Knowledge Graph Construction

David Chaves-Fraga, Ontology Engineering Group
Universidad Politécnica de Madrid, Spain
Data Integration Team

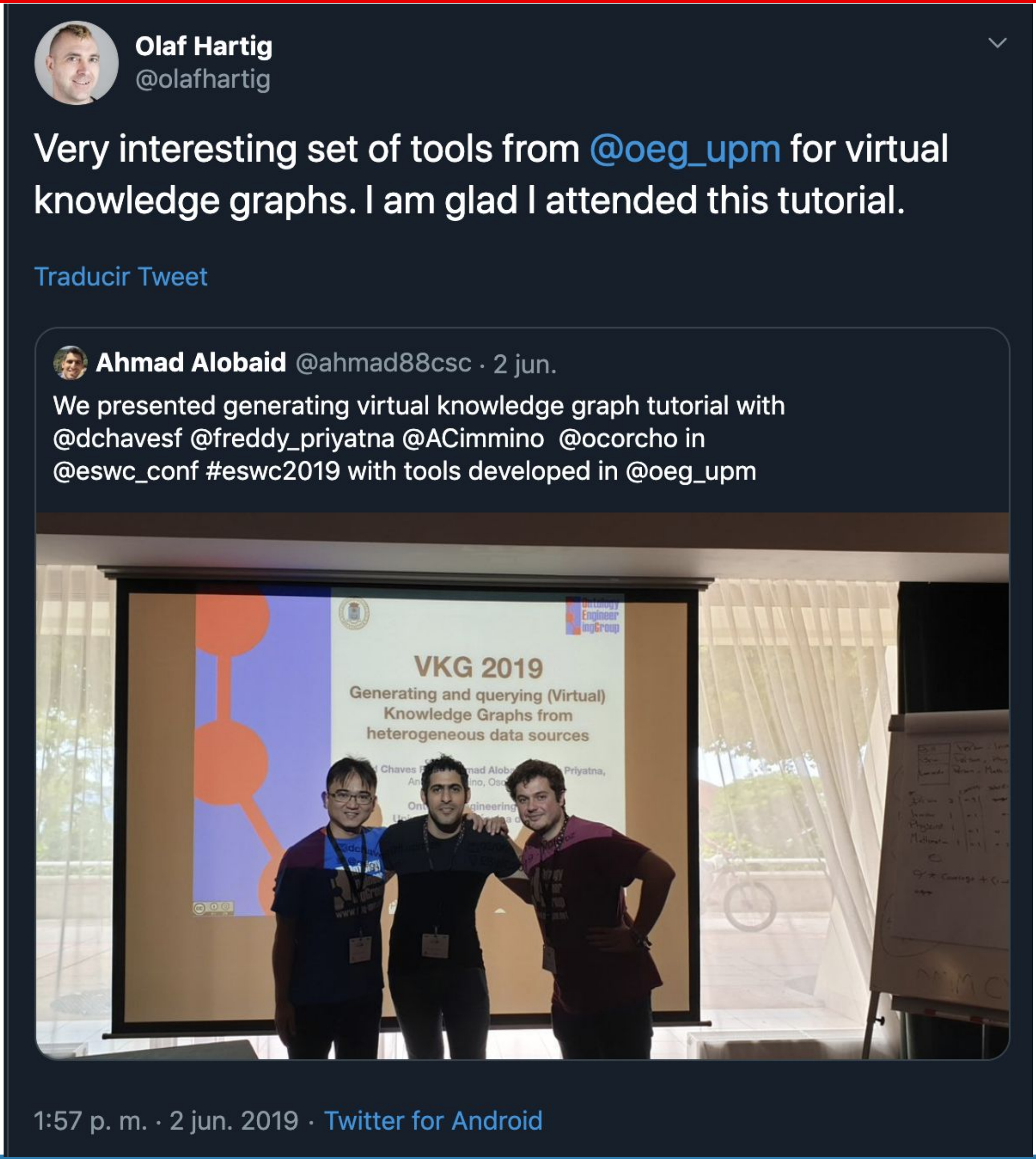
dchaves@fi.upm.es
@dchavesf

12/09/2019
OEG

Last talk 1 year ago...

https://dchaves.oeg-upm.net/

VKG2019 Tutorial



Olaf Hartig
@olafhartig

Very interesting set of tools from @oeg_upm for virtual knowledge graphs. I am glad I attended this tutorial.

Traducir Tweet

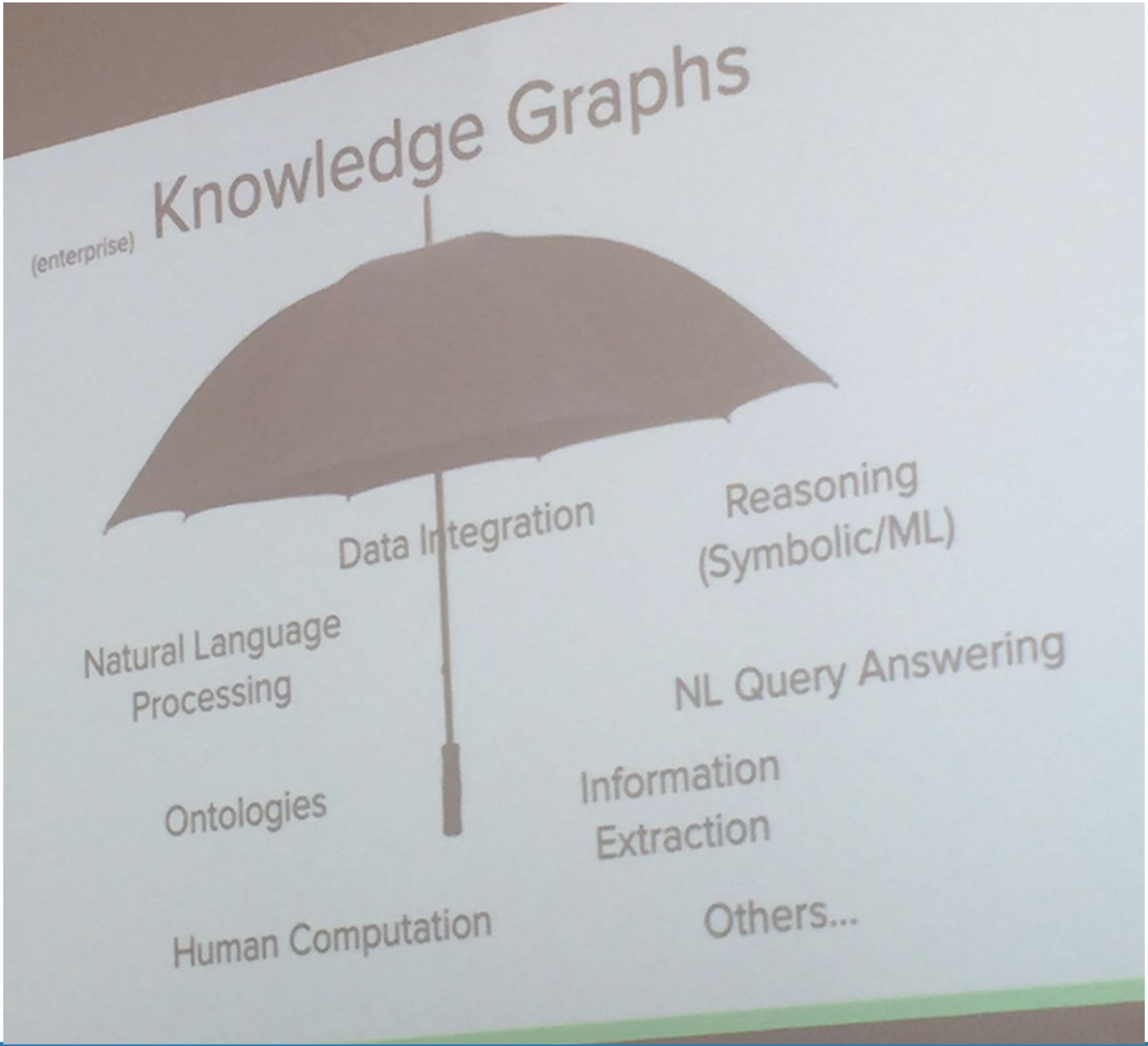
Ahmad Alobaid @ahmad88csc · 2 jun.

