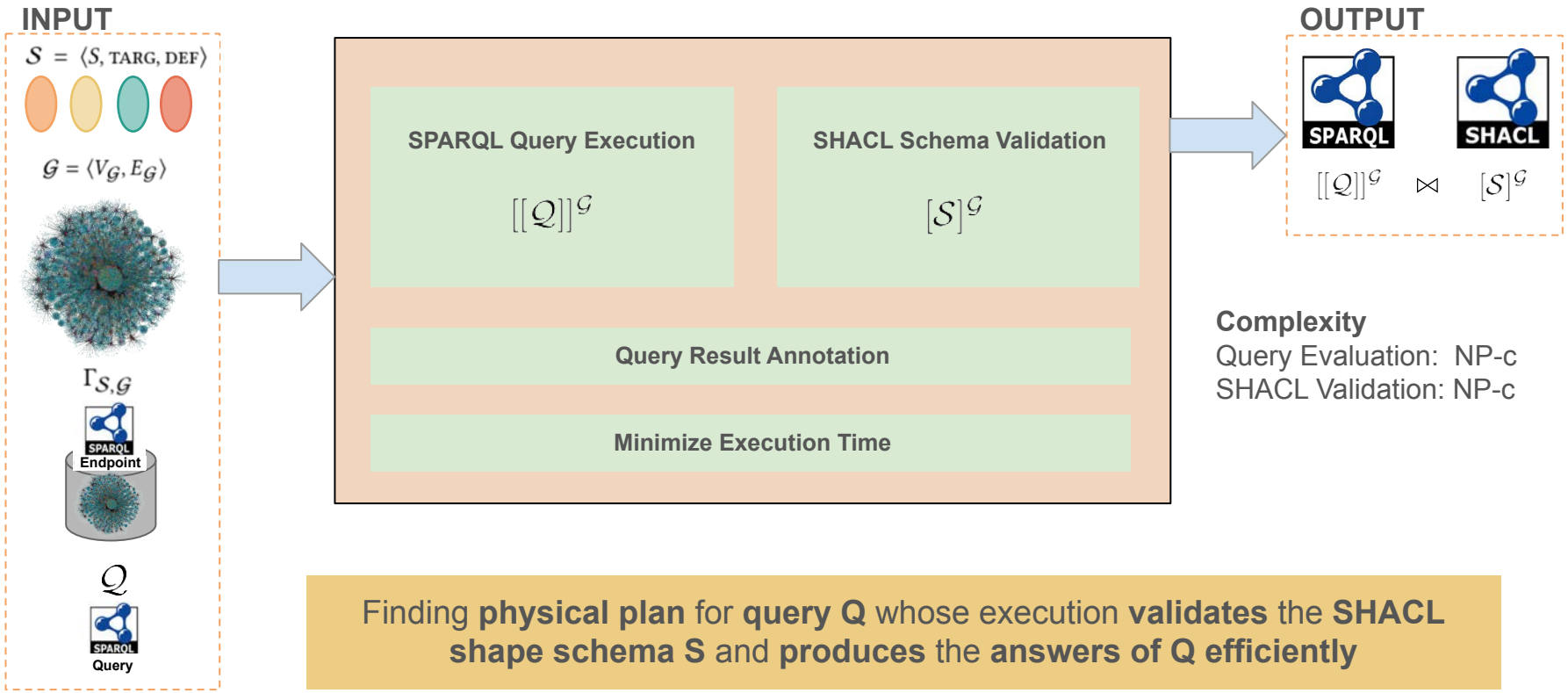
- well-formed ShEx schemas
- SPARQL triple pattern reordering
  - hierarchical structure of ShEx shapes
  - shapes included in other shapes are ranked higher
  - triple patterns with unique predicates are ranked highest

Rabbani et al.

- Extension of SHACL with statistics
- Cost-based query optimizer
- Precomputation time reduced

No work on **explainable SPARQL query results** so far.

