Statistics about Data Shape Use in RDF Data

Sven Lieber, Ben De Meester, Anastasia Dimou, and Ruben Verborgh

Ghent University – imec – IDLab,
Department of Electronics and Information Systems,
Technologiepark-Zwijnaarde 122, 9052 Ghent, Belgium
{firstname.lastname}@ugent.be

Abstract. Statistics about constraint use in RDF data bring insights in common practices to address data quality. However, we only have such statistics for OWL axioms, not for constraint languages, such as SHACL or ShEx, that have recently become more popular. We extended previous work on axiom statistics to provide evidence of constraint type use. In this poster[1] we present preliminary statistics about the use of SHACL core constraints in data shapes found on GitHub. We found that class, datatype and cardinality constraints are predominantly used, similar to the dominant use of domain and range in ontologies. Less-used constraint types need further attention in visualization or modeling tools to address data quality issues. More constraints of SHACL but also ShEx need to be included to deepen the understanding. Data quality researchers and tool designers can make informed decisions based on the provided statistics.

Keywords: SHACL · Statistics · RDF · Constraints · Montolo

1 Introduction

Recently, RDF constraint languages, such as SHACL [5] or ShEx [7], have been developed to model restrictions in the form of constraints on data. Statistics for OWL ontologies showed that only a subset of possible axioms are commonly used [6], but such evidence does not yet exist for constraints which poses a gap and leaves users to anticipate possible use cases or cover whole specifications.

Insights about used constraint types can be taken from generated constraints or curated repositories. Astrea [3] and OSLO [4] which generate shapes from existing sources cover specific subsets of SHACL, but this is due to limited mapping and not because of evidence of broad use. To the best of our knowledge, only small repositories of SHACL constraints with less than 5 entries exist[2] [3].

In this poster paper, we present preliminary statistics generated by a constraint type extension of our Montolo framework [6] to collect RDF Data Cube compliant statistics about axiom use. Following the same approach, we used the

[2] https://schreckl.inspirito.de/

[3] http://shacl-play.sparna.fr/catalog

vocabulary of Montolo[4] to create definitions for all SHACL core constraints and created statistics for identified data shapes from GitHub.

Our work provides insights in constraint type use and is extendible with respect to constraint types of other RDF constraint languages. Preliminary results, the created corpus of SHACL shapes as well as the tool to download the shapes are available with a persistent identifier (DOI: 10.5281/zenodo.3988930[5]) and under an open license[6] to attract more research.

2 Constraint Type Statistics

We explain the framework to collect constraint type statistics, which sources we consider and present preliminary results before we discuss the results.

Framework We briefly describe the framework to collect constraint type statistics and the selection of SHACL data shapes. Montolo uses an extension of LODStats [1] to define statistical modules to detect (patterns of) RDF terms[7]. We created a statistical module for each core constraint of SHACL to detect SHACL serializations of constraint types, e.g., `sh:class` or `sh:minCount`. Additionally, we created definitions for SHACL core constraints with the Montolo vocabulary.

We searched for the term "SHACL" in GitHub and manually selected repositories which contain valid SHACL shapes that do not appear as simple examples. We also considered common SHACL shapes, such as Schema.org's SHACL[8] and SHACL constraints of SHACL itself[9]. We implemented a tool to download data shapes and merge the ones that conceptually belong together, e.g. because they are in the same repository; the tool is part of the accompanying resource of this paper.

Results In total, we analyzed the SHACL RDF files of 13 projects containing 1,978 NodeShapes. Two of the projects, the aforementioned OSLO and the SHACL version of schema.org are similar to the Astrea examples, i.e. data shapes generated based on a subset of SHACL. We describe statistics about constraint types of potentially manually curated SHACL shapes while comparing it with generated SHACL shapes of OSLO, schema.org and Astrea.

All constraint types are used (Fig. 2) but constraint types regarding cardinality, class and datatype of properties are most frequently used by total number (Fig. 1). Class and datatype constraints are primarily found in our corpus which likewise is generated by Astrea, OSLO and SHACL of schema.org. This suggests that class and datatype constraints are main use cases for constraint types which find common use; it appears similar to the dominance of domain

[4] `http://w3id.org/montolo/ns/montolo-voc`
[5] `https://zenodo.org/record/3988930`
[6] `https://creativecommons.org/publicdomain/zero/1.0/`
[7] `https://github.com/IDLabResearch/lovstats`
[8] `http://datashapes.org/schema`
[9] `https://www.w3.org/ns/shacl-shacl`

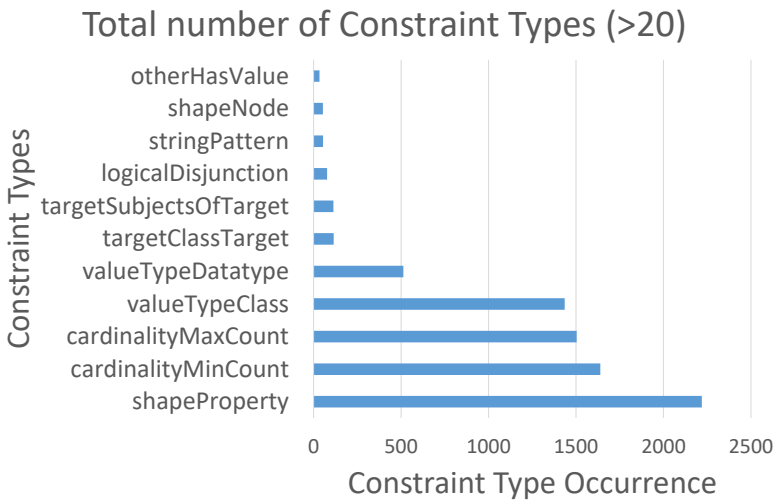

Fig. 1: Constraints on properties, their cardinality and datatype or class are most frequently used in manually curated data shapes (excluding OSLO & schema.org' SHACL). Constraint types used less than 20 times are not shown.