We presented generating virtual knowledge graph tutorial with @dchavesf @freddy_priyatna @ACimmino @ocorcho in @eswc_conf #eswc2019 with tools developed in @oeg_upm

VKG 2019
Generating and querying (Virtual) Knowledge Graphs from heterogeneous data sources
Ontology Engineering Group

1:57 p. m. · 2 jun. 2019 · Twitter for Android

KGB Workshop



(enterprise) Knowledge Graphs

Data Integration

Reasoning (Symbolic/ML)

Natural Language Processing

NL Query Answering

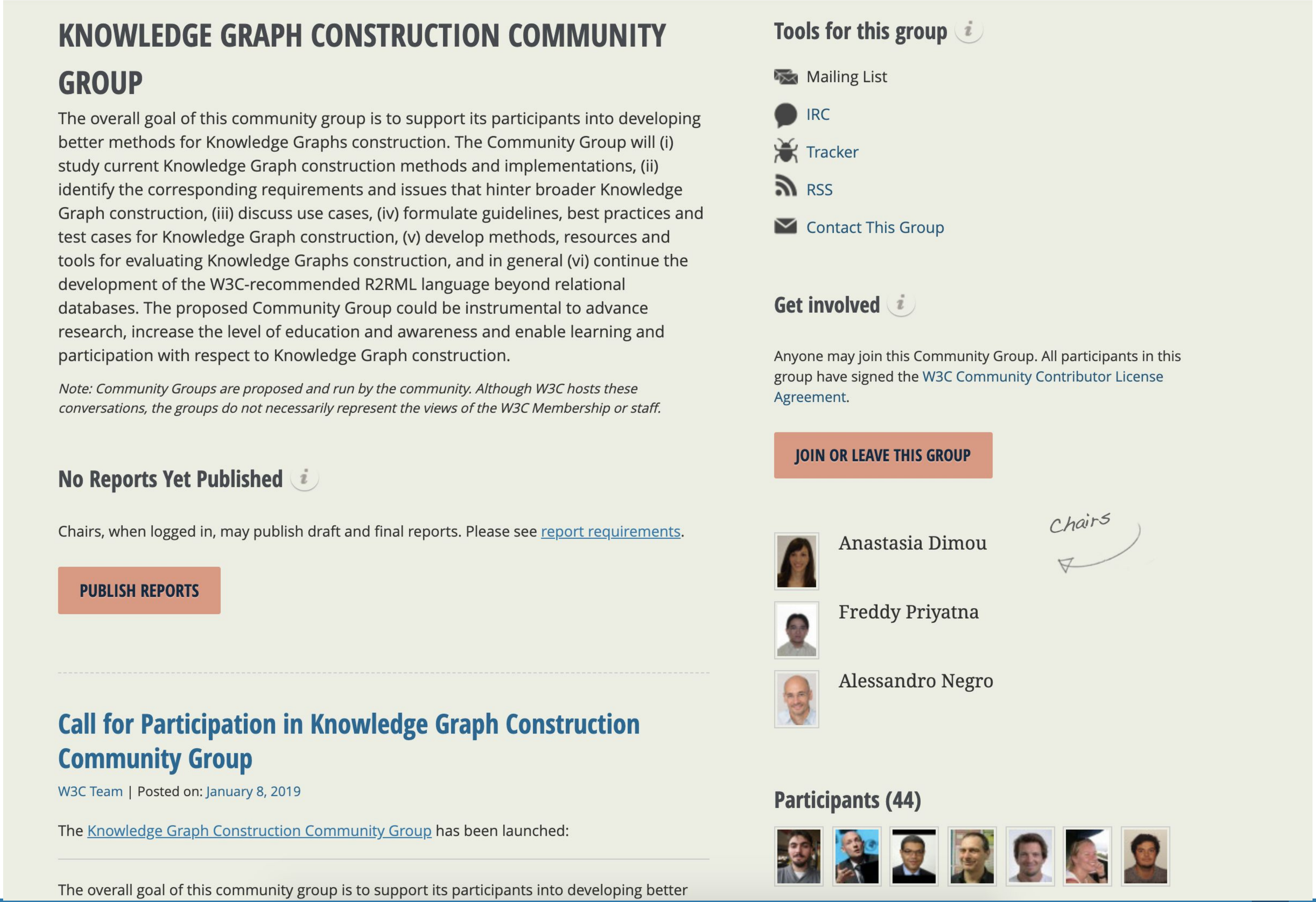
Ontologies

Information Extraction

Human Computation

Others...

W3C Knowledge Graph Construction CG



KNOWLEDGE GRAPH CONSTRUCTION COMMUNITY GROUP

The overall goal of this community group is to support its participants into developing better methods for Knowledge Graphs construction. The Community Group will (i) study current Knowledge Graph construction methods and implementations, (ii) identify the corresponding requirements and issues that hinter broader Knowledge Graph construction, (iii) discuss use cases, (iv) formulate guidelines, best practices and test cases for Knowledge Graph construction, (v) develop methods, resources and tools for evaluating Knowledge Graphs construction, and in general (vi) continue the development of the W3C-recommended R2RML language beyond relational databases. The proposed Community Group could be instrumental to advance research, increase the level of education and awareness and enable learning and participation with respect to Knowledge Graph construction.

Note: Community Groups are proposed and run by the community. Although W3C hosts these conversations, the groups do not necessarily represent the views of the W3C Membership or staff.

No Reports Yet Published

Chairs, when logged in, may publish draft and final reports. Please see report requirements.

PUBLISH REPORTS

Call for Participation in Knowledge Graph Construction Community Group

W3C Team | Posted on: January 8, 2019

The Knowledge Graph Construction Community Group has been launched:

The overall goal of this community group is to support its participants into developing better

Tools for this group

- Mailing List
- IRC
- Tracker
- RSS
- Contact This Group

Get involved

Anyone may join this Community Group. All participants in this group have signed the W3C Community Contributor License Agreement.

