Engaging Content
Engaging People

Assessing the Quality of R2RML Mappings

Ademar Crotti Junior, Jeremy Debattista and Declan O'Sullivan

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union
European Union
European Regional Development Fund
SFI Science Foundation Ireland

The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

Outline

Motivation

- Data quality
- Data mapping
- Motivating example

Our approach

Evaluation

Conclusions and future work

Data quality

Data quality is a complex multidimensional concept involving various aspects by which one can **characterize the quality** of a dataset for a particular task

Data quality problems, such as inaccuracy, incompleteness, and inconsistency, imply **limitations to the full exploitation of data**

In most cases, data quality frameworks assess the final datasets and not the artefacts used to produce them i.e. the **mappings**

Data mapping

Mappings relate source and target elements

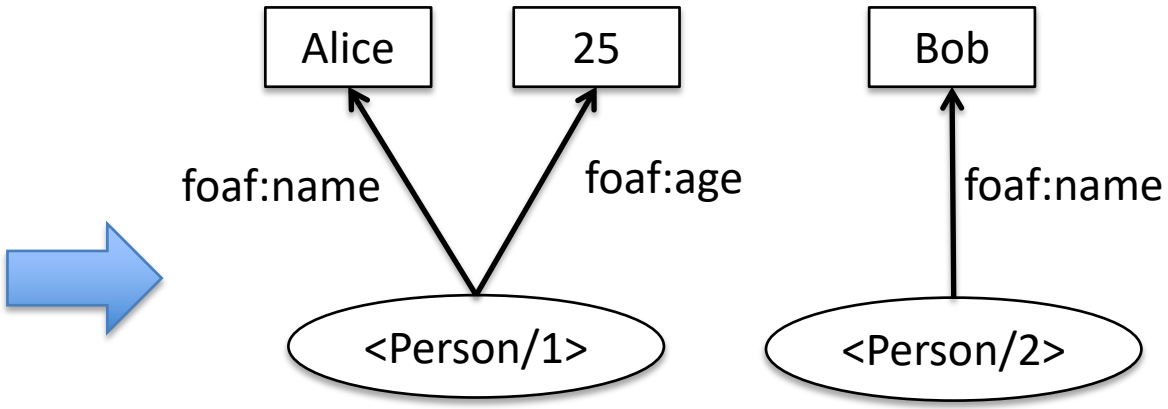
In the Semantic Web, mappings are commonly used to **declaratively** define the **transformations** needed to represent **non-RDF resources as RDF**

These mappings are defined using **mapping languages**

Data mapping

Person

ID	NAME	AGE
1	Alice	25
2	Bob	NULL




```
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
<TriplesMap1>
  rr:logicalTable [ rr:tableName "Person"];
  rr:subjectMap [
    rr:template "http://www.ex.com/Person/{ID}";
    rr:class foaf:Person
  ];
  rr:predicateObjectMap [
      rr:predicate foaf:name; rr:objectMap [rr:column "NAME" ]
  ];
  rr:predicateObjectMap [
      rr:predicate foaf:age; rr:objectMap [rr:column "AGE" ]
  ].
```

R2RML (RDB to RDF Mapping Language)

W3C-Recommended RDF-based mapping language to map relational databases into RDF

R2RML mappings are composed of triples maps with:

- One logical table
- One subject map
- Zero or more predicate object maps

```
<TriplesMap1>
  rr:logicalTable [ rr:tableName "Person"];
  rr:subjectMap [
    rr:template "http://www.ex.com/Person/{ID}";
    rr:class foaf:Person
  ];
  rr:predicateObjectMap [
      rr:predicate foaf:name; rr:objectMap [rr:column "NAME" ]
  ].
```