and range axioms for ontologies [6]. Disjunction constraints (`sh:or`) are used by more than 75% of the analyzed repositories and to a large extent by the automatically generated SHACL for schema.org. This can be explained by the flexibility of schema.org: properties are specified to expect one of several possible types. However, disjunction is almost non-existent in Astrea, showing that the selected ontologies barely contain `owl:unionOf` statements. Value range constraints (`sh:minExclusive`, `sh:maxInclusive`, etc) are barely found in our corpus and are neither generated for the Astrea examples nor OSLO, suggesting less future use, similar for other constraint types.

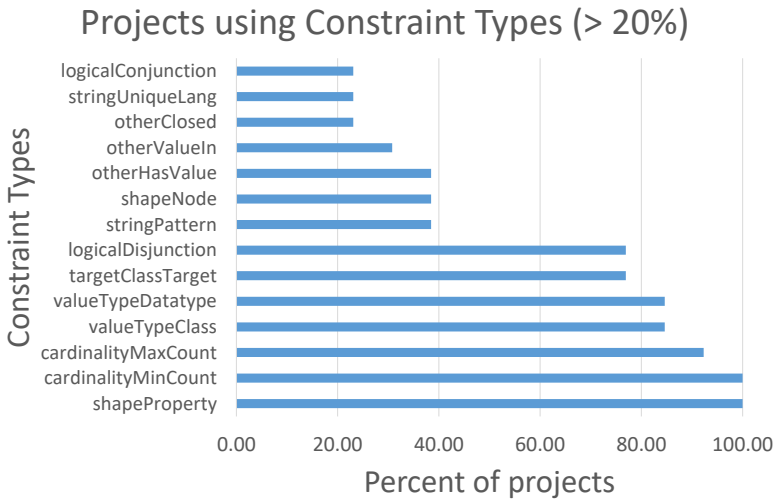

Fig. 2: All constraint types are used, however, datatype, class and cardinality constraints of properties are most often used. Constraint types which are used in less than 20% of the projects are not shown.

Discussion Constraint types complement ontology restrictions yet both show a similar use pattern. Our previous study on restrictions in ontologies found that taxonomic relationships (`rdfs:domain`, `rdfs:range`, `rdfs:subClassOf`) are extensively used whereas restrictions on literals were barely found. We see a similar pattern of constraint use compared to axiom use: relationships between concepts restricted to certain classes or datatypes. However, the current analysis suggests that with respect to literals at least string patterns (`sh:pattern`) find some use in shapes which complements missing literal restrictions use of ontologies.

However, we see more potential in the use of constraints with respect to literals. One out of seven RDF statements in large knowledge graphs contains a literal as object [2]. Several string or literal value range constraint types are defined by SHACL and ShEx which can be used to impose precise restrictions on literals. We have no insights with which tools the shapes were created yet this might be important. Current tools might focus too much on classes and datatypes while neglecting other constraint types. Appropriate tools with user-friendly interfaces are crucial and should be available such that users are made aware of possible constraint types and are assisted in using them.

Conclusion and Future Work Our preliminary results identified cardinality, class, datatype and disjunction constraints as commonly used. Developers of tools related to RDF constraints become able to iteratively implement their tools as they can cover first these commonly used constraint types. However, to exploit the existing data quality potential, developers should not neglect other constraint types completely especially regarding literal values. Future work can extend the statistics by including ShEx and extending the sample size. We currently work on visual notations for RDF constraints[10] which will benefit from this and future insights in constraint type use.

References

1. Auer, S., Demter, J., Martin, M., Lehmann, J.: LODStats - An Extensible Framework for High-Performance Dataset Analytics. In: EKAW (2012)
2. Beek, W., Ilievski, F., Debattista, J., Schlobach, S., Wielemaker, J.: Literally better: Analyzing and improving the quality of literals. Semantic Web (2018)
3. Cimmino, A., Fernández-Izquierdo, A., García-Castro, R.: Astrea: Automatic Generation of SHACL Shapes from Ontologies. The Semantic Web (2020)
4. De Paepe, D., Thijs, G., Buyle, R., Verborgh, R., Mannens, E.: Automated UML-based ontology generation in OSLO^2. In: The Semantic Web: Satellite Events (2017)
5. Knublauch, H., Kontokostas, D.: Shapes Constraint Language (SHACL). Recommendation, World Wide Web Consortium (2017)
6. Lieber, S., De Meester, B., Dimou, A., Verborgh, R.: MontoloStats – Ontology Modeling Statistics. In: Proceedings of the 10$^{\text{th}}$ K-Cap Conference (2019)
7. Prud'hommeaux, E., Labra Gayo, J.E., Solbrig, H.: Shape expressions: an RDF validation and transformation language. In: Proceedings of the 10$^{\text{th}}$ International Conference on Semantic Systems. New York, NY, United States (2014)

[10] `https://w3id.org/imec/unshacled/spec/shape-vowl` and `https://w3id.org/imec/unshacled/spec/shape-uml`