JOIN OR LEAVE THIS GROUP

Chairs

Anastasia Dimou

Freddy Priyatna

Alessandro Negro

Participants (44)

Contributions (2018-2019)

Tools/Engines (Optimizing KG Construction):

- Morph Suite: **Morph-CSV**, Morph-GraphQL, Morph-RDB
- **SDM-RDFizer**

Mapping Languages:

- **Mapping Translation**
- **RML Test Cases**

Evaluations:

- **What are the parameters that affect a KG Construction?**
- GTFS-Bench: The [OBDI/VKG] benchmark
- Evaluation of KG Construction Engines

Morph Suite

https://morph.oeg-upm.net/

Morph-CSV

Enhancing OBDA query translation* over Open Tabular Data

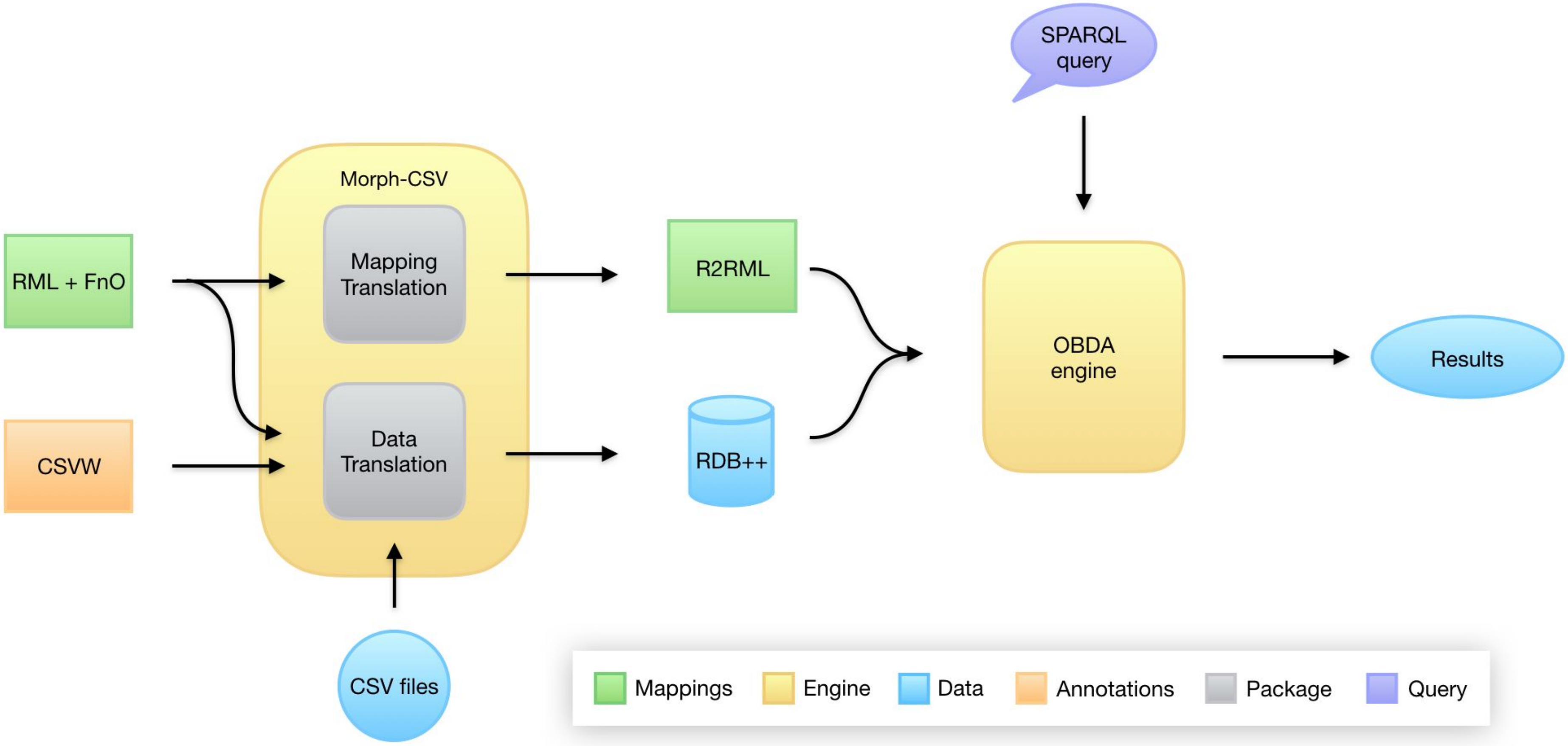
Exploit mapping/annotations to improve query completeness and query performance:

- Generate corresponding SQL schema from input files
- Apply transformation and cleaning functions
- Identify indexes
- Translate input mappings to R2RML
- Morph-CSV can be embedded in the top of any R2RML-compliant engine

Paper: Written and submitted 2 times → next ESWC/SWJ

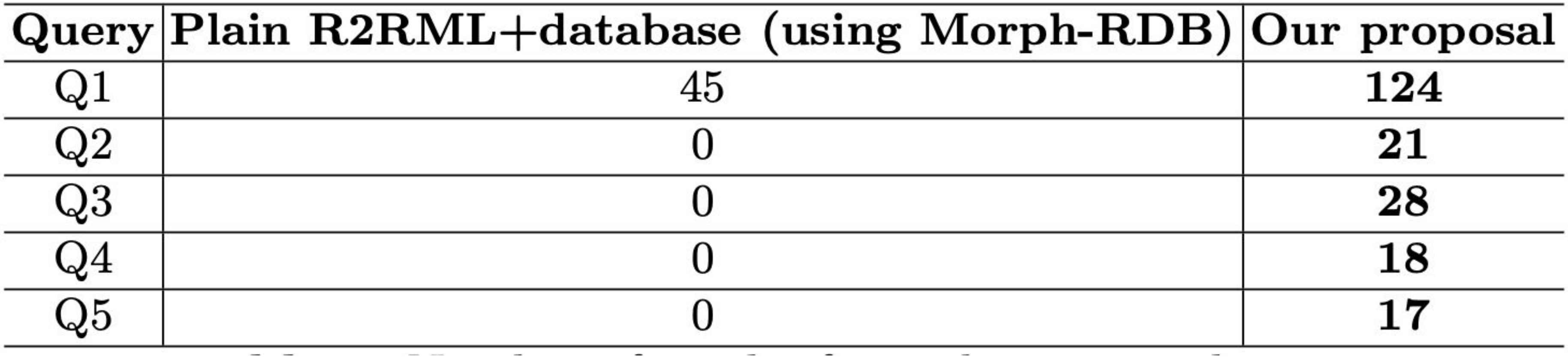
* OBDA query translation = Virtual Knowledge Graph Creation

Morph-CSV


SPARQL query
Morph-CSV
Mapping Translation
Data Translation
RML + FnO
CSVW
R2RML
RDB++
OBDA engine
Results
CSV files
🟩 Mappings
🟨 Engine
🟦 Data
🟧 Annotations
⬜ Package
🟪 Query

