Test-driven Assessment of [R2]RML Mappings to Improve Dataset Quality

Anastasia Dimou[1], Dimitris Kontokostas[2], Markus Freudenberg[2], Ruben Verborgh[1], Jens Lehmann[2], Erik Mannens[1], Sebastian Hellmann[2], and Rik Van de Walle[1]

[1] Ghent University - iMinds - Multimedia Lab, Belgium
{firstname.lastname}@ugent.be
[2] Universitat Leipzig, Institut fur Informatik, AKSW, Germany
{lastname}@informatik.uni-leipzig.de

Abstract. RDF dataset quality assessment is currently performed primarily after data is published. Incorporating its results, by applying corresponding adjustments to the dataset, happens manually and occurs rarely. In the case of (semi-)structured data (e.g., CSV, XML), the root of the violations often derives from the mappings that specify how the RDF dataset will be generated. Thus, we suggest shifting the quality assessment from the RDF dataset to the mapping definitions that generate it. The proposed test-driven approach for assessing mappings relies on RDFUnit test cases applied over mappings specified with RML. Our evaluation is applied to different cases, e.g., DBpedia, and indicates that the overall quality of an RDF dataset is quickly and significantly improved.

Keywords: Linked Data Mapping, Data Quality, RML, R2RML, RDFUnit

1 Introduction

Although more and more data is published as Linked Data, there are significant variations in *quality* [6], commonly conceived as "fitness for use" for a certain application or use case. Similar violation patterns reoccur frequently, and most encountered violations are related to the dataset's schema, namely the vocabularies or ontologies used to annotate the data [4]. When datasets stem originally from semi-structured formats (e.g., csv, XML), the schema is derived from the set of classes and properties specified by the mappings which are applied repeatedly. Consequently, the same violations are repeated in the dataset as well. Lately, combinations of different ontologies and vocabularies are used to annotate data [5]. This increases the likelihood of such violations, as they often derive from incorrect usage or incorrect combinations of schemas in the mappings.

Taking mappings of data to RDF as a software engineering task, a set of unit test cases can be assigned to the mappings to ensure the correct generation of RDF datasets from input data. Incorporating quality assessment as part of the mapping is essential to prevent same violations to appear repeatedly within the dataset and over distinct entities. After all, structural adjustments can still be

applied in this phase, as violations are identified at their root. Furthermore, if mappings are assessed, every other new data source also mapped using them, directly benefit from the improvements. Therefore, we proposed a uniform solution [1] that incrementally assesses the quality of an RDF dataset, covering both the mappings and the dataset itself. In this work, we aim to elaborate more on how RDFUnit patterns [4] for dataset test cases were arose to cover RML mappings [2], too.

2 [R2]RML Mappings Quality Assessment with RDFUnit

Our solution relies on the RDF mapping language (RML) [2] that allows specifying mapping definitions expressed in RDF, and the RDFUnit validation framework due to its associated test-case-based architecture [3]. For our proof-of-concept implementation[3], RDFUnit test cases are applied to mappings defined with RML.

RML extends R2RML[4], the W3C recommended language for defining mappings of data in relational databases to RDF, and covers also mappings from sources in different semi-structured formats, such as CSV and JSON [2]. RML documents [2] specify how the input data can be represented in RDF. The main building blocks of RML documents are Triples Maps that defines how triples are generated and consists of three main parts: the Logical Source, the Subject Map and zero or more Predicate-Object Maps. Term Maps define how RDF terms (IRI, blank node or literal) is generated and can be *constant-valued* that always generates the same RDF term, or *reference-valued* that is the data value of a referenced data fragment in a given Logical Source, or *template-valued term map* that is a valid string template that can contain referenced data fragments of a given Logical Source.

RDFUnit [4] is an RDF validation framework inspired by test-driven software development. In RDFUnit, every vocabulary, ontology, dataset or application can be associated by a set of data quality test cases. The test case definition language of RDFUnit is SPARQL, convenient to directly query for identifying violations. For rapid test case instantiation, a pattern-based SPARQL-template engine, running over a library of common patterns[5], is supported where variables can be easily bind into patterns. RDFUnit has a *Test Auto Generator* (TAG) component. TAG searches for schema information and automatically instantiates new test cases.

As [R2]RML mappings can be processed as RDF documents, because of their native RDF representation and viewpoint (written as the generated triples), the same set of schema validation patterns normally applied on the RDF dataset is also applicable on the mappings that state how it is generated. Nevertheless, instead of validating the triple's predicate against its subject and object, the predicate is extracted from the Predicate Map and is validated against the Term Maps that generate the subject and object. To achieve this, the properties and

[3] https://github.com/mmlab/RMLValidator
[4] http://www.w3c.org/TR/R2RML
[5] https://github.com/AKSW/RDFUnit/blob/master/configuration/patterns.ttl

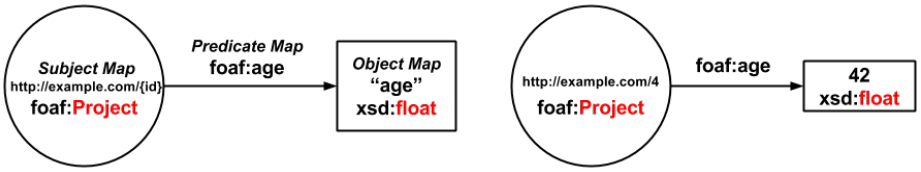

Fig. 1. (i) RML mapping (*left*) (ii) and corresponding triples (*left*) as generated.

