

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

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

In: Proceedings of the 4th International Workshop on Methods for Establishing Trust of (Open) Data.(METHOD2015), 2015.

http://trustingwebdata.org/method2015/papers/METHOD_2015_paper_4.pdf

# 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

Ghent University - iMinds - Multimedia Lab, Belgium
{firstname.lastname}@ugent.be

**Abstract.** Assessing the trustworthiness of a dataset is of crucial importance on the Web of Data. In the case of Linked Data derived from (semi-)structured data, the trustworthiness of a dataset can be assessed partly through their mapping definitions, which are often neglected. However, an approach was proposed to assess and refine such mapping definitions, which was proven to be more effective than assessing and refining the quality of a dataset directly. In this paper, we derive important provenance from mappings quality assessment and refinement, enabling us to assess the relative trustworthiness of the datasets they generate.

## 1 Introduction

The ever increasing adoption of Linked Data causes data owners to look for ways to efficiently publish their data on the Web, e.g. through an automatic mapping process. Even though these mappings are mainly performed by data owners, data consumers must assess the quality of the datasets they consider to use and decide whether they trust the dataset for further use or not.

In most cases, Linked Data Quality Assessment (QA) is focused and applied to data that is already published and it is performed by each interested party whenever necessary. In the case of (semi-)structured data mapped to the RDF data model however, the most crucial moment to assess them for their quality is after the data is mapped and before it is published, allowing the data publisher to better react to any issues. To this end, we proposed a new approach to automatically assess and refine mapping documents to improve dataset quality [1]. In this paper, we expand upon this idea and incorporate the capture of standardized provenance information (in W3CPROV [6]) during the quality and assessment process of mapping documents, helping a data consumer interpret the relative trustworthiness of mapped data.

## 2 Related Work

To the best of our knowledge, no other approaches currently exist to expose the provenance of mapping data to RDF, or of its quality assessments and refinements in interoperable form. Although trust assessment approaches exist in

literature, we found none specific to mappings. The closest related work we found is TRAMP [4], a system to help understand the transformations performed in complex schema mappings, through their provenance. TRAMP is mostly focused on relational database schemas, and predates the W3C PROV standard.

## 3 Mapping Assessment and Refinement Workflow

In previous work, we contributed to the RMLValidator [1], a uniform, iterative, incremental assessment and refinement workflow that produces a high-quality RDF dataset. The RMLValidator is based on applying the RDFUnit validation framework [5] to mappings described with the RDF mapping language (RML) [2].

RDFUnit is a validation framework for RDF, inspired by the unit tests commonly applied in software development. In RDFUnit, the SPARQL language is used to define a set of data quality test cases (generic of user-defined) for every vocabulary, ontology, dataset or application. By using SPARQL, violations can be easily identified because they can be directly queried for. The RMLValidator incrementally assesses the quality of an RDF dataset, covering both the mappings and the dataset itself, by using the following workflow: The RML mapping definitions are assessed against quality assessment measures. The violations identified during this *Mapping Quality Assessment (*MQA*)* are reported and are taken into consideration to refine the definitions. The MQA may be repeated until the mapping definitions can not be further refined. The refined version is then used to generate an RDF representation of either a sample of the data or the complete data. The generated RDF dataset is then assessed, using the same quality assessment framework. This *Dataset Quality Assessment (*DQA*)* can also be repeated until a final, refined version of the mapping definitions is generated. The latter is then used to perform the actual mapping.

## 4 Provenance of Mappings & Trust Assessment

Provenance can be manually asserted or automatically collected while mapping the data [3]. To expose it in an interoperable, machine-interpretable way, W3C recommended the PROV specification [6]. Additionally, there are several occasions in the workflow described in Section 3 where provenance may be logged. On a high level, these four stages are: A) mapping the original data to RDF using the original mapping document; B) assessing and refining the quality of the mapping document on its own (MQA); C) assessing and refining the mapping document quality even further through a data sample (DQA); D) mapping the data to new RDF using the improved mapping document. In Figure 1, we provide a general overview of the provenance that can be logged during these four stages. The symbols used correspond to those used in the W3C PROV specifications (ellipses for `prov:Entity`, and rectangles for `prov:Activity`, and directed arrows for relations between them). Note that in PROV, the direction of the relations – and thus, the arrows in Figure 1 – is inverse to that of the actual workflow, which

might seem counter-intuitive at first glance. In the remainder of this section, we describe each step in detail.

