Assessing Quality of R2RML Mappings for OSi's Linked Open Data Portal

Alex Randles[1*][0000-0001-6231-3801] and Declan O'Sullivan[1][0000-0003-1090-3548]

[1] ADAPT Centre, Trinity College Dublin, Dublin 2, Ireland
{alex.randles,declan.osullivan}@adaptcentre.ie

Abstract. As the number of geospatial Linked Data datasets being published grows, so does the need to ensure their quality and trustworthiness. The quality assessment of these datasets is most often assessed after the dataset has been published, however, due to the authoritative nature of geospatial data, we propose bringing quality assessment earlier into the Linked Data generation process itself. In order to create these datasets, artifacts are required to be defined called 'uplift mappings'. These uplift mappings use the R2RML specification language to define the relationship between the non-RDF geospatial data and its Linked Data RDF representation. This paper describes a mapping quality framework which will assess and refine the quality of the R2RML uplift mappings using a number of quality metrics. We demonstrate the use of our framework in the publication pipeline for Ordnance Survey Ireland's (OSi) Linked Open Data portal for geospatial data, http://data.geohive.ie. The use of the R2RML quality framework early in the publication pipeline provides significant confidence in the quality of the resulting linked data geospatial data published through the portal.

Keywords: Geospatial data, Linked data, Data Quality, Uplift mappings.

1 Introduction

Increasingly geospatial data is being exposed using W3C's Linked Data[1] approach, which allows this data to be easily consumed in a machine-readable manner using standard web technologies, thus making the interlinking of multiple data sources much easier. However, due to the expectation that geospatial data provided by National Mapping Agencies are authoritative, a high level of quality control is required throughout the creation process.

An example of one such project involves a collaboration between the Science Foundation Ireland ADAPT Research Centre[2] and Ordinance Survey Ireland (OSi)[3].

[1] https://www.w3.org/standards/semanticweb/data

[2] ADAPT homepage at http://www.adaptcentre.ie

[3] OSi homepage at https://www.osi.ie/

The resulting Linked Open Data portal available at data.geohive.ie (see **Fig. 1**) involves taking selected geospatial data stored using a relational database model called Prime2 and making it available as Linked Open Data [1]. Prime2 stores information on over 45,000,000 spatial objects representing key geospatial features in Ireland. Converting the relational data stored in Prime2 into the RDF format needed for Linked Open Data, required the creation of the OSi Spatial Ontology[4] , as a suitable ontology was not found, to accurately represent their geospatial data. R2RML[5] uplift mappings are created by domain experts to specify how the geospatial data in relational format is to be transformed into RDF according to the OSi spatial ontology.

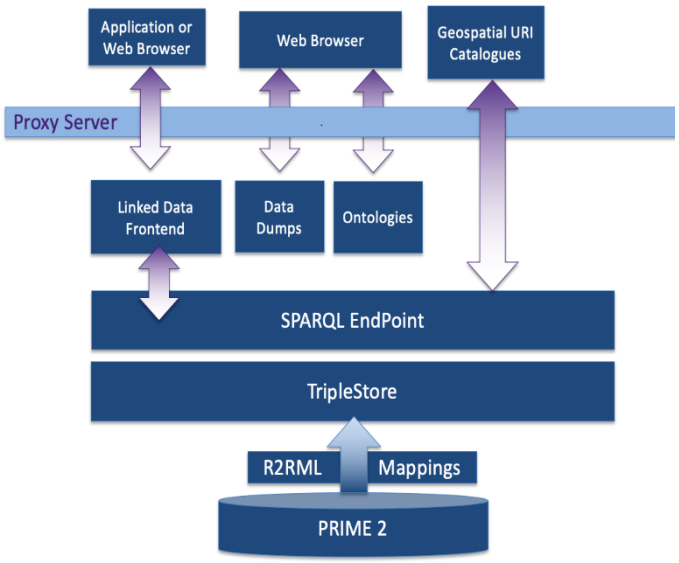

Fig. 1. data.geohive.ie Technical Architecture

In this paper, we describe the quality improvement which can be offered by our R2RML quality framework in the production of the R2RML uplift mappings. Assessing and refining the quality of the uplift mappings used to create the geospatial Linked Open Data will prevent errors within the uplift mappings causing significant number of quality issues within the resulting RDF dataset [1]. The R2RML quality framework allows users to produce higher quality mappings and datasets, while also facilitating the maintenance and reuse of those mappings [2].