classes are identified and their namespaces are used to retrieve the schemas and generate the test cases as if they were the actual dataset. The expected value, as derived from the Predicate Map, is compared to the defined one, as derived from the corresponding Object Map. For example, the extracted predicate is `foaf:age` which normally expects an integer datatype, but the Term Map that generates the object is defined to have a float. Its mapping document follows:

```
<#Mapping> rr:subjectMap [rr:template "http://example.com/{id}"; rr:class foaf:Project];
      rr:predicateObjectMap [rr:predicate foaf:age; rr:objectMap [rml:reference "age"]].
```

Corresponding RDFUnit test cases and patterns were defined to apply to the mappings, adjusting the assessment queries.[6] The defined test cases cover all possible alternative ways of defining equivalent mappings that generate the same triples. RDFUnit can annotate test cases by requesting additional variables and binding them to specific result properties. The test case patterns applied to the aforementioned example and its instantiation are indicatively presented. The following is the `where` clause of a test case that assesses the datatype and is applied to the dataset:

```
?resource %%P1%% ?c.
FILTER (DATATYPE(?c) != %%D1%%)
```

```
?resource foaf:age ?c.
FILTER (DATATYPE(?c) != xsd:int)
```

The following is the `where` clause of the same test case applied to the mapping:

```
?resource rr:predicateObjectMap ?poMap .
?poMap rr:predicate %%P1%% ;
       rr:objectMap ?objM .
?objM rr:datatype ?c .
FILTER (?c != %%D1%%)
```

```
?resource rr:predicateObjectMap ?poMap .
?poMap rr:predicate foaf:age ;
       rr:objectMap ?objM .
?objM rr:datatype ?c .
FILTER (?c != xsd:int)
```

3 Evaluation and Discussion

The assessed datasets and corresponding mappings, as well as the assessment results are summarized in Table 1: DBpedia mappings[7], after the mappings were converted from wikitext markup to RML[8], and its dataset were assessed. DBLP mappings were assessed, after the mappings were converted from D2RQ to RML[9],

[6] https://github.com/AKSW/RDFUnit/blob/master/data/tests/Manual/www.w3.org/ns/r2rml/rr.tests.Manual.ttl

[7] http://mappings.dbpedia.org/

[8] https://goo.gl/GPB1Ar

[9] https://github.com/RMLio/D2RQ_to_R2RML.git

	dataset assessment				*mapping assessment*				
dataset	**size**	**time**	**#fail.**	**#viol.**	**size**	**time**	**#fail.**	**#viol.**	**triples**
DBpEn	62M	16h	1,128	3.2M	115K	11s	1	160	255K
DBpNL	21M	1.5h	683	815K	53K	6s	1	124	106K
DBLP	12M	12h	7	8.1M	368	12s	2	8	8M
iLastic[10]	150K	12s	23	37K	825	15s	3	26	37K
CDFLG[11]	0.6K	7s	15	678	558	13s	4	16	631
CEUR-WS[12]	2.4K	6s	7	783	702	5s	3	12	783

Table 1. The number of triples (size), number of test cases, evaluation time, failed test cases and total individual violations appear for both *dataset* and *mapping assessment*.

and the corresponding dataset were assessed, too. The evaluation results show that the required quality assessment time is significantly reduced, especially in the case of medium/large datasets, when assessing the mappings compared to the complete RDF dataset. That improvement happens because the dataset assessment requires examining each triple separately to identify, for instance, that 12M triples violated the predicate's range, whereas mapping assessment requires only 1 triple to be examined. The effectiveness of mapping assessments is also high: in all cases where the dataset generation fully relies on its mappings, the majority of violations identified can be addressed.

Acknowledgements. This paper's research activities were funded by Ghent University, iMinds, the Institute for the Promotion of Innovation by Science and Technology in Flanders, the Fund for Scientific Research-Flanders and by grants from the EU's 7th & H2020 Programmes for projects ALIGNED (GA 644055), GeoKnow (GA 318159) and LIDER (GA 610782).

References

1. A. Dimou, D. Kontokostas, M. Freudenberg, R. Verborgh, J. Lehmann, E. Mannens, S. Hellmann, and R. Van de Walle. Assessing and Refining Mappings to RDF to Improve Dataset Quality. In *Proceedings of the 14th International Semantic Web Conference*, Oct. 2015.
2. A. Dimou, M. Vander Sande, P. Colpaert, R. Verborgh, E. Mannens, and R. Van de Walle. RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In *Workshop on Linked Data on the Web*, 2014.
3. D. Kontokostas, M. Brümmer, S. Hellmann, J. Lehmann, and L. Ioannidis. NLP data cleansing based on Linguistic Ontology constraints. In *Proc. of the Extended Semantic Web Conference 2014*, 2014.
4. D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, and A. Zaveri. Test-driven Evaluation of Linked Data Quality. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 747–758, 2014.
5. M. Schmachtenberg, C. Bizer, and H. Paulheim. Adoption of the Linked Data Best Practices in Different Topical Domains. volume 8796 of *LNCS*. Springer, 2014.
6. A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality Assessment for Linked Data: A Survey. *Semantic Web Journal*, 2015.

[10] `http://explore.ilastic.be/`

[11] `http://ewi.mmlab.be/cd/all`

[12] `http://rml.io/rml/data/SPC2015/`