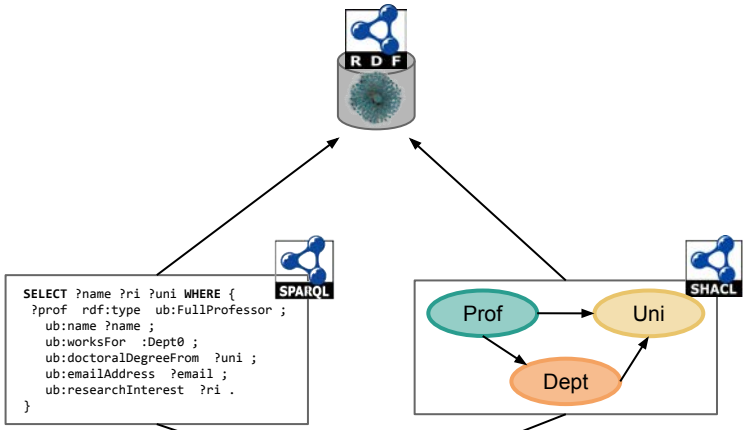
Problem Statement



Complexity
Query Evaluation: NP-c
SHACL Validation: NP-c

Finding **physical plan** for **query Q** whose execution **validates** the **SHACL shape schema S** and **produces** the **answers of Q efficiently**

Proposed Approach



Query Decomposition

- subject star-shaped decomposition
- one star ≈ one class

SHACL Validation

- interleaved validation
- subset of shape schema

Query Result Annotation

- add SHACL validation result as metadata
- explainability

Novelty of the approach:

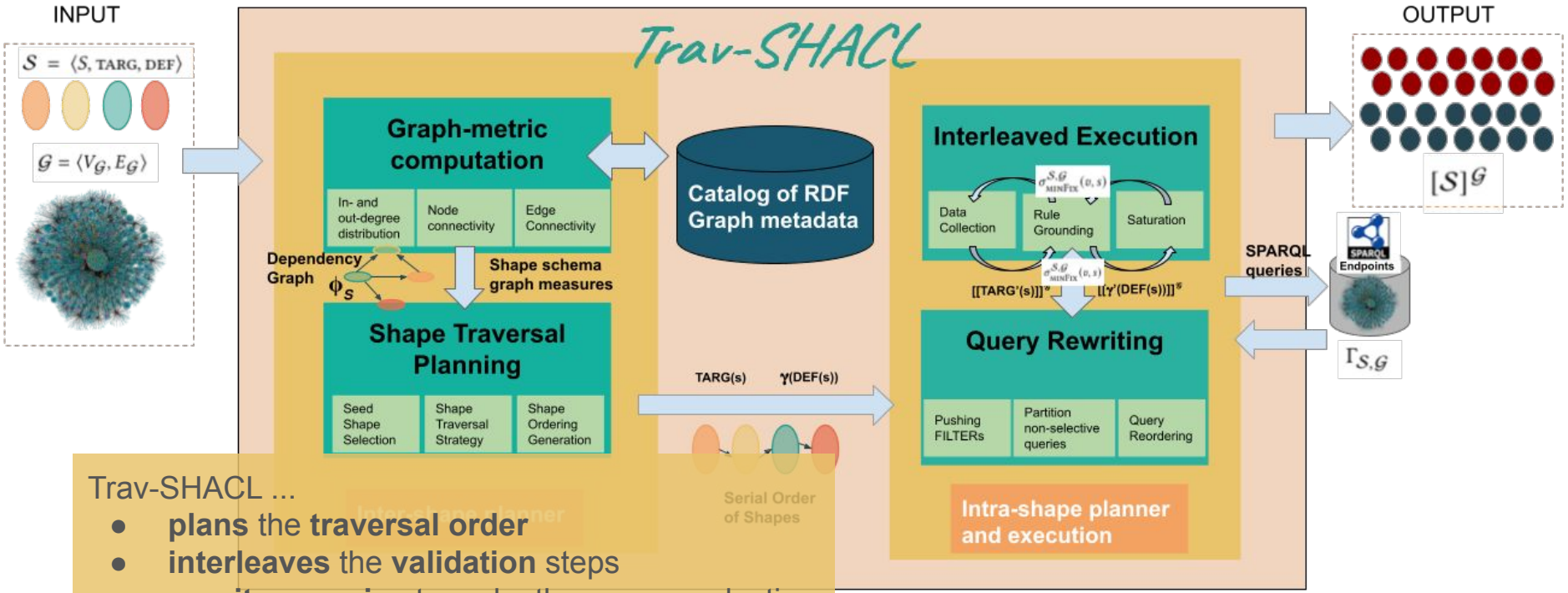
- **identification** of **query plan** able to **combine** query answering with integrity constraint validation
- **explainability** of SPARQL query results
- **optimizations** in SHACL validation