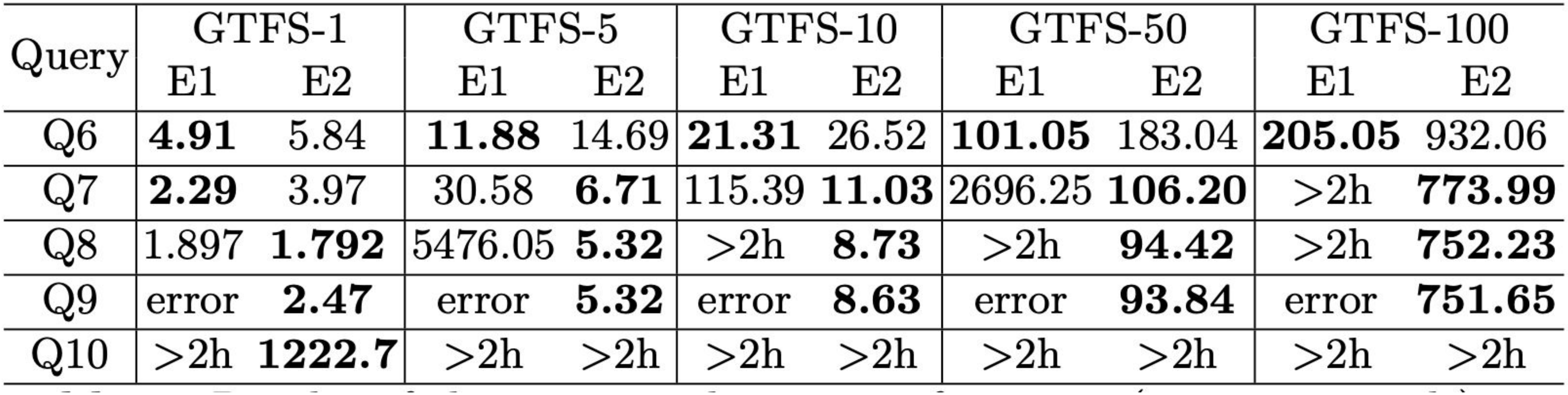
Morph-CSV results

Completeness



Query	Plain R2RML+database (using Morph-RDB)	Our proposal
Q1	45	**124**
Q2	0	**21**
Q3	0	**28**
Q4	0	**18**
Q5	0	**17**

Performance



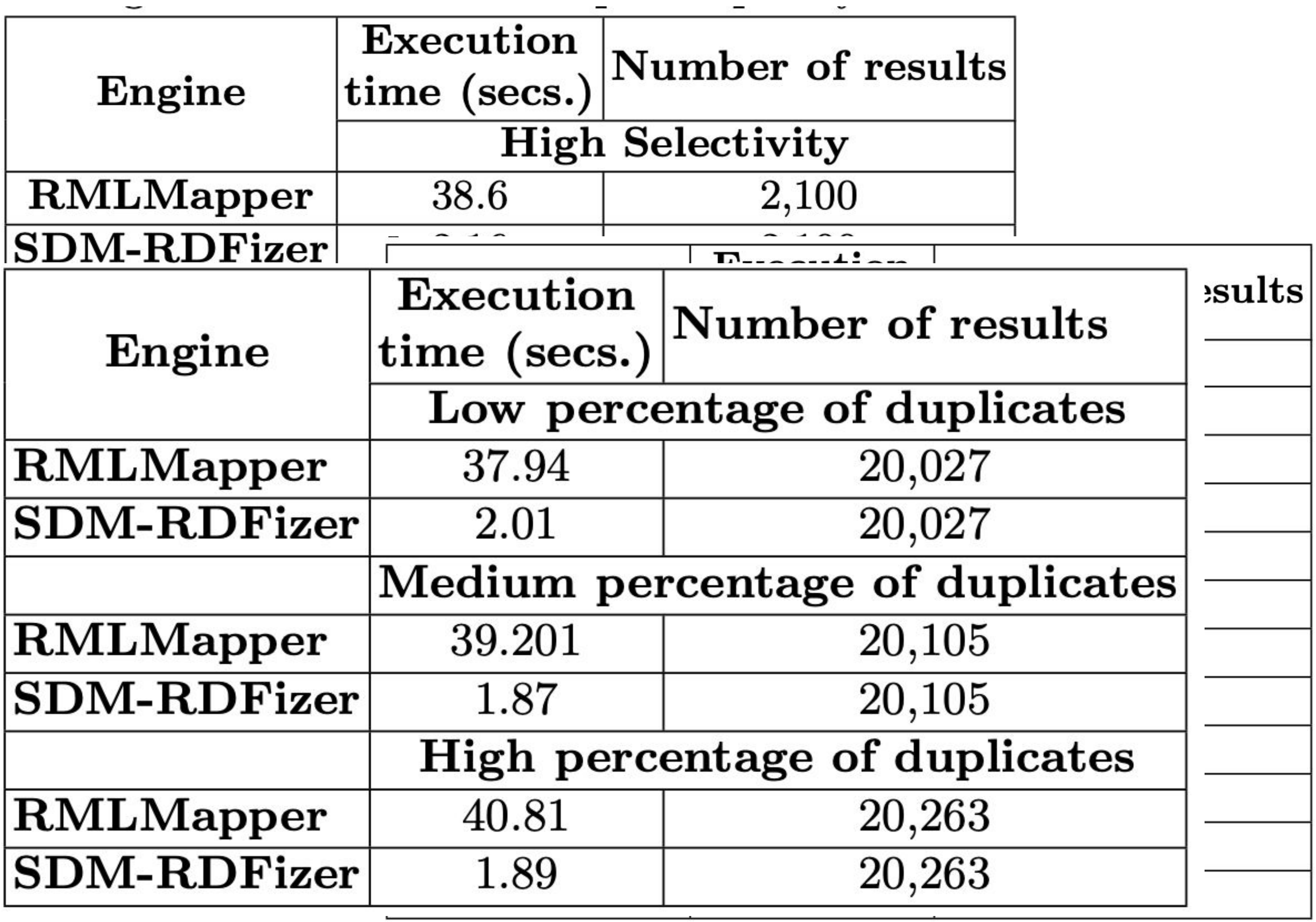
Query	GTFS-1		GTFS-5		GTFS-10		GTFS-50		GTFS-100	
	E1	E2	E1	E2	E1	E2	E1	E2	E1	E2
Q6	**4.91**	5.84	**11.88**	14.69	**21.31**	26.52	**101.05**	183.04	**205.05**	932.06
Q7	**2.29**	3.97	30.58	**6.71**	115.39	**11.03**	2696.25	**106.20**	>2h	**773.99**
Q8	1.897	**1.792**	5476.05	**5.32**	>2h	**8.73**	>2h	**94.42**	>2h	**752.23**
Q9	error	**2.47**	error	**5.32**	error	**8.63**	error	**93.84**	error	**751.65**
Q10	>2h	**1222.7**	>2h	>2h	>2h	>2h	>2h	>2h	>2h	>2h

SDM-RDFizer: an interpreter **of mapping rules** that allow the transformation of (**un)structured data** into **RDF knowledge graphs**

- Mapping rules defined in RML (100% RML compliant)
- Data structures and relation alegra operators for:
  - Efficient join execution (column1.table1=column2.table2)
  - Efficient management of duplicates
- Continuous behaviour
- Paper: TBD

https://github.com/SDM-TIB/SDM-RDFizer

SDM-RDFizer results



Engine	Execution time (secs.)	Number of results
	High Selectivity	
RMLMapper	38.6	2,100
SDM-RDFizer		