The remainder of this paper is organized as follows: Section 2 describes a general overview of our R2RML quality framework. Section 3 demonstrates a walkthrough of our framework executed on an example from OSi's set of geospatial R2RML map-

[4] OSi Spatial Ontology at `http://ontologies.geohive.ie/osi`
[5] R2RML specification at `https://www.w3.org/TR/r2rml/`

pings. Section 4 discusses related work. Section 5 concludes our paper and discusses future work.

2 Mapping Quality Framework

In this section we briefly describe the mapping quality framework under development which assesses and refines the quality of R2RML mappings used to generate RDF datasets. The rationale for choosing R2RML as our target language for mapping quality assessment and refinement is that it is the W3C recommendation for mapping relational databases to RDF datasets and has wide uptake.

We previously designed a Mapping quality framework [2] using SHACL constraints language[6], which can be used to validate all data in RDF format. Within this previous framework, a machine-readable report on R2RML mappings is generated using SHACL's validation report vocabulary. Furthermore, SPARQL queries are then used to update and refine the mappings, since they are defined in RDF format. However, SHACL is not designed specifically for validating mappings, and we concluded that a new framework design which is domain specific would allow users to capture more detailed provenance and metadata relating to the quality information related to the mappings.

Our updated framework design is split into two main stages: mapping assessment and mapping refinement. The framework is designed using a web-based Python application which can execute SPARQL queries on the mapping using the RDFLib[7] library, allowing the framework to query and update the mappings. Furthermore, the machine-readable reports generated by our framework are defined in a domain specific vocabulary called the Mapping Quality Vocabulary (MQV)[8] [3] which we developed to enable quality metadata and provenance information relating to the assessment and refinement of mappings to be captured and published.

Our framework design involves the users uploading an R2RML mapping and an optional local ontology. A local ontology refers to an ontology which is not available remotely. After these have been uploaded by the users, each remote vocabulary used within the mapping is fetched and stored in a local cache to speed up execution time. These vocabularies are queried by the framework in order to generate vocabulary specific quality metrics. Furthermore, the quality metrics are designed such that the framework can provide suggested semi-automatic refinements to rectify identified violations to quality metrics[9]. These refinements can be selected by the users and executed on the mapping in the framework in order to produce a refined quality-

[6] SHACL specification at `https://www.w3.org/TR/shacl/`

[7] RDFLib documentation at `https://rdflib.readthedocs.io/en/stable/`

[8] Mapping Quality Vocabulary Specification available at `https://alex-randles.github.io/MQV/`

[9] Quality metrics and refinements at `https://docs.google.com/spreadsheets/d/1165CWRjE3gDxyLy3qL9BBukHu7oR_zXVznGvxubUxbU/edit?usp=sharing`

improved R2RML mapping. Moreover, the framework uses MQV to capture metadata and provenance relating to the quality assessment and refinement of the mappings.

3 Demonstration Walkthrough

In this section we present a running example to demonstrate the quality assessment and refinement of a sample R2RML mapping[10]. The mapping has been extracted from the R2RML mappings used to generate OSi's linked data for data.geohive.ie, in this case related to geometry of a townland. For illustrative purposes this R2RML mapping has been edited to include an undefined property (geo:asWTK), rather than the correct defined property (geo:asWKT). If this minor spelling mistake was not spotted before execution, it could easily result in each triple generated from the townland relational table using this R2RML mapping to be incorrectly represented in the resulting linked data dataset. **Fig. 2** shows a screenshot of our frameworks user interface after it has assessed the quality of the sample mapping. The framework highlights the predicate and object that violate one of the quality metrics for R2RML mappings in red under the "Location" heading on right hand side of the Figure. This enables a user to quickly identify the issue in the R2RML mapping.

