

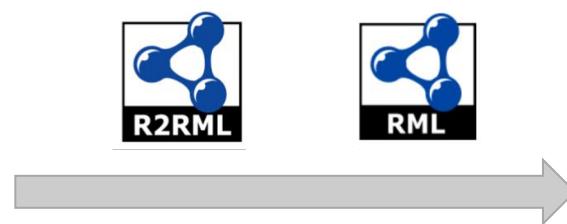


Knowledge Graph Construction with R2RML and RML: An ETL System-based Overview

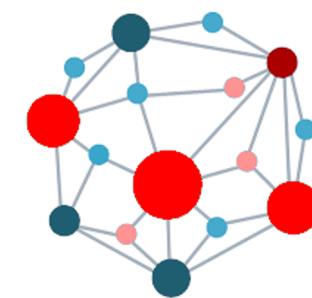
**Julián Arenas-Guerrero, Ana Iglesias-Molina,
Jhon Toledo, Luis Pozo-Gilo, Daniel Doña,
Oscar Corcho, David Chaves-Fraga**
Ontology Engineering Group,
Universidad Politécnica de Madrid, Spain

Mario Scrocca
Cefriel – Politecnico di Milano, Italy

Heterogeneous Data Sources



RDF Materialized Knowledge Graph



Knowledge graph construction with **declarative mapping rules**

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



Which one **fits** best in my **use case**?

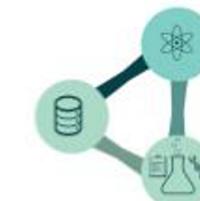
- Open source engines
- RML engines selected based on the **implementation report**
- r2rml4net excluded from R2RML engines as it only supports SQL Server



ontop morph



-F



SDM-RDFizer



Data Formats

Mapping Languages

Relational DBMS

Data Errors

Ontology Input

Input Data Sources

Functions

Chunk Processing

Output Formats

Named Graphs

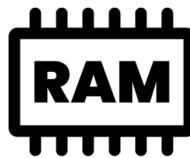
Triplestore Output

We use mapping languages test cases to asses the conformance of the engines.

We use an existing **benchmark** to assess:



Execution time



Memory consumption



Correctness of the results

Mapping Languages Conformance

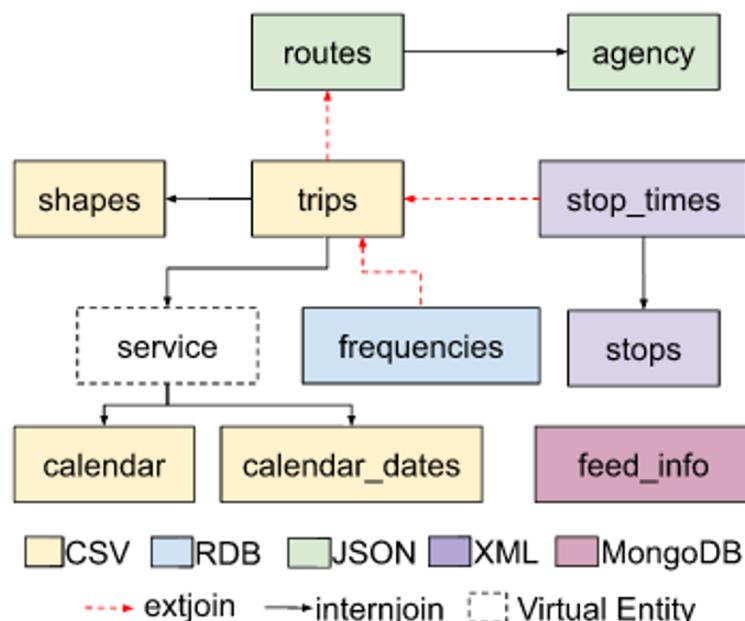


	PostgreSQL		MySQL	
	passed	failed	passed	failed
Morph-RDB	27	35	31	31
Ontop	59	3	45	17
DB2Triples	13	49	24	38
R2RML-F	13	49	24	38



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

<https://www.w3.org/TR/rdb2rdf-test-cases/>
<https://rml.io/test-cases/>



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

Chaves-Fraga, David, et al. "GTFS-Madrid-Bench: A benchmark for virtual knowledge graph access in the transport domain." *Journal of Web Semantics* 65 (2020): 100596. <https://doi.org/10.1016/j.websem.2020.100596>.