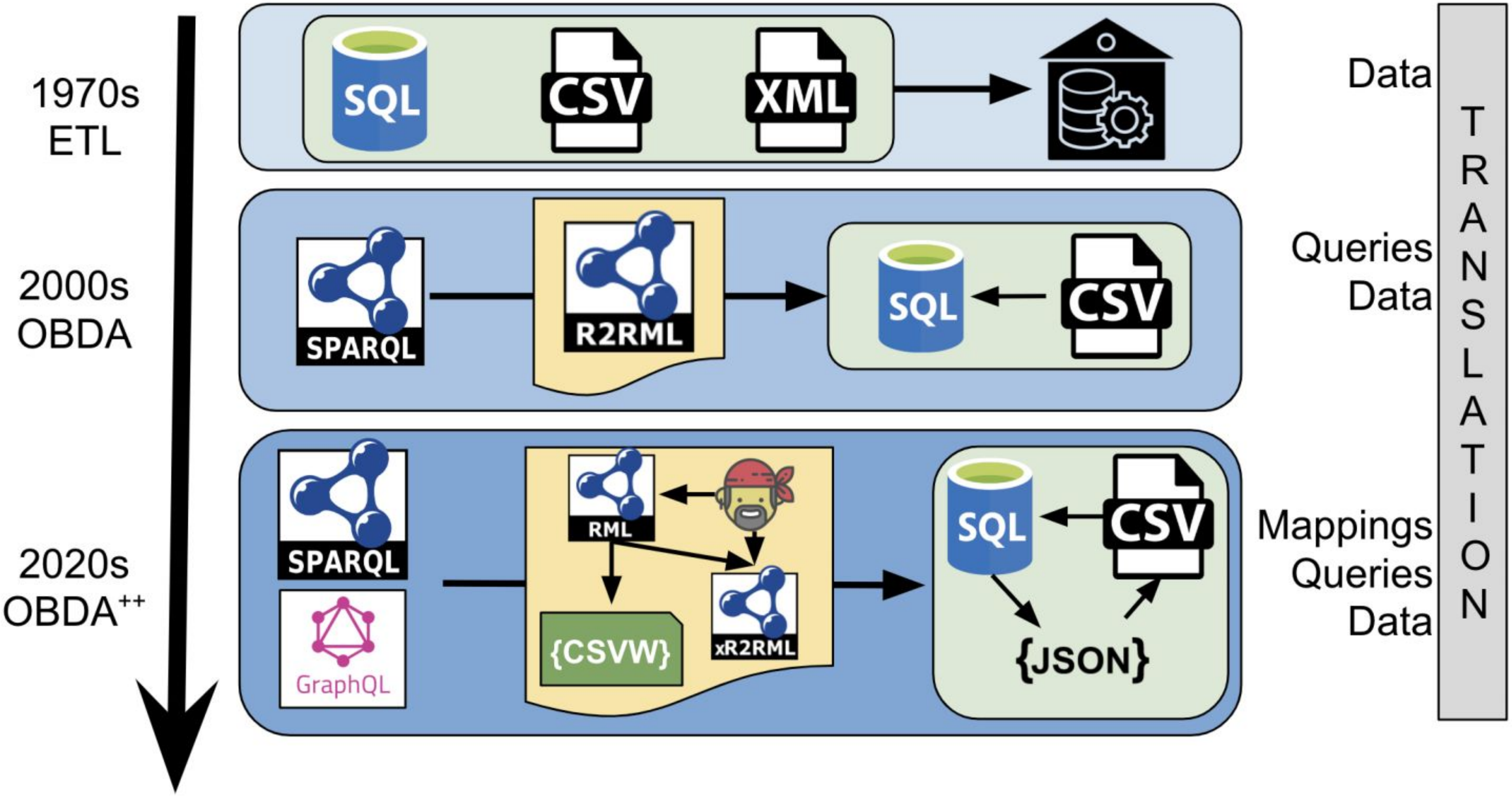
Engine	Execution time (secs.)	Number of results
	Low percentage of duplicates	
RMLMapper	37.94	20,027
SDM-RDFizer	2.01	20,027
	Medium percentage of duplicates	
RMLMapper	39.201	20,105
SDM-RDFizer	1.87	20,105
	High percentage of duplicates	
RMLMapper	40.81	20,263
SDM-RDFizer	1.89	20,263

Mapping Translation Concept

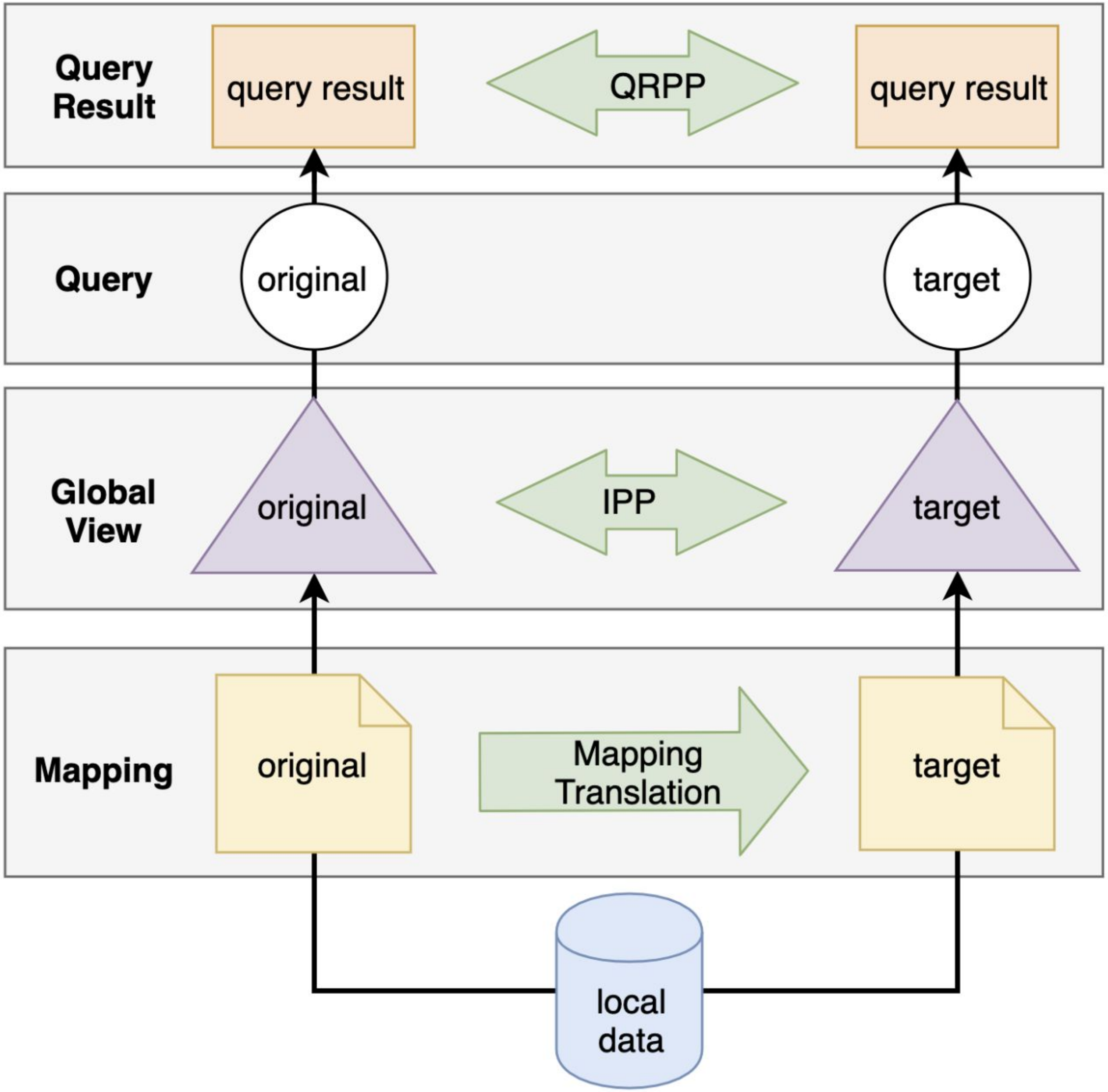
- Multiple use cases on KG Generation from Heterogeneous data sources (not same as RDB)
- Emergence of ad-hoc mapping languages to solve ad-hoc problems
- 1 mapping language → 1 tool

Corcho, O., Priyatna, F., Chaves-Fraga, D.: **Towards a New Generation of Ontology Based Data Access.** In: Semantic Web Journal (2019)

Mapping Translation Concept



Mapping Translation Concept



Use cases

- Maintainability: YARRRML, RMLC-Iterator
- Declarative2Programmed: Morph-GraphQL
- Enhance access to Tabular Data: Morph-CSV
- Understanding the semantics of mappings:
    - R2RML and Direct Mappings
    - OBDA Mappings from Ontop

RML Test Cases

- Emergence of KG engines parsing RML mappings
- R2RML Test cases for RDB
- How to test the conformance of the engines?