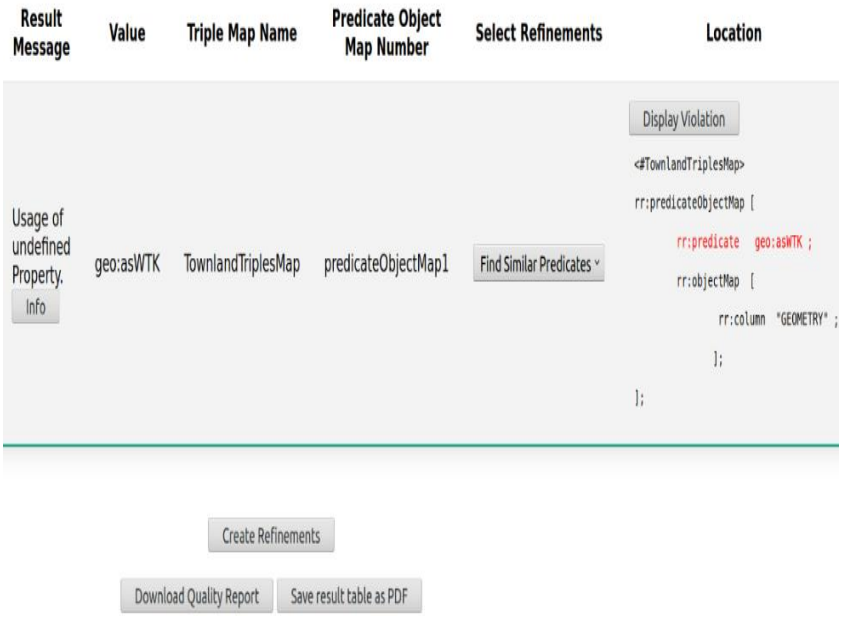

Fig. 2. Screenshot of R2RML Mapping Quality Assessment Framework: Violation reporting & Refinement selection

[10] Sample R2RML Mapping at `https://github.com/alex-randles/GeoLD2021-Paper-Examples/blob/main/sample_mapping.ttl`

A machine-readable quality report[11] shown in **Listing 1** is generated using the Mapping Quality Vocabulary (MQV). This report describes the violation (`ex:violation-0`) which was shown in human-readable format in **Fig. 2**. This quality report details important information relating to the violation. Such as quality metric (`mqv:metricD2`) which detected the violation, its location within the mapping (`<#TownlandTriplesMap>`) and a result message which describes the violation in a human-readable format (`"Usage of undefined Property."`).

```
ex:violation-0     a  mqv:MappingViolation ;
    mqv:hasLocation   "predicateObjectMap1" ;
    mqv:hasValue      geo:asWTK ;
    mqv:inTripleMap   <#TownlandTriplesMap> ;
    mqv:isDescribedBy mqv:metricD2 ;
    mqv:resultMessage "Usage of undefined Property." ;
    mqv:wasRefinedBy  ex:refinement-0 .
```

Listing 1: Extract of quality report generated

After quality violations have been detected within an R2RML mapping, they should be refined to prevent violations within the mapping replicating within the Linked Data dataset generated [1]. Refining the violation detected within this mapping which relates to an 'undefined property' can be accomplished either semi-automatically or manually. Semi-automatic refinement involves the framework suggesting several properties similar to the undefined property and allowing the users the option to input a new property into the framework. Manual refinement involves the users editing the mapping manually using a text editor or similar. If the user chooses to semi-automatically refine the mapping using our framework, a refined mapping and validation report[12] will be output. The validation report details the refinement (`ex:refinement-0`) associated with the violation detected within the mapping. Furthermore, the refinement is associated with the SPARQL query (`mqv:hasRefinementQuery`) which created the refined mapping.