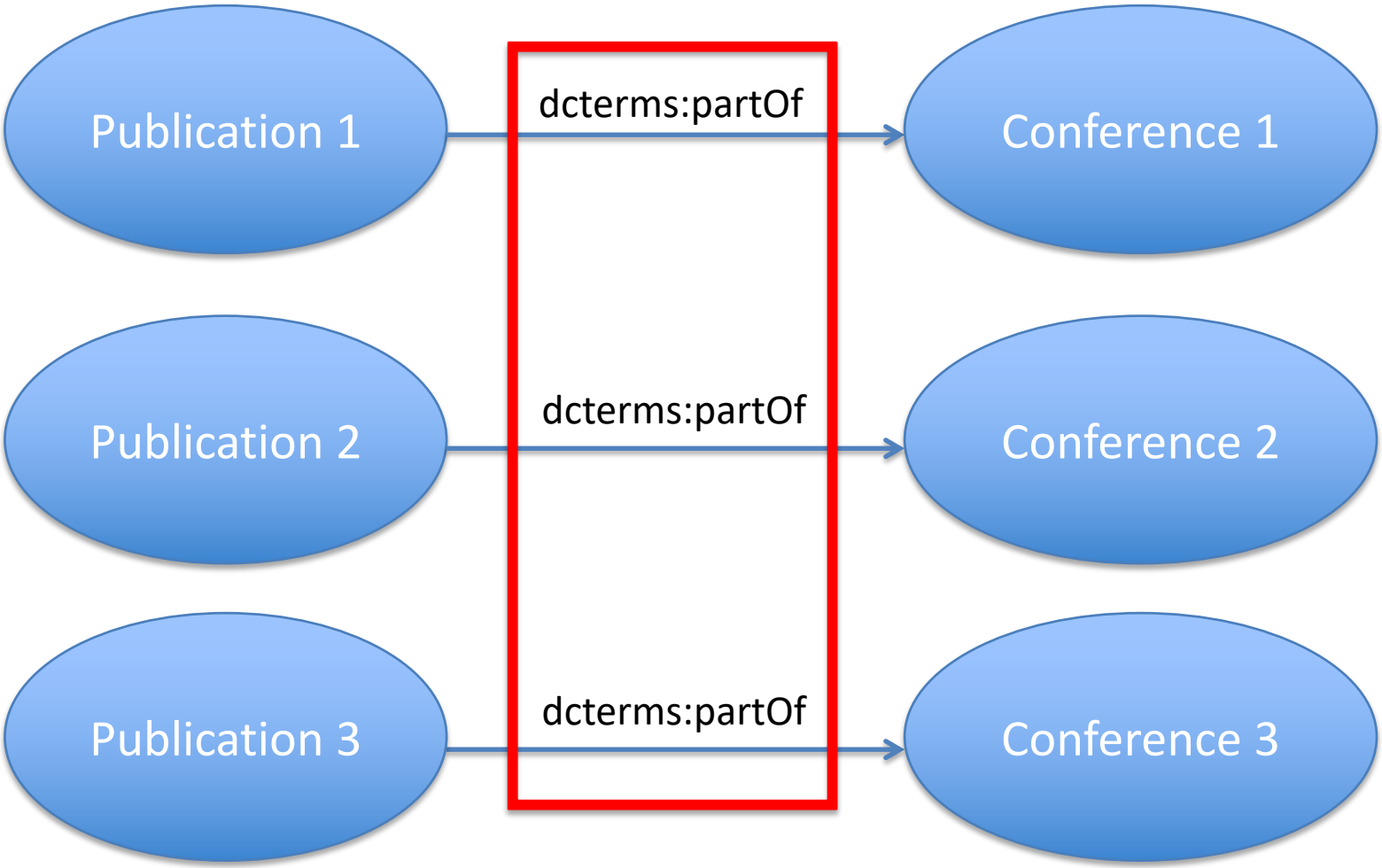
Motivating example - DBLP mapping


```
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix dcterms: <http://purl.org/dc/terms/> .

<TriplesMap1>
...
 rr:predicateObjectMap [
   rr:predicate dcterms:partOf ;
   rr:objectMap [
     rr:parentTriplesMap <#Publications> ;
        rr:joinCondition [
                rr:child "crossref" ;
                rr:parent "dblp_key" ;
        ];
     ];
 ] .
```

Undefined property!