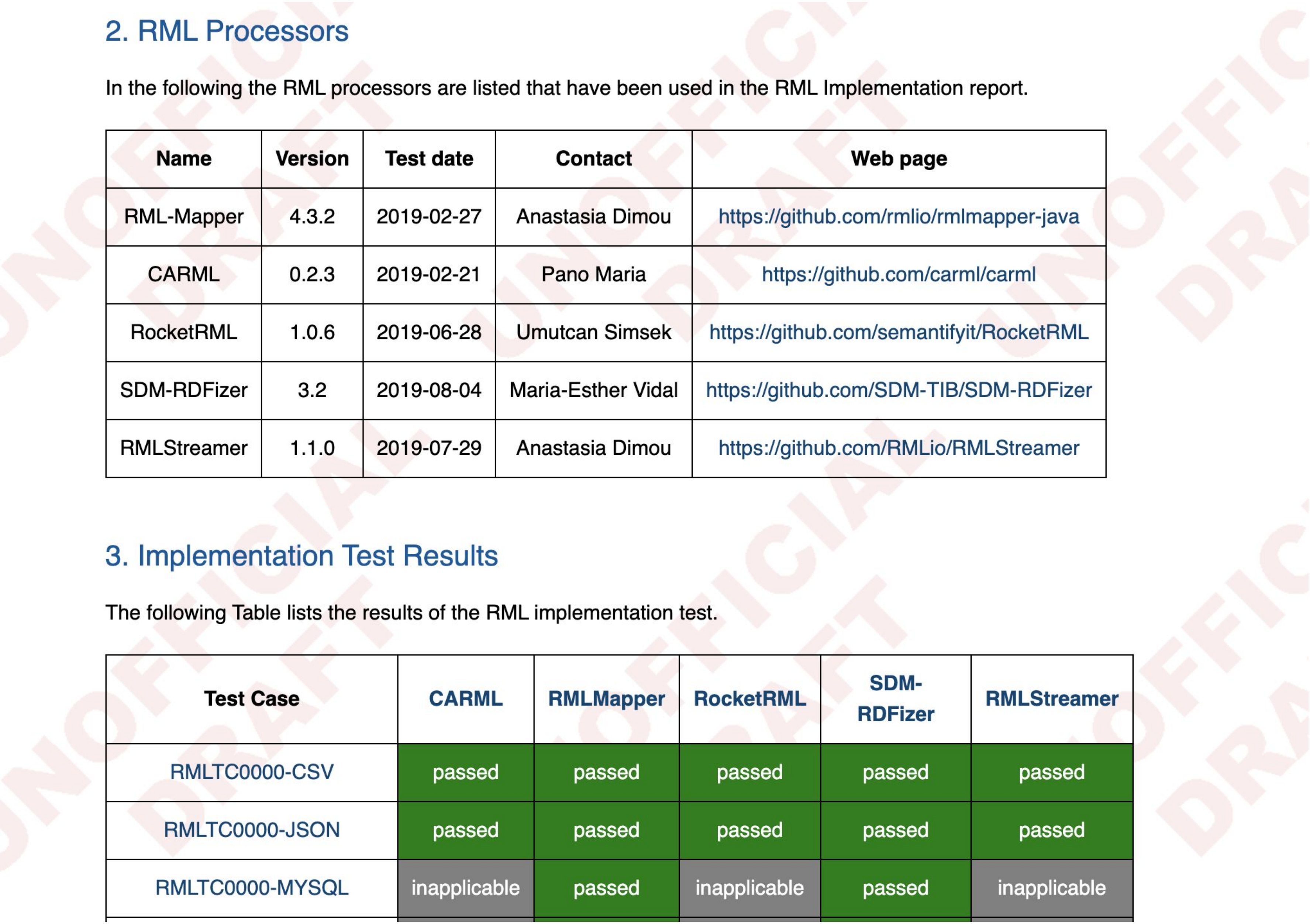
Heyvaert, P., Chaves-Fraga, D., Priyatna, F., Corcho, O., Mannens, E., Verborgh, R., Dimou, A.: **Conformance Test Cases for the RDF Mapping Language (RML)**. In: 1st Iberoamerican Knowledge Graphs and Semantic Web Conference.

RML Test Cases

297 Test Cases covering:

- CSV, JSON, XML, MySQL, PostegreSQL, SQLServer
- Translated from R2RML Test Cases
- Semantically described by:
  - Evaluation and Report Language (EARL) 1.0 Schema
  - Test case manifest vocabulary
  - Test Metadata vocabulary
  - Data Catalog vocabulary
- Each Test Case includes:
  - Data + Mapping + Expected Output KG
- Website: http://rml.io/test-cases/
- OnGoing work: including FnO Test Cases

RML Implementation report



2. RML Processors

In the following the RML processors are listed that have been used in the RML Implementation report.

Name	Version	Test date	Contact	Web page
RML-Mapper	4.3.2	2019-02-27	Anastasia Dimou	https://github.com/rmlio/rmlmapper-java
CARML	0.2.3	2019-02-21	Pano Maria	https://github.com/carml/carml
RocketRML	1.0.6	2019-06-28	Umutcan Simsek	https://github.com/semantifyit/RocketRML
SDM-RDFizer	3.2	2019-08-04	Maria-Esther Vidal	https://github.com/SDM-TIB/SDM-RDFizer
RMLStreamer	1.1.0	2019-07-29	Anastasia Dimou	https://github.com/RMLio/RMLStreamer

3. Implementation Test Results

The following Table lists the results of the RML implementation test.

