Enabling Dataset Trustworthiness by Exposing the Provenance of Mapping Quality Assessment and Refinement

Tom De Nies, Anastasia Dimou, Ruben Verborgh, Erik Mannens, and Rik Van de Walle

Contents

Context: assessing trust of RDF datasets

Mapping semi-structured data to RDF

Mapping quality assessment and refinement workflow

Capturing provenance of the workflow

Deriving Trust

Context

How do we decide to **trust an RDF dataset** or not?

One important aspect is: *where did it come from?*

In a lot of cases, the **RDF data was mapped** from semi-structured data using a mapping language.

In our lab, we developed such a language:
The RDF Mapping Language http://rml.io

Mapping semi-structured data to RDF

W3C R2RML exists to map databases to RDF

To map all other data formats, there’s **RML**

The cool thing: **RML definitions are RDF themselves**
→ they can be *queried* using SPARQL

The problem: **not all mappings are perfect** right away

Mapping quality assessment and refinement workflow

Evaluate data quality **during the mapping** stage

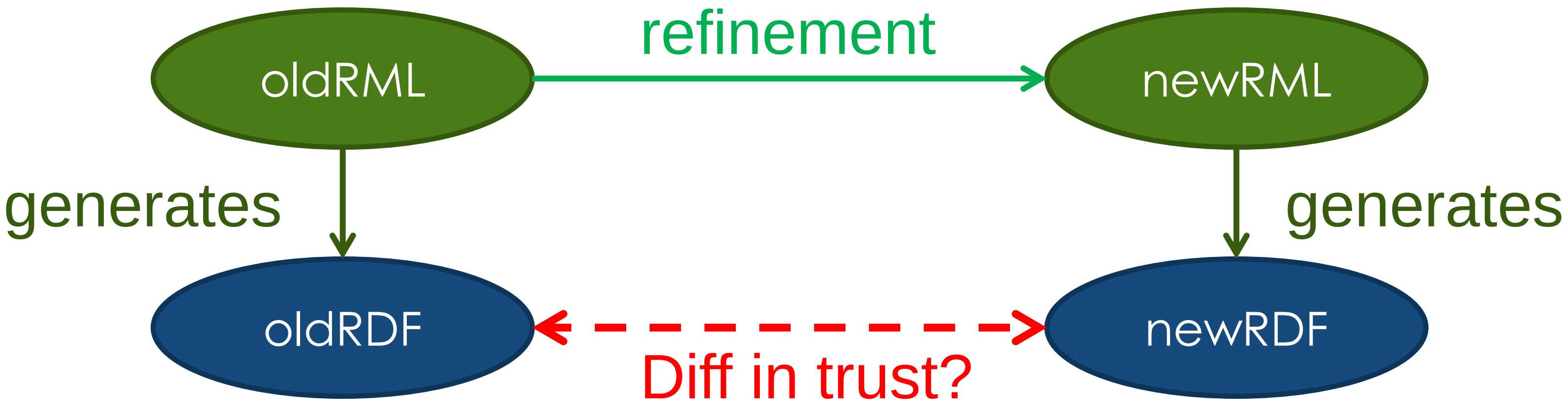
Based **on RDFUnit tests** for mapping documents instead of data

Turns out to be **much more efficient** for mapping documents than for data (seconds vs. hours)

Generates **violations** (warnings and errors),
based on which the mapping definitions are **refined**

Capturing provenance of the workflow

Goal:
evaluate the **difference (delta) in trust** between the new and old dataset



Deriving Trust

Query the provenance for **violations** and see:

- **How many** there are
- **How bad** they are (e.g., errors can be worse than warnings)

The cool thing: **it's all RDF**, so it can be done with standard reasoning tools (N3, SPARQL, …)