Knowledge Graph Construction with R2RML and RML: An ETL System-based Overview

Julián Arenas-Guerrero, Ana Iglesias-Molina, Jhon Toledo, Luis Pozo-Gilo, Daniel Doña, Oscar Corcho, David Chaves-Fraga
Ontology Engineering Group, Universidad Politécnica de Madrid, Spain

Mario Scrocca
Cefriel – Politecnico di Milano, Italy

julian.arenas.guerrero@upm.es
7/6/2021
Heterogeneous Data Sources

RDF Materialized Knowledge Graph

Knowledge graph construction with **declarative mapping rules**
Motivation

... but there are many engines available...

Which one fits best in my use case?
- Open source engines
- RML engines selected based on the implementation report
- r2rml4net excluded from R2RML engines as it only supports SQL Server

https://rml.io/implementation-report/
Qualitative Analysis

Data Formats
Relational DBMS
Input Data Sources
Functions
Chunk Processing
Named Graphs
Mapping Languages
Ontology Input
Output Formats
Triplestore Output

Input Data Sources

Data Errors

Functions

Chunk Processing

Named Graphs

Output Formats
We use an existing benchmark to assess:

- **Execution time**
- **Memory consumption**
- **Correctness** of the results
### Mapping Languages Conformance

#### R2RML

<table>
<thead>
<tr>
<th></th>
<th>PostgreSQL</th>
<th></th>
<th>MySQL</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>passed</td>
<td>failed</td>
<td>passed</td>
<td>failed</td>
</tr>
<tr>
<td>Morph-RDB</td>
<td>27</td>
<td>35</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>Ontop</td>
<td>59</td>
<td>3</td>
<td>45</td>
<td>17</td>
</tr>
<tr>
<td>DB2Triples</td>
<td>13</td>
<td>49</td>
<td>24</td>
<td>38</td>
</tr>
<tr>
<td>R2RML-F</td>
<td>13</td>
<td>49</td>
<td>24</td>
<td>38</td>
</tr>
</tbody>
</table>

#### RML

<table>
<thead>
<tr>
<th></th>
<th>CSV</th>
<th></th>
<th>JSON</th>
<th></th>
<th>XML</th>
<th></th>
<th>PostgreSQL</th>
<th></th>
<th>MySQL</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>passed</td>
<td>failed</td>
<td>passed</td>
<td>failed</td>
<td>passed</td>
<td>failed</td>
<td>passed</td>
<td>failed</td>
<td>passed</td>
<td>failed</td>
</tr>
<tr>
<td>RMLMapper</td>
<td>38</td>
<td>1</td>
<td>38</td>
<td>2</td>
<td>36</td>
<td>2</td>
<td>53</td>
<td>7</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>CARML</td>
<td>25</td>
<td>14</td>
<td>23</td>
<td>17</td>
<td>23</td>
<td>15</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RocketRML</td>
<td>29</td>
<td>10</td>
<td>30</td>
<td>10</td>
<td>29</td>
<td>9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SDM-RDFizer</td>
<td>24</td>
<td>15</td>
<td>23</td>
<td>17</td>
<td>21</td>
<td>17</td>
<td>11</td>
<td>49</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>RMLStreamer</td>
<td>39</td>
<td>0</td>
<td>40</td>
<td>0</td>
<td>38</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Chimera</td>
<td>36</td>
<td>3</td>
<td>37</td>
<td>3</td>
<td>35</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

[https://www.w3.org/TR/rdb2rdf-test-cases/](https://www.w3.org/TR/rdb2rdf-test-cases/)
[https://rml.io/test-cases/](https://rml.io/test-cases/)
Madrid Benchmark

Data formats:
- RDB
- CSV
- JSON
- XML
- CUSTOM

Data scaling factors:
- 1, 10, 100, 1000

24h timeout
128GB max memory

RMLStreamer excluded:
- No duplicate elimination

https://github.com/oeg-upm/kgc-eval/

Execution Times

GTFS₁₁

GTFS₁₀

GTFS₁₀₀

SDM-RDFizer

Ontop, R2RML-F, db2triples
## Correctness of the Results

<table>
<thead>
<tr>
<th></th>
<th>Ontop</th>
<th>Morph-RDB</th>
<th>db2triples</th>
<th>R2RML-F</th>
<th>SDM-RDFizer</th>
<th>RML Mapper</th>
<th>Chimera</th>
<th>Rocket RML</th>
<th>CARML</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GTFS-1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RDB</td>
<td>395953</td>
<td>454661</td>
<td>395953</td>
<td>395953</td>
<td>395953</td>
<td>397622</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CSV</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>JSON</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>397622</td>
<td>395953</td>
<td>395953</td>
<td>397622</td>
</tr>
<tr>
<td>XML</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>397622</td>
<td>395953</td>
<td>395953</td>
<td>395953</td>
</tr>
<tr>
<td>CUSTOM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>397622</td>
<td>395953</td>
<td>395953</td>
<td>395953</td>
</tr>
</tbody>
</table>

KGC with [R2]RML: An ETL System-based Overview
1. There are **few systems with high coverage of the features** considered in our qualitative analysis.

2. Several engines have a **medium-low conformance** w.r.t. the mapping languages specification.
   - Ontop, RMLMapper, RMLStreamer, Chimera

1. Most of the engines report performance and **scalability problems** for large input data sources.
   - SDM-RDFizer, Ontop
Knowledge Graph Construction with R2RML and RML: An ETL System-based Overview

Julián Arenas-Guerrero, Ana Iglesias-Molina, Jhon Toledo, Luis Pozo-Gilo, Daniel Doña, Oscar Corcho, David Chaves-Fraga
Ontology Engineering Group, Universidad Politécnica de Madrid, Spain

Mario Scrocca
Cefriel – Politecnico di Milano, Italy

julian.arenas.guerrero@upm.es
7/6/2021