*A) Provenance of Original Mapping Document* In Figure 1A, we see which provenance elements are generated when the original data is mapped[1]. In this step, the original data is retrieved, and used by the mapping activity, which also uses a mapping document in order to generate RDF. This provenance corresponds to the following PROV-O triples:

```
:originalData a prov:Entity .
:dataRetrieval a prov:Activity ;  prov:used :originalData .
:data a prov:Entity ;  prov:wasGeneratedBy :dataRetrieval .
:mapDoc a prov:Entity .
:mapping a prov:Activity ;  prov:used :data, :mapDoc .
:rdf a prov:Entity ;  prov:wasGeneratedBy :mapping .
```

*B) Provenance of Mappings Quality Assessment and Refinement* The next step is to assess and refine the quality of the aforementioned mapping document. The first MQA activity might generate a number of violations. These are then used by the first refinement activity, which generates a new mapping document. To see whether this new mapping document actually represents an improvement over the old one, its quality is assessed again. This second MQA activity generates a new set of violations, which can then be compared to the previous ones, and used for a second refinement activity which generates a second new mapping document. This process is typically repeated until an optimal situation is achieved (e.g., when the violations remain the same). In Figure 1B, two such repetitions are shown. The PROV-O exposed by this process is:

```
:mapDoc a prov:Entity .
:qualityAssessment1 a prov:Activity ;  prov:used :mapDoc .
:violation1 a prov:Entity, :violationTypeA ;
        prov:wasGeneratedBy :qualityAssessment1 .
:violation2 a prov:Entity, :violationTypeA ;
        prov:wasGeneratedBy :qualityAssessment1 .
:violation3 a prov:Entity, :violationTypeB ;
        prov:wasGeneratedBy :qualityAssessment1 .
:refinement1 a prov:Activity ;
        prov:used :mapDoc, :violation1, :violation2, :violation3 .
:refinedMapDoc1 a prov:Entity ; prov:wasGeneratedBy :refinement1 .
:qualityAssessment2 a prov:Activity ;  prov:used :refinedMapDoc1 .
:violation4 a prov:Entity, :violationTypeB ;
        prov:wasGeneratedBy :qualityAssessment2 .
:violation5 a prov:Entity, :violationTypeC ;
        prov:wasGeneratedBy :qualityAssessment2 .
:refinement2 a prov:Activity ;
        prov:used :refinedMapDoc1, :violation4, :violation5.
:refinedMapDoc2 a prov:Entity ; prov:wasGeneratedBy :refinement2 .
```

[1] Note that in order to assess and refine a mapping, step A) does not actually need to be executed, but it is essential to record the provenance of the data published using the original mapping.