Motivating example - DBLP mapping



Motivating example - DBLP mapping

Assessing entire datasets is **resource and time consuming**

Each **mapping violation** may become **exponentially** larger in the resulting dataset

If the **mappings** are not fixed (**root cause**), any **quality assessment** executed to the published datasets would be **overwritten** when mappings are reused – for new or updated data

Assessing the quality of R2RML mappings

Our approach proposes the use of quality assessment frameworks to **also** cover the mapping process

- Identify the origin in order to fix violations before dataset generation
- Avoids the propagation of violations
- Assist data providers into producing high quality datasets

The **earlier** data **quality issues** are **identified and fixed the better**

The proposed approach allows for the definition of **quality metrics to assess the mappings** used to generate datasets

Implementation

Extension to the **Luzzu Framework** in order to also assess **mappings**

Luzzu is a **scalable, extensible, and customizable** Linked Data **quality assessment framework**

This extension currently **supports R2RML** in which a data structure is exposed to third party implemented metrics

Luzzu also allows for metrics to generate **detailed quality reports** together with **metadata** on the execution of the metrics

Implementation – Mapping Quality Metrics

The following quality metrics, which are classified in the representational category have been implemented.

- Usage of undefined classes
- Usage of undefined properties
- Usage of blank nodes
- Usage of RDF reification

Limitations

Certain violations can only be identified after the dataset generation

These are caused because of the input data or the ontology and vocabularies being used

For example, a mapping defines a datatype which may cause an error in the final dataset depending on the input values (e.g. xsd:date)

Evaluation

The metrics implemented in our Luzzu extension were used to evaluate mappings from two real-world use cases

MusicBrainz. MusicBrainz is an open music encyclopedia containing information about artists, releases and recordings

DBLP. The Computer Science bibliography collects open bibliographic information from major computer science journals and proceedings

Results

Mapping Quality Metric	MusicBrainz	DBLP
Usage of undefined classes	66.6%	40%
Usage of undefined properties	82.6%	76.9%
Usage of blank nodes	100%	100%
Usage of RDF reification	100%	100%

Results

MusicBrainz

- All classes and properties for the Modular Unified Tagging Ontology, which is used in the mappings, were found to be undefined

DBLP

- All classes and properties for the ontology with URI `http://swrc.ontoware.org/ontology#`, which is used in the mappings, were found to be undefined
- The property `dcterms:partOf` was found to be undefined. The correct property is `dcterms:isPartOf`
- The property `dcterms:tableOfContent` was found to be undefined. The correct property is `dcterms:tableOfContents`

Example of quality report for DBLP mappings


```
@base <https://w3id.org/lodquator/resource/> .

# ... other prefixes ...

<ba4e8bf9-7e40-4e19-9b62-fb96fce429d2>
    a qpro:QualityProblem ;
    qpro:isDescribedBy dqm:UndefinedPropertiesMetric ;
    qpro:problemStructure qpro:ModelContainer ;
    qpro:problematicThing <469a3186-8d9f-48e3-9027-
8458d887dca8> .

<469a3186-8d9f-48e3-9027-8458d887dca8>
    qpro:exceptionDescription dqm-prob:UndefinedProperty ;
    ex:undefinedProperty dcterms:partOf ;
    ex:onMapping <../TriplesMapPublications> ; ... .
```

Conclusions

Several **quality assessment frameworks** have been proposed in literature, however, in most cases, these **remain independent of the mapping process**

The goal of our **proposed approach** is to **allow the assessment** of the **mappings** used to generate RDF datasets

Assessing mappings is expected to be **more cost efficient** due to the number of triples being assessed and time taken

We have **demonstrated** our approach by **extending Luzzu, implementing 4 metrics**, and **evaluating** it using **two real-world** sets of mappings

Future work

Support for other mapping languages such as RML

Implementation of other quality metrics to cover other dimensions and categories

Integration of the proposed approach to mapping editors such that this mapping assessment may be done at design time

The end

Thank you.