Data formats:

RDB
CSV
JSON
XML
CUSTOM

Data scaling factors:

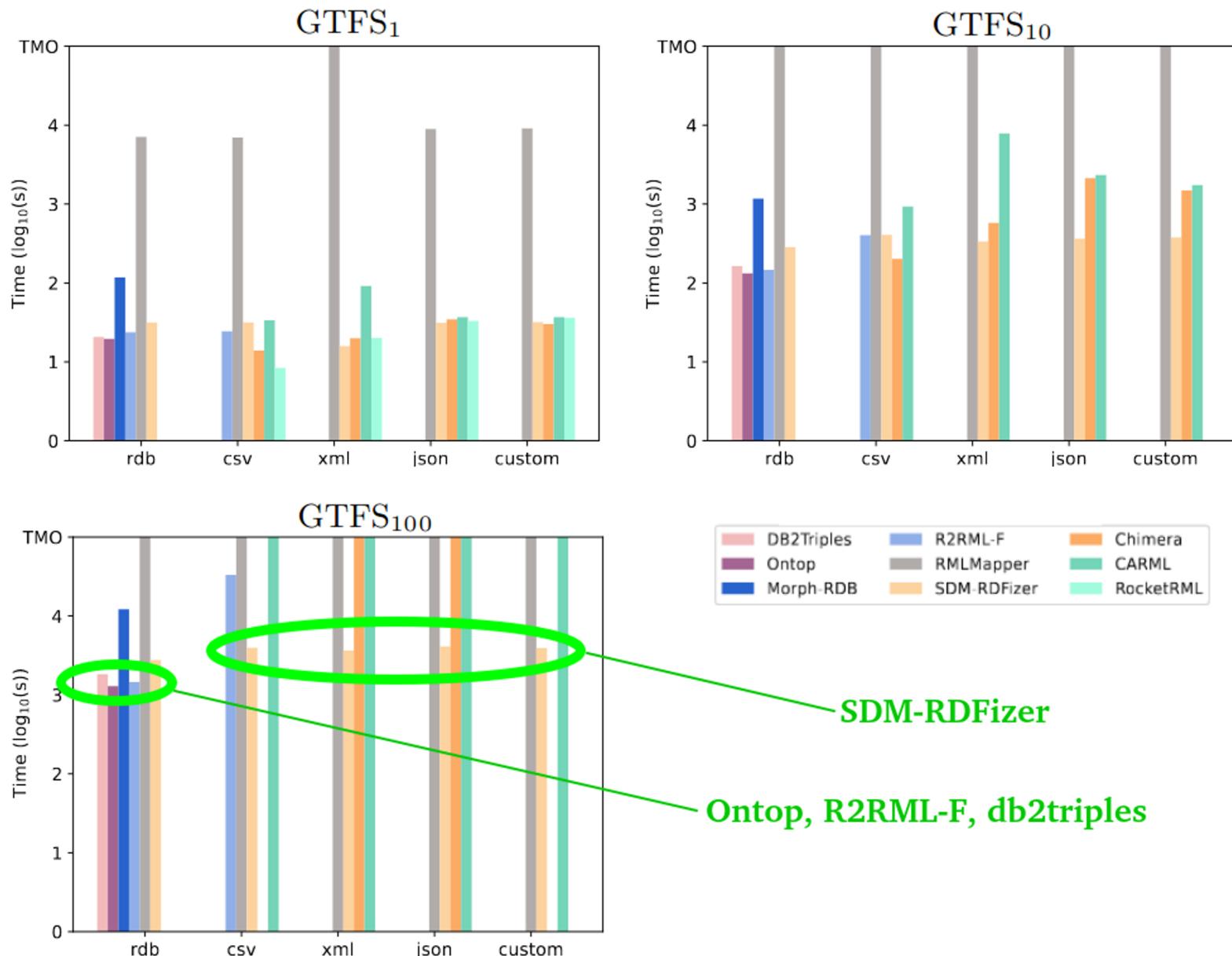
1, 10, 100, 1000

24h timeout

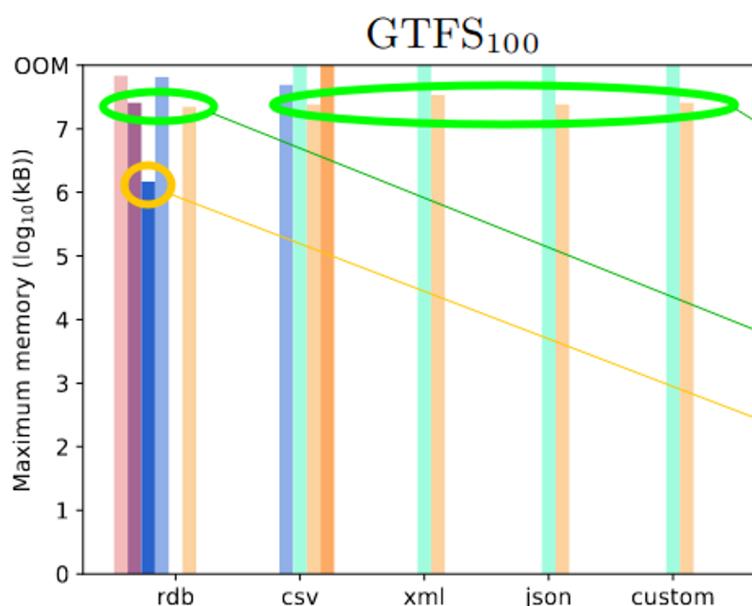
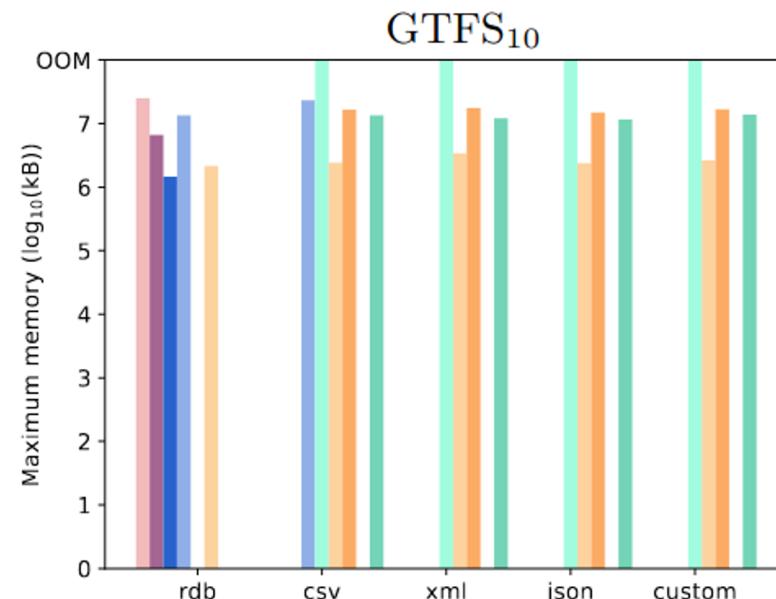
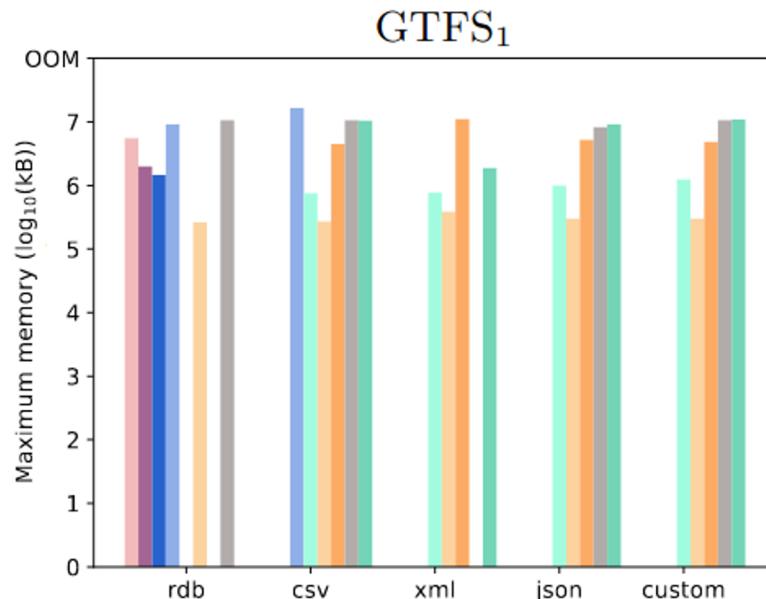
128GB max memory

RMLStreamer excluded:
No duplicate
elimination

Execution Times



Memory Consumption



Legend:

- DB2Triples
- Ontop
- Morph-RDB
- R2RML-F
- RocketRML
- SDM-RDFizer
- RMLMapper
- Chimera
- CARML

SDM-RDFizer

Ontop, SDM-RDFizer

Morph-RDB

Correctness of the Results

	Ontop	Morph-RDB	db2triples	R2RML-F	SDM-RDFizer	RML-Maper	Chimera	Rocket-RML	Rocket-CARML
GTFS-1									
RDB	395953	454661	395953	395953	395953	397622	-	-	-
CSV	-	-	-	395953	395953	397622	395953	395953	395953
JSON	-	-	-	-	395953	397622	395953	397622	397622
XML	-	-	-	-	395953	397622	395953	395953	395953
CUSTOM	-	-	-	-	395953	397622	395953	395953	395953

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