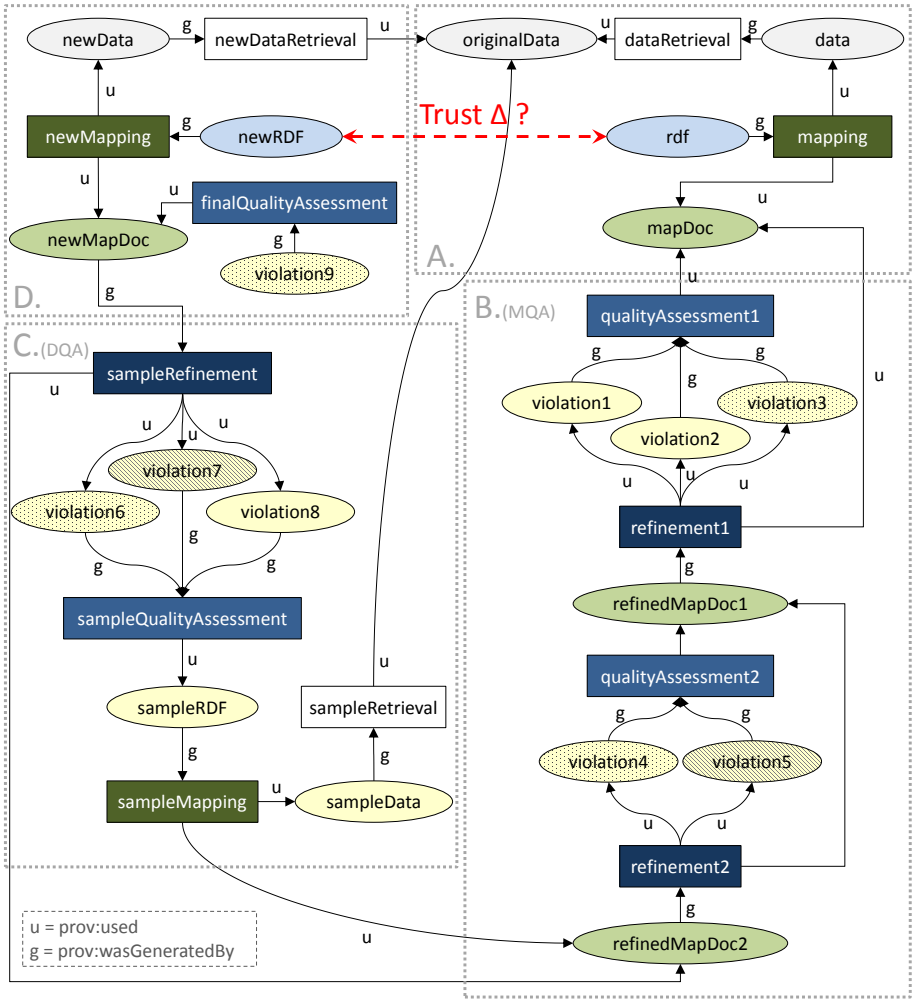


**Fig. 1.** Overview of the provenance of data generated while using, assessing and refining a mapping document. The figure describes A. the normal mapping situation (without refinement); B. the quality assessment and refinement workflow of the mapping as such (MQA); C. the DQA of the mapping (through the mapping of a data sample); D. the mapping of new data using the final refined mapping document. Note that in a realistic scenario, the MQA in step B might be repeated a number of times to refine the mapping document optimally (two repetitions are shown here). Additionally, note that the original data does not change, only its RDF representations generated using the different mappings. By reasoning over the provenance of these RDF representations, a different level of trust (delta) may be assigned to each of them.

Note that, apart from their `prov:Entity` type, each violation also has its own specific violation type. In Figure 1B, this is indicated by the background pattern of the entities. This gives us information about how effective the first refinement step was. For example, in Figure 1B we can see that although `refinement1` eliminated the violation type of `violation1` and `violation2`, the violation type of `violation3` is still present in `violation4` after the refinement, and a new violation type is even introduced through `violation5`. This means that the new mapping document has less violations than the old one, however we do not know if they are more or less severe. This is important for the trust assessment of the refined mapping document, as is further discussed in Section 4.

*C) Provenance of Dataset Quality Assessment* The next step in the process is to retrieve a sample of the original data (or even the entire dataset), and use this in a sample mapping activity, together with the refined mapping document from step B. This process is illustrated by Figure 1C. The sample mapping activity generates a sample of RDF data, which is then used by a sample DQA activity. This DQA activity generates violations, which are then used in a sample refinement activity of the mapping document, if possible. This generates a final, refined mapping document to be used in the final mapping step. The triples generated by this step are described by the following PROV-O:

```
:sampleRetrieval a prov:Activity ; prov:used :originalData .
:sampleData a prov:Entity ; prov:wasGeneratedBy :sampleRetrieval .
:sampleMapping a prov:Activity ; prov:used :sampleData, :refinedMapDoc2 .
:sampleRDF a prov:Entity ; prov:wasGeneratedBy :sampleMapping .
:sampleQualityAssessment a prov:Activity ; prov:used :sampleRDF .
:violation6 a prov:Entity, :violationTypeB ;
        prov:wasGeneratedBy :sampleQualityAssessment .
:violation7 a prov:Entity, :violationTypeC ;
        prov:wasGeneratedBy :sampleQualityAssessment .
:violation8 a prov:Entity, :violationTypeA ;
        prov:wasGeneratedBy :sampleQualityAssessment .
:sampleRefinement a prov:Activity ;
        prov:used :refinedMapDoc2, :violation6, :violation7, :violation8 .
:newMapDoc a prov:Entity ; prov:wasGeneratedBy :sampleRefinement .
```