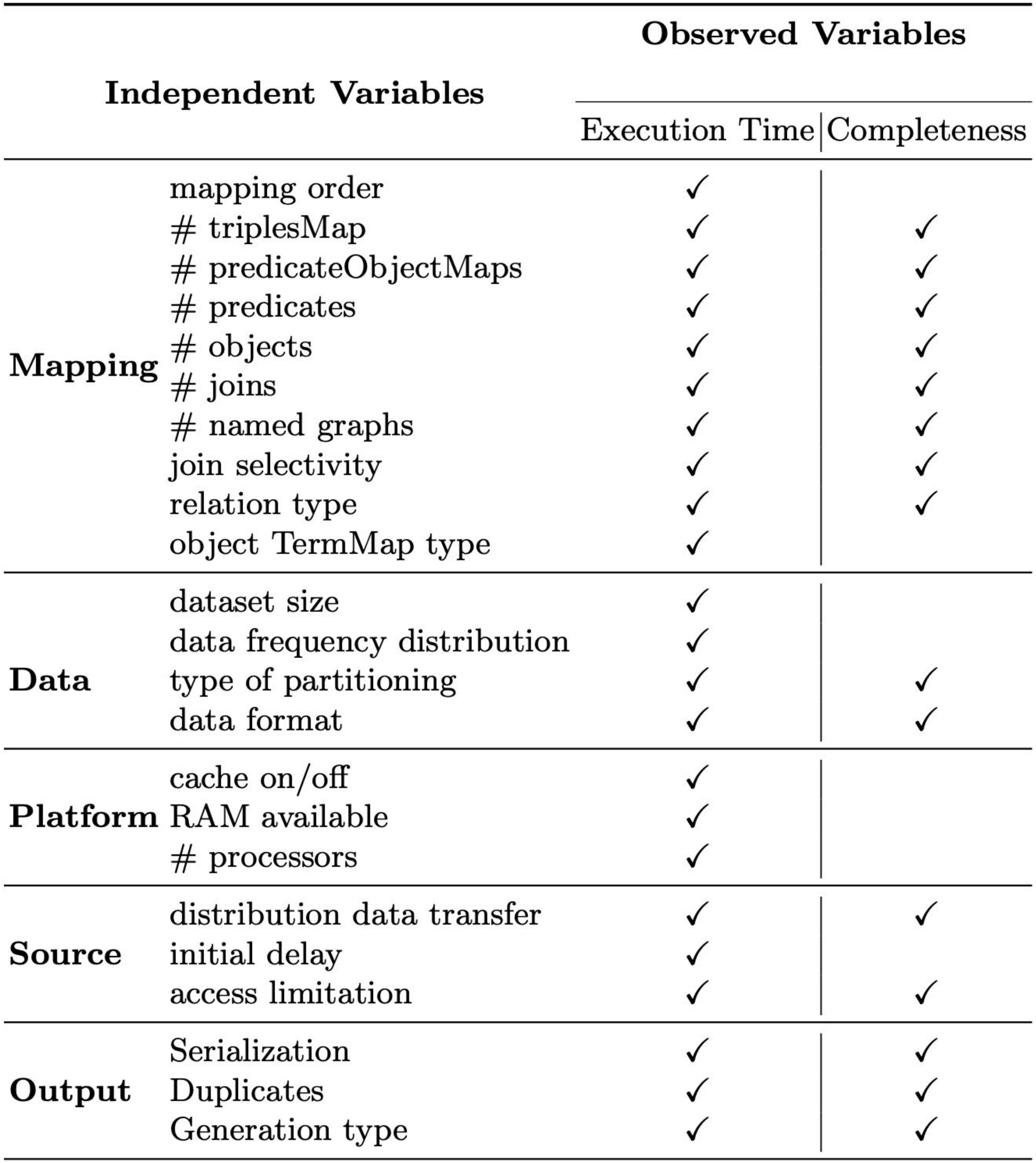
Test Case	CARML	RMLMapper	RocketRML	SDM-RDFizer	RMLStreamer
RMLTC0000-CSV	passed	passed	passed	passed	passed
RMLTC0000-JSON	passed	passed	passed	passed	passed
RMLTC0000-MYSQL	inapplicable	passed	inapplicable	passed	inapplicable

Parameters affecting KGC

- Emergence of tools that process mapping rules for knowledge graph creation
- No standard benchmark to test their performance and completeness
- Multiple variables involved in the process
- Evaluations focused on data size

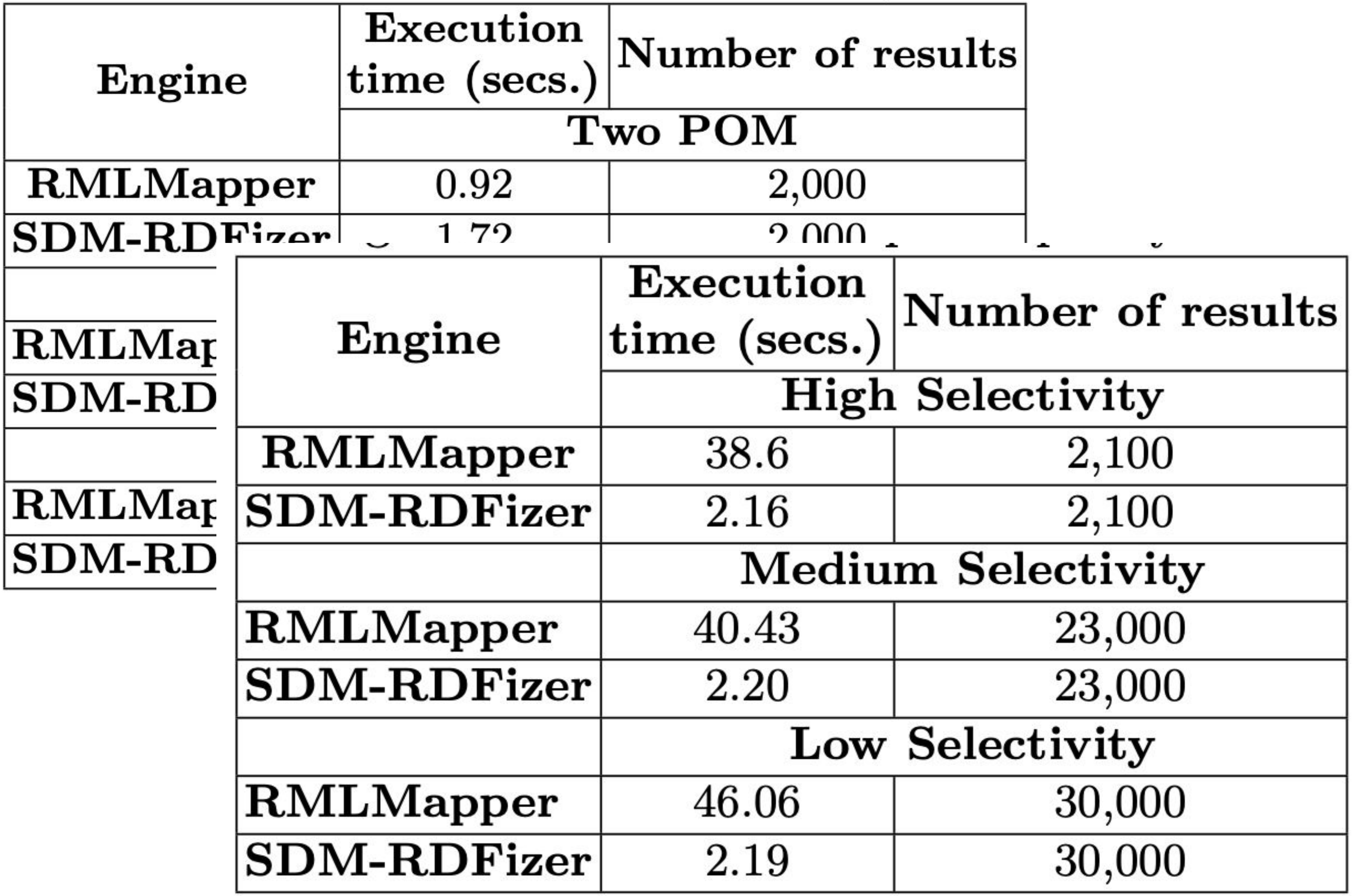
David Chaves-Fraga, Kemele M. Endris, Enrique Iglesias, Oscar Corcho, and Maria-Esther Vidal. **What are the Parameters that Affect the Construction of a Knowledge Graph?**. Accepted at the 18th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2019).

