SHACL – based validation of FC Barcelona players data from DBpedia

Author: Vladimír Vavruška

Subject: 4IZ441

Repository: https://github.com/Vavakas16/SHACL-validation-project

Introduction

This semestral project applies SHACL (Shapes Constraint Language) as a validation language for assessing the quality of real-world linked data. The focus is on RDF data describing professional football players of FC Barcelona extracted from DBpedia. Sports-related data are a suitable target for constraint-based validation because they are often populated semi-automatically or aggregated from multiple sources, whcih makes them inconsistently across entities.

The project addresses the absence of explicit guarantees regarding the structural and semantic consistency of football player data in DBpedia. The objective of this project is to design a SHACL shape graph that formalizes reasonable constraints for football players and to assess the extent to which real DBpedia data conform to these constraints.

DATA Extraction and description

The dataset used in this project was obtained from the public DBpedia SPARQL endpoint. The extraction focuses football players associated with FC Barcelona during the 2024–25 season. Players were identified using season-related resources and linked to the club via the *dbo:team* property.

The extracted data were converted into a standalone RDF dataset in Turtle format to enable validation. Each player entity is modeled as an instance of *dbo:Athlete* and described using these properties:

- *foaf:name,*
- *dbo:birthDate*
- *dbo:position,*
- *dbo:height*,
- *dbo:number*
- *dbo:Team*

The dataset intentionally reflects the original state of DBpedia data without manual correction. This ensures that the validation results genuinely represent real-world data quality issues rather than artifacts introduced during preprocessing.

Vocabularies

The primary vocabulary used is the DBpedia Ontology (dbo:), which provides classes and properties for describing athletes, teams, and personal attributes. The FOAF vocabulary (foaf:) is used for representing human-readable names of players. SHACL is then used as the validation language for defining structural and datatype constraints for processing Turtle-formatted RDF graphs.

SHACL Shape Graph

The core contribution of this project is the design of a SHACL shape about football player data. The shape targets all RDF subjects that are associated with a team via the *dbo:team* property, thereby focusing validation on player-like entities.

Several types of constraints were defined:

- **Team**: Every validated entity must have *dbo:team* equal to dbr:FC_Barcelona. This ensures that only relevant players are considered.
- **name**: Every player must atleast 1 name, the language tag must be @en
- **birthDate**: Every players must have exactly 1 birtDate. The datatype is xsd:Date
- **position**: PLayers must have at least one position, which is nodeKind: sh:IRI
- **number**: player must have set number, which is a datatype xsd:nonnegativeInteger
- **height**: Player must have set height, which is either xsd:decima or xsd:double
- warnings on *foaf:name* and *dbo:position* and *dbo:number* if multiple values found

The shape graph is designed to balance strictness and realism: it enforces core semantic expectations while tolerating known modeling practices in DBpedia.

Validation Results and Analysis

The SHACL validation produced a detailed validation report containing both violations and warnings, reflecting different levels of data quality issues in the analyzed dataset. The distinction between these two severity levels allows for a more nuanced interpretation of the results.

Missing Essential player information - Violations

Violations primarily correspond to missing mandatory properties that are considered essential for describing a football player. For several entities, most notably *dbr:Ferran_Torres, dbr:Diego_Kochen*, and *dbr:Frenkie_de_Jong*, multiple *sh:MinCount* violations were detected. These violations indicate the absence of required properties:

- foaf:name
- dbo:birthDate
- dbo:position
- dbo:number
- dbo:height

Despite being typed as *dbo:Athlete* and linked to FC Barcelona, these entities lack basic descriptive attributes. This suggests that DBpedia contains partially populated or placeholder entities that are structurally incomplete.

From a data quality perspective, such entities are problematic, as they cannot be reliably used in applications that expect complete player profiles.Another violation is the absence of a shirt number (*dbo:number*) for *dbr:Aleix_Garrido*.

Warnings

Warnings were issued in cases where the data deviate from strict modeling assumptions but still reflect plausible real-world scenarios. A frequent warning relates to multiple values for *foaf:name*. Players such as *dbr:Dani_Rodríguez*, *dbr:Pablo_Torre*, *dbr:Quim_Junyent*, *dbr:Marc_Bernal*, and others have more than one name value, typically representing full names and commonly used short names.

These situations violate the *sh:MaxCount* constraint but were intentionally classified as warnings rather than violations, as they do not represent incorrect data but rather alternative naming practices.

Another important group of warnings concerns multiple player positions (*dbo:position*). Players such as *dbr:Ansu_Fati, dbr:Pau_Víctor, dbr:Dani_Olmo*, and *dbr:Eric_García* have more than one position assigned. While this violates the maximum cardinality constraint, it accurately reflects the fact that football players often play in multiple roles. These warnings therefore expose a tension between simplified schema assumptions and real-world complexity.

Conclusion

The original contribution of this project lies in the formulation of realistic SHACL constraints tailored to football player data and in the systematic evaluation of real-world DBpedia content against these constraints.

The validation results demonstrate how SHACL can be used not only as a strict conformance checker but also as an analytical tool for data quality assessment. Overall, the validation results reveal two dominant patterns. Violations point to structural incompleteness of some DBpedia entities, where essential attributes are missing entirely. Warnings, on the other hand, highlight modeling variability, especially in the representation of names and playing positions.

These findings confirm that SHACL validation is most valuable when interpreted analytically. Rather than merely labeling data as valid or invalid, the validation results provide insight into how Linked Open Data are modeled in practice and where schema expectations diverge from real-world semantics.

Sources

https://dbpedia.org/sparql/

https://shacl.org/playground/