Online vs Offline Validation

	online	offline
speed	❌	✔
who	✔ everyone	❌ data provider
adaptivity	✔	❌

Offline validation is **no option** in this scenario

Results So Far: Trav-SHACL (1/2)



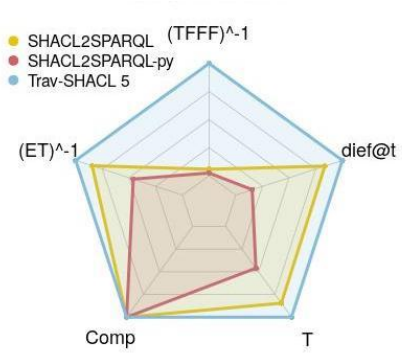
Trav-SHACL ...

- **plans** the **traversal order**
- **interleaves** the **validation** steps
- **rewrites queries** to make them more selective

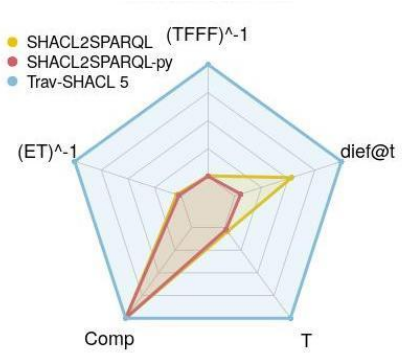
Mónica Figuera, Philipp D. Rohde, and Maria-Esther Vidal. Trav-SHACL: Efficiently Validating Networks of SHACL Constraints, The Web Conference (WWW 2021). https://doi.org/10.1145/3442381.3449877

Results So Far: Trav-SHACL (2/2)

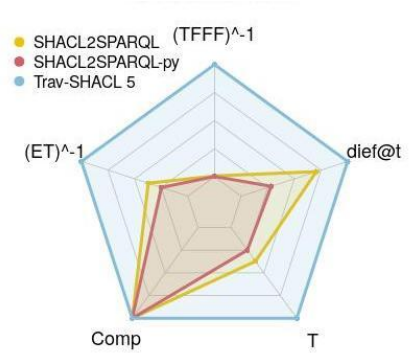
Schema 2 SKGs



Schema 1 MKGs



Schema 3 LKGs



Metrics

dief@t: continuous efficiency at time t

(TFFF)^-1: Time for First Answer (sec)

(ET)^-1: Execution Time (sec)

Comp: sum of (in)validated entities

T: Throughput (answer/sec)

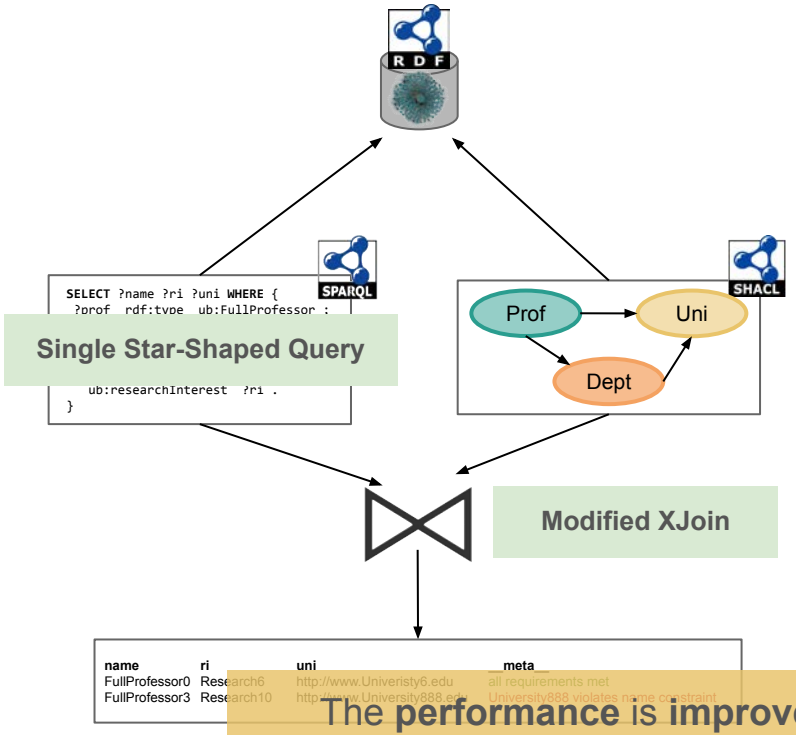
- # Constraint query mappings:
  - **839K** in SHACL2SPARQL,
  - **468K** in Trav-SHACL.
- # Constraint query mappings:
  - **703K** in SHACL2SPARQL,
  - **814** in Trav-SHACL.
- # Constraint query mappings:
  - **22.94M** in SHACL2SPARQL,
  - **6.19M** in Trav-SHACL.