Parameters affecting KGC



Independent Variables		Observed Variables	
		Execution Time	Completeness
Mapping	mapping order	✓	
	# triplesMap	✓	✓
	# predicateObjectMaps	✓	✓
	# predicates	✓	✓
	# objects	✓	✓
	# joins	✓	✓
	# named graphs	✓	✓
	join selectivity	✓	✓
	relation type	✓	✓
	object TermMap type	✓	
Data	dataset size	✓	
	data frequency distribution	✓	
	type of partitioning	✓	✓
	data format	✓	✓
Platform	cache on/off	✓	
	RAM available	✓	
	# processors	✓	
Source	distribution data transfer	✓	✓
	initial delay	✓	
	access limitation	✓	✓
Output	Serialization	✓	✓
	Duplicates	✓	✓
	Generation type	✓	✓

Parameters affecting KGC



Engine	Execution time (secs.)	Number of results
	Two POM	
RMLMapper	0.92	2,000
SDM-RDFizer	1.72	2,000
RMLMap		
SDM-RD		
RMLMap		
SDM-RD		

Engine	Execution time (secs.)	Number of results
	High Selectivity	
RMLMapper	38.6	2,100
SDM-RDFizer	2.16	2,100
	Medium Selectivity	
RMLMapper	40.43	23,000
SDM-RDFizer	2.20	23,000
	Low Selectivity	
RMLMapper	46.06	30,000
SDM-RDFizer	2.19	30,000

Parameters affecting KGC

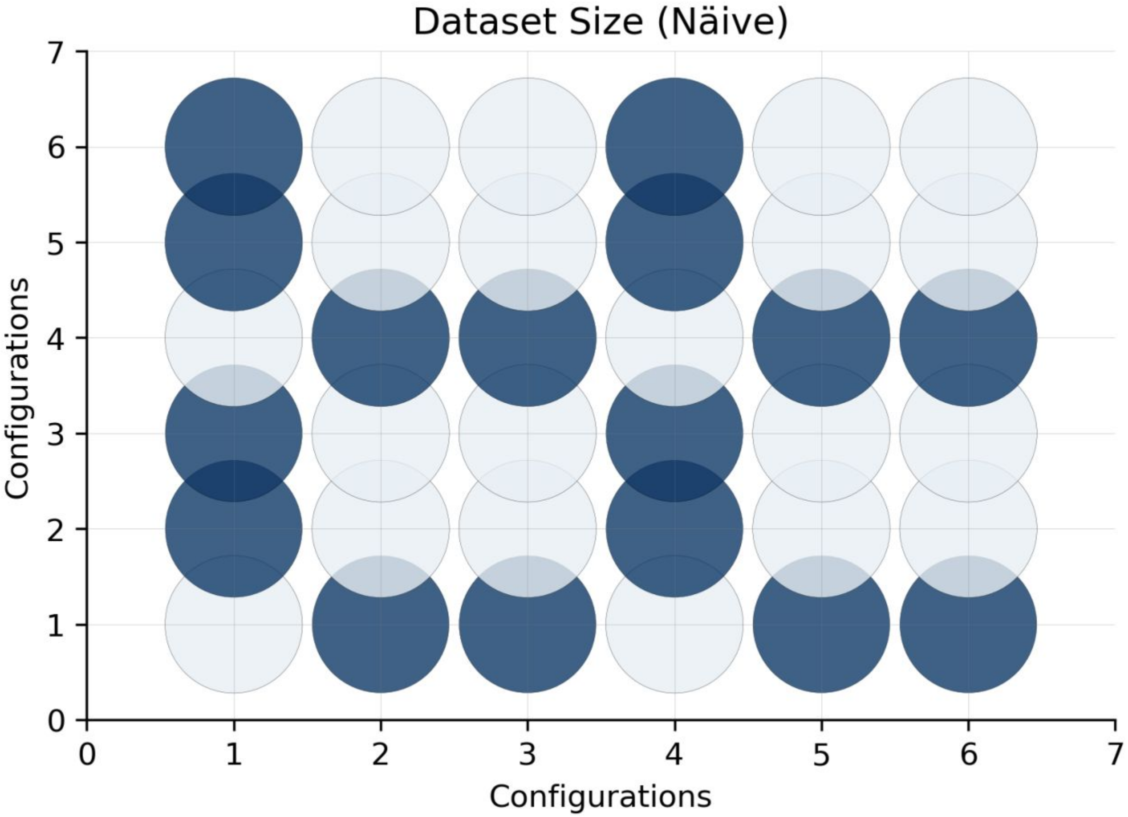


Fig. 2: Comparison of Knowledge Graph Creation Tool on Different Dataset Sizes (Naïve). The first three configurations, i.e. 1, 2, and 3 in x-axis and y-axis, correspond to SDM-RDFizer on datasets 1K, 10K, and 50K, respectively. The last three configurations, i.e. 4, 5, and 6 on x-axis and y-axis, correspond to RMLMapper 1K, 10K, and 50K, respectively. Grey bubbles correspond to correlation value of 1.0; blue bubbles show a positive correlation. The number of blue bubbles suggests that both systems exhibit similar behaviour.

Parameters affecting KGC

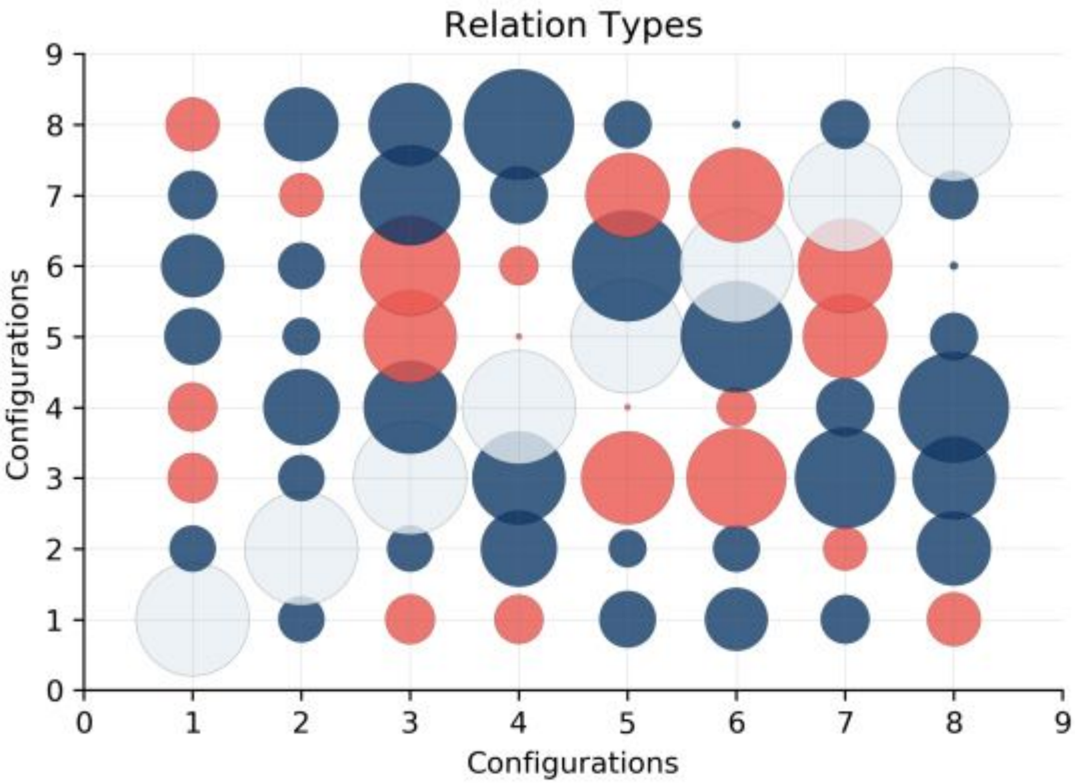