4 Related work

EvaMap [4] is a framework which generates a global quality score for each mapping and provides feedback to the users, however, this feedback is not machine-readable. A test driven approach [5] which extends an existing framework called RDFUnit[13] in order to execute SPARQL queries on the mappings. The quality report generated can

[11] Quality report at `https://github.com/alex-randles/GeoLD2021-Paper-Examples/blob/main/quality_report.ttl`

[12] Validation report at `https://github.com/alex-randles/GeoLD2021-Paper-Examples/blob/main/validation_report.ttl`

[13] `http://rdfunit.aksw.org/`

be represented using the RDFUnit ontology which has not been designed for the purpose of capturing mapping provenance and metadata. Resglass [6] is a framework which uses a rule-driven methodology to rank mapping rules based on a score. Furthermore, no machine-readable quality report is generated and the rules are inspected by experts based on the scores. Another approach [1] extends an existing quality assessment tool called Luzzu[14]. This approach doesn't refine the violations detected within the mappings.

5 Conclusion and Future Work

Exposing geospatial data in RDF format requires artifacts to be defined called mappings, which define the relationship between the data sources. Creating suitable mappings requires the knowledge of domain experts [7]. However, this creation process is error prone and can result in poor quality geospatial Linked data being published. Furthermore, the authoritative nature of this data requires high quality for consumers.

Introducing mapping quality assessment and refinement into the geospatial Linked data publication process will result in higher quality and more trustworthy data being published and consumed by third parties. This paper describes and demonstrates a mapping quality framework which implements several quality metrics and refinements which focus on common quality issues found within mappings.

Future work will include the implementation of further metrics and refinements which will allow the framework more expressive capabilities in improving the quality of mappings. Furthermore, an extensive system and user evaluation of the framework, as well as improvements based on evaluation results.

Acknowledgements. This research was conducted with the financial support of the SFI AI Centre for Research Training under Grant Agreement No. 18/CRT/6223 at the ADAPT SFI Research Centre at Trinity College Dublin. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant #13/RC/2106.

References

1. Junior, A.C., Debattista, J., O'Sullivan, D.: Assessing the Quality of R2RML Mappings. In: Kaffee, L.-A., Endris, K.M., Vidal, M.-E., Comerio, M., Sadeghi, M., Chaves-Fraga, D., and Colpaert, P. (eds.) Joint Proceedings of the 1st International Workshop On Semantics ForTransport and the 1st International Workshop on Approaches for MakingData Interoperable co-located with 15th Semantics Conference (SEMANTiCS2019), Karlsruhe, Germany, September 9, 2019. CEUR-WS.org (2019).
2. Randles, A., Crotti Junior, A., O'Sullivan, D.: A Framework for Assessing and Refining the Quality of R2RML mappings. In: Proceedings of the 22nd International Conference on Information Integration and Web-Based Applications & Services.

[14] https://github.com/Luzzu/Framework/tree/V5

Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3428757.3429089.

3. Randles, A., Crotti Junior, A., O'Sullivan, D.: Towards a vocabulary for mapping quality assessment. To Appear Proc. 15th Int. Work. Ontol. Matching 19th Int. Semant. Web Conf. (ISWC), 2020. (2020).
4. Moreau, B., Serrano-Alvarado, P.: Assessing the Quality of RDF Mappings with EvaMap. In: 17th Extended Semantic Web Conference (ESWC2020) (2020).
5. 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: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 133–149. Springer Verlag (2015). https://doi.org/10.1007/978-3-319-25010-6_8.
6. Heyvaert, P., De Meester, B., Dimou, A., Verborgh, R.: Rule-driven inconsistency resolution for knowledge graph generation rules. Semant. Web. 10, (2019). https://doi.org/10.3233/SW-190358.
7. Mcglinn, D.K., Brennan, R., Debruyne, C., Meehan, A., McNerney, L., Clinton, E., Kelly, P., O'Sullivan, D.: Publishing authoritative geospatial data to support interlinking of building information models. Autom. Constr. 124, 103534 (2021). https://doi.org/10.1016/j.autcon.2020.103534.