World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:15, No:12, 2021

Validation of Mapping Historical Linked Data to International Committee for Documentation (CIDOC) Conceptual Reference Model Using Shapes Constraint Language

Authors : Ghazal Faraj, András Micsik

Abstract : Shapes Constraint Language (SHACL), a World Wide Web Consortium (W3C) language, provides well-defined shapes and RDF graphs, named "shape graphs". These shape graphs validate other resource description framework (RDF) graphs which are called "data graphs". The structural features of SHACL permit generating a variety of conditions to evaluate string matching patterns, value type, and other constraints. Moreover, the framework of SHACL supports high-level validation by expressing more complex conditions in languages such as SPARQL protocol and RDF Query Language (SPARQL). SHACL includes two parts: SHACL Core and SHACL-SPARQL. SHACL Core includes all shapes that cover the most frequent constraint components. While SHACL-SPARQL is an extension that allows SHACL to express more complex customized constraints. Validating the efficacy of dataset mapping is an essential component of reconciled data mechanisms, as the enhancement of different datasets linking is a sustainable process. The conventional validation methods are the semantic reasoner and SPARQL queries. The former checks formalization errors and data type inconsistency, while the latter validates the data contradiction. After executing SPARQL queries, the retrieved information needs to be checked manually by an expert. However, this methodology is time-consuming and inaccurate as it does not test the mapping model comprehensively. Therefore, there is a serious need to expose a new methodology that covers the entire validation aspects for linking and mapping diverse datasets. Our goal is to conduct a new approach to achieve optimal validation outcomes. The first step towards this goal is implementing SHACL to validate the mapping between the International Committee for Documentation (CIDOC) conceptual reference model (CRM) and one of its ontologies. To initiate this project successfully, a thorough understanding of both source and target ontologies was required. Subsequently, the proper environment to run SHACL and its shape graphs were determined. As a case study, we performed SHACL over a CIDOC-CRM dataset after running a Pellet reasoner via the Protégé program. The applied validation falls under multiple categories: a) data type validation which constrains whether the source data is mapped to the correct data type. For instance, checking whether a birthdate is assigned to xsd:datetime and linked to Person entity via crm:P82a_begin_of_the_begin property. b) Data integrity validation which detects inconsistent data. For instance, inspecting whether a person's birthdate occurred before any of the linked event creation dates. The expected results of our work are: 1) highlighting validation techniques and categories, 2) selecting the most suitable techniques for those various categories of validation tasks. The next plan is to establish a comprehensive validation model and generate SHACL shapes automatically.

Keywords : SHACL, CIDOC-CRM, SPARQL, validation of ontology mapping
Conference Title : ICDEMLA 2021 : International Conference on Data Engineering and Machine Learning Applications
Conference Location : Amsterdam, Netherlands
Conference Dates : December 02-03, 2021