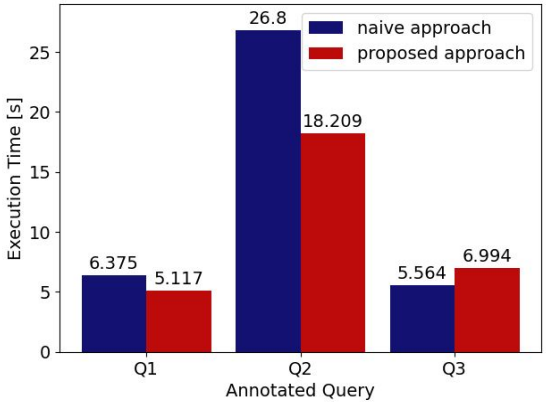
- **Trav-SHACL always delivers results continuously**,
- generates the first answer faster,
- finishes the execution faster,
- scales up to large knowledge graphs.

Impact of the interleaved execution

Mónica Figuera, Philipp D. Rohde, and Maria-Esther Vidal. Trav-SHACL: Efficiently Validating Networks of SHACL Constraints, The Web Conference (WWW 2021). https://doi.org/10.1145/3442381.3449877

Results So Far: Query Annotation





RQ: Is the performance improved by applying the proposed approach?

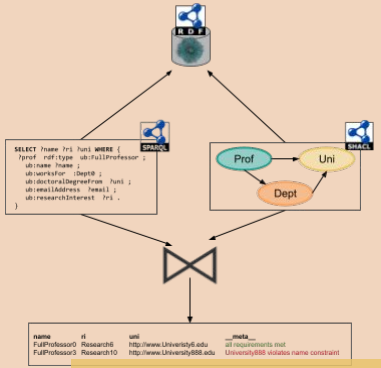
WatDiv

- 10 million triples
- 3-5 triple patterns per query
- less than 100 query results

The **performance** is **improved** but more studies are needed.

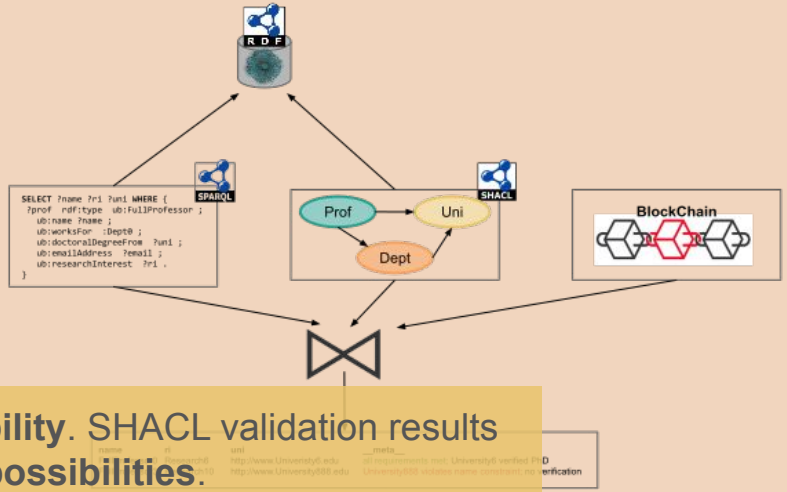
Research Plan

SPARQL Query Result Annotation



- extension of the semantics of SPARQL
- more studies on behavior
- extension to queries with any BGP
- integration into federated query engine