*D) Final Mapping* The final mapping, as shown in Figure 1D, is performed in the same way as described in Section 4, except that the new mapping activity uses the new mapping document instead of the original one to generate new RDF representation from the original data. Additionally, to ensure that we have a complete provenance trace, a final MQA step is performed to find out which violations remain. The PROV-O triples that are generated during this final step are:

```
:newDataRetrieval a prov:Activity ; prov:used :originalData .
:newData a prov:Entity ; prov:wasGeneratedBy :newDataRetrieval .
:finalQualityAssessment a prov:Activity ; prov:used :newMapDoc .
:violation9 a prov:Entity ; prov:wasGeneratedBy :finalQualityAssessment .
:newMapping a prov:Activity ; prov:used :newData, :newMapDoc .
:newRDF a prov:Entity ; prov:wasGeneratedBy :newMapping .
```

*Trust Interpretation* We now have all the provenance recorded to make an assessment of the difference in trustworthiness between the RDF generated using the original mapping document, and the RDF generated using the new one. This can be achieved by creating reasoning rules and/or queries over the provenance, combined with semantic information available on the various violations.

For example, the provenance of the mapping workflow in Figure 1 allows us to make trust statements based on two criteria. On the one hand, we can simply observe the number of refinements and violations to suggest a trust assessment to the consumer. In the example, this would tell us that it is probably better to trust `:newRDF`, which has only one violation, than `:rdf`, which has three. On the other hand, it could be that `:violation9` is actually much more severe than `:violation1`,`:violation2` and `:violation3` combined. Therefore, we also propose to create semantics-based rules, which go deeper and report on the types of refinement performed, and the gravity of the violations (e.g., errors or warnings).

## 5 Discussion and Future Work

We showed that it is feasible to enable the inference of trust assessments by exposing the provenance of a mapping quality assessment and refinement workflow. However, a number of challenges remain. Richer semantics are needed to describe the results (e.g., violations) of quality assessments, and the implications they have on the trustworthiness of generated data. Here, a part of the responsibility lies with the quality assessment approaches. There have been promising initiatives in this direction, such as the Test-Driven Data Validation Ontology[2] associated with the RDFUnit system, which we will investigate in future work.

## References


[1] Dimou, A., Kontokostas, D., Freudenberg, M., Verborgh, R., Lehmann, J., Mannens, E., Hellmann, S., Van de Walle, R.: Assessing and Refining Mappings to RDF to Improve Dataset Quality. In: Proceedings of the 14th ISWC (2015)

[2] Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In: Workshop on Linked Data on the Web (2014)

[3] Dimou, A., Vander Sande, M., De Nies, T., Mannens, E., Van de Walle, R.: RDF Mapping Rules Refinements according to Data Consumers' Feedback. In: 23rd International Conference on World Wide Web Companion. pp. 249–250 (2014)

[4] Glavic, B., Alonso, G., Miller, R.J., Haas, L.M.: Tramp: Understanding the behavior of schema mappings through provenance. VLDB Endowment 3(1-2) (2010)

[5] Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., Zaveri, A.: Test-driven Evaluation of Linked Data Quality. In: 23rd International Conference on World Wide Web (2014)

[6] Moreau, L., Missier, P., W3C Provenance Working Group: PROV-DM: The PROV Data Model. W3C Recommendation 30 April (2013)


[2] `http://rdfunit.aksw.org/ns/core#`