DBpedia Mappings Quality Assessment*

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

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

Abstract. The root of schema violations for RDF data generated from (semi-)structured data, often derives from mappings, which are repeatedly applied and specify how an RDF dataset is generated. The DBpedia dataset, which derives from Wikipedia infoboxes, is no exception. To mitigate the violations, we proposed in previous work to validate the mappings which generate the data, instead of validating the generated data afterwards. In this work, we demonstrate how mappings validation is applied to DBpedia. DBpedia mappings are automatically translated to RML and validated by RDFUnit. The DBpedia mappings assessment can be frequently executed, because it requires significantly less time compared to validating the dataset. The validation results become available via a user-friendly interface. The DBpedia community takes them into consideration to refine the DBpedia mappings or ontology and thus, increase the dataset quality.

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

1 Introduction

Although more and more data is published as Linked Data, there are significant variations in *quality* [5], commonly conceived as "fitness for use" for a certain application or use case. 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, if those mappings contain inaccuracies, the same violations are repeated over and over in the dataset. Incorporating quality assessment *as part of the mapping* activity is therefore essential to prevent most recurring schema-based violations. To this end, we have proposed a uniform approach for assessing the mappings and dataset quality [1]. We implemented our approach based on the RDFUnit validation framework [3] and the RML mapping language [2]. Our solution incrementally assesses the quality of an RDF dataset, covering both the mappings and the

* 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 grants from the EU's 7th & H2020 Programmes for the projects ALIGNED (GA 644055), GeoKnow (GA 318159) and LIDER (GA 610782).

dataset itself. Since RML mappings are expressed in RDF, the RDFUnit validation framework can apply its test cases to RML mappings similarly to how it applies them to RDF datasets. Assessing an RDF dataset requires a lot of time, thus it cannot be frequently executed, and, when it happens, the violations' root is not intuitively detected. On the contrary, directly assessing mappings that generates a dataset requires significantly less time and the violation root is detected.

In this work, we demonstrate how we incorporated our solution in the DBpedia validation workflow. DBpedia mappings are automatically translated to RML and subsequently assessed using RDFUnit. In this demo, the validation results will be shown via a user friendly interface and users can directly contribute to improve the DBpedia mappings. Once they update a mapping or the DBpedia ontology, users will be able to trigger a new validation round and immediately see the updated validation results, without the violation they just corrected.

2 Expressing DBpedia Mappings with RML

DBpedia [4] provides a *collaborative mapping approach* of Wikipedia infoboxes to the DBpedia ontology[3]. The mappings are maintained and edited through the DBpedia mappings wiki[4], using the same wiki markup syntax as Wikipedia to define the mappings. However, the quality of wikitext-based mappings cannot be assessed directly, and certainly not in the same way as the resulting dataset.

RML covers mappings from sources in different (semi-)structured formats. Furthermore, it is highly scalable towards other structures and formalizations. Taking advantage of this, we introduced *wikitext serialisation* as a new Reference Formulation. A Reference Formulation is used to indicate the grammar which should be used to refer to data of a certain structure and format. 674 distinct mapping documents for English, 463 for Dutch and a total of 4,468 for all languages supported in the DBpedia mappings wiki are translated to RML and are available at `http://mappings.dbpedia.org/server/mappings/en/pages/rdf/`.

A DBpedia mapping follows for the Infobox Person[5]:

```
{{TemplateMapping
| mapToClass = Person
| mappings =
  {{PropertyMapping | templateProperty = name | ontologyProperty = foaf:name }}
  {{PropertyMapping | templateProperty = birth_date | ontologyProperty = birthDate }}
  {{PropertyMapping | templateProperty = birth_place | ontologyProperty = birthPlace }}}}
```

and its corresponding RML mapping, after being translated to RML is [6]:

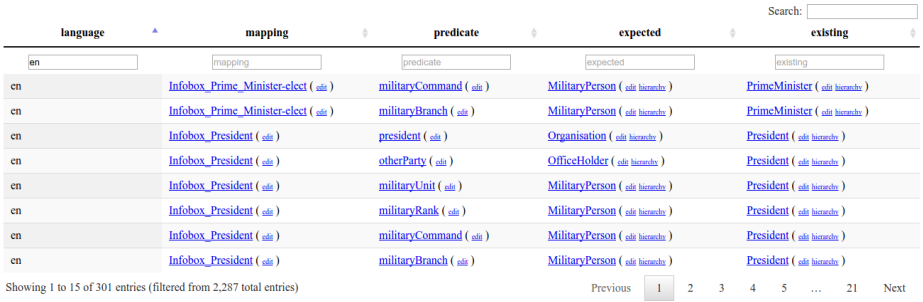
```
<http://mappings.dbpedia.org/server/mappings/en/Infobox_person>
    rr:subjectMap [ rr:class dbpedia:Person ; rr:termType rr:IRI ;
        rr:constant "http://dbpedia.org/resource/Template:Infobox_person" ] ;
    rr:predicateObjectMap [ rr:predicate dbpedia:birthPlace ;
        rr:objectMap [ a rr:ObjectMap ;  rml:reference "birth_place". ] ] .
```

[3] `http://wiki.dbpedia.org/Ontology`

[4] `http://mappings.dbpedia.org`

[5] `http://mappings.dbpedia.org/index.php?title=Mapping_en:Infobox_person&action=edit`

[6] The example is adjusted to improve reading. A full RML transformation can be found at `http://mappings.dbpedia.org/server/mappings/en/pages/rdf/Mapping_en%3AInfobox_person`



Search:

language	mapping	predicate	expected	existing
en	mapping	predicate	expected	existing
en	Infobox_Prime_Minister-elect (edit)	militaryCommand (edit)	MilitaryPerson (edit hierarchy)	PrimeMinister (edit hierarchy)
en	Infobox_Prime_Minister-elect (edit)	militaryBranch (edit)	MilitaryPerson (edit hierarchy)	PrimeMinister (edit hierarchy)
en	Infobox_President (edit)	president (edit)	Organisation (edit hierarchy)	President (edit hierarchy)
en	Infobox_President (edit)	otherParty (edit)	OfficeHolder (edit hierarchy)	President (edit hierarchy)
en	Infobox_President (edit)	militaryUnit (edit)	MilitaryPerson (edit hierarchy)	President (edit hierarchy)
en	Infobox_President (edit)	militaryRank (edit)	MilitaryPerson (edit hierarchy)	President (edit hierarchy)
en	Infobox_President (edit)	militaryCommand (edit)	MilitaryPerson (edit hierarchy)	President (edit hierarchy)
en	Infobox_President (edit)	militaryBranch (edit)	MilitaryPerson (edit hierarchy)	President (edit hierarchy)

Showing 1 to 15 of 301 entries (filtered from 2,287 total entries)

Previous 1 2 3 4 5 … 21 Next

Fig. 1. Screenshot of a violations list presented to the DBpedia community. For every violating mapping, the predicate with the existing RDF term, according to the corresponding DBpedia mapping, and the expected value, according to the DBpedia ontology are presented.

3 DBpedia Mappings Quality Assessment

Since RML mappings can be processed as RDF documents, and are written from the viewpoint of the generated triples, the same set of schema validation patterns normally applied to the RDF dataset is also applicable to the mappings that state how the dataset is generated. RDFUnit was extended to also support quality assessment over RML mappings [1]. Indicatively, instead of validating each triple's predicate from the final RDF dataset against its subject and object, the predicate is extracted from the Predicate Map, that defines what the triple's predicate will be in RML, and is validated against the Term Maps that define how the subject and object will be generated. The expected value, as derived from the DBpedia ontology, is compared to the specified one, as derived from the corresponding mapping. To achieve this, the schemas and their namespaces are retrieved and the test cases are generated as if they were the actual dataset. For instance, an extracted predicate expects a Literal as object according to the DBpedia ontology, but the mapping that defines how the object is generated specifies that a resource should be generated instead; in this case a violation is reported.

To systematically validate DBpedia mappings and have up-to-date reports, we created a script[7] to trigger all DBpedia mappings validation and is executed every night. The script exports the DBpedia mapping violations as a JSON file that, in turn, is visualized (cf. Figure 1) using a user-friendly interface which is available at `http://mappings.dbpedia.org/validation`. The assessment and report generation is automated, streamlined, and frequently executed. The DBpedia community uses the violations list as feedback to correct violating mappings or enhance the DBpedia ontology and, thus, improves the dataset's quality.

DBpedia Mappings and Dataset Assessment

We compared the DBpedia 2014 release assessment to the DBpedia mappings assessment. English and Dutch DBpedia mappings as well as DBpedia mappings of

[7] `https://github.com/AKSW/RDFUnit/blob/master/rdfunit-examples/src/main/java/org/aksw/rdfunit/examples/DBpediaMappingValidator.java`

all 27 supported languages were validated. The results show that the quality assessment time is significantly reduced when assessing the mappings compared to the complete RDF dataset. It takes only 11 seconds to assess the English DBpedia mappings, while assessing the whole DBpedia dataset takes 16 hours, because the dataset assessment requires examining each triple separately to identify, for instance 12M triples violating the range of foaf:primaryTopic. Mapping assessment requires only 1 triple to be examined. Indicatively, the evaluation of all mappings for all 27 language editions resulted in a total of 1316 domain-level violations.

	dataset assessment			*mapping assessment*		
dataset	**#triples**	**time**	**#viol.**	**#triples**	**time**	**#viol.**
DBpEn	62M	16.h	3.2M	115K	11s	160
DBpNl	21M	1.5h	815K	53K	6s	124
DBpAll	–	–	–	511K	32s	1,316

Table 1. For each of the DBpedia *dataset* and *mapping assessment*, the number of triples, evaluation time and total individual violations appear respectively.

The latest DBpedia releases rely on results of this work[8]. We currently incorporate the RML toolchain in the DBpedia extraction framework[9] and plan to integrate the mapping validation in the editing step and, thus, prevent the creation of violating mappings. This will enable the complete assessment and refinement workflow use [1] to automatically improve the DBpedia dataset quality.

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. J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Kleef, S. Auer, and C. Bizer. DBpedia - a Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Sem. Web Journal*, 2014.
5. A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality Assessment for Linked Data: A Survey. *Semantic Web Journal*, 2015.

[8] https://github.com/dbpedia/mappings-tracker/issues/57

[9] https://github.com/dbpedia/dbpedia-gsoc/wiki/2016-Integrating-RML-in-the-DBpedia-extraction-framework