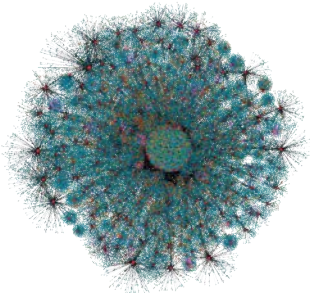
Additional Annotations



- verification of result via blockchain

Extension of SPARQL for **explainability**. SHACL validation results only one of **many possibilities**.

Lessons Learned

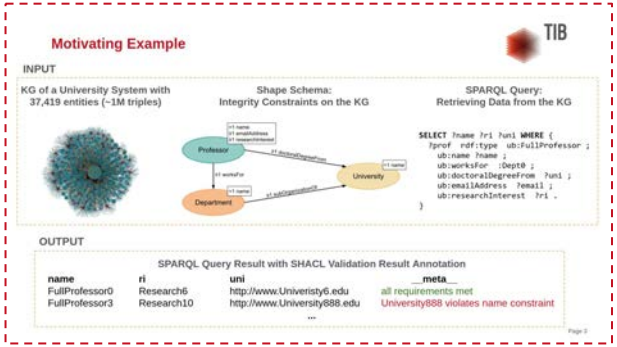


- Benchmark needed
- Validation computationally expensive

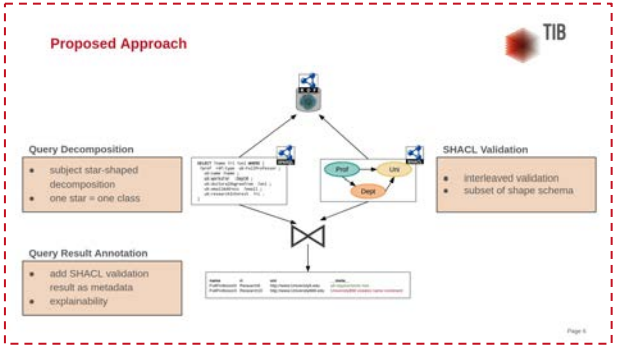
- Explainability needed
- Lower performance is the price

- Many KGs have low data quality
- Means to improve quality are needed

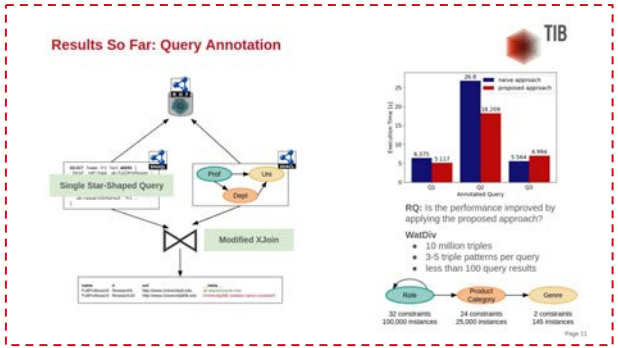
Conclusions



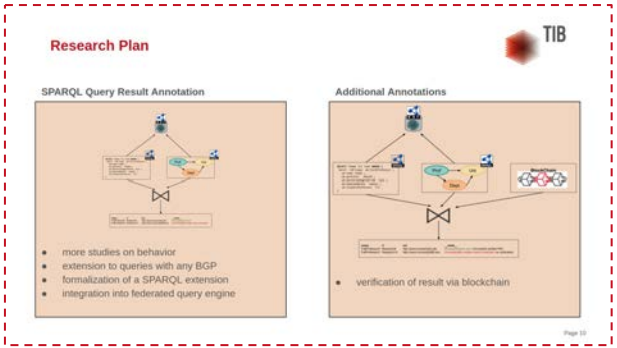
Explainable SPARQL Query Results



Annotation with SHACL Validation



Promising Results from Prototype



Extension of SPARQL for Explainability

Leibniz Universität Hannover

TIB

Thanks for your attention!

Contact:
Philipp D. Rohde
philipp.rohde@tib.eu @philipp_rohde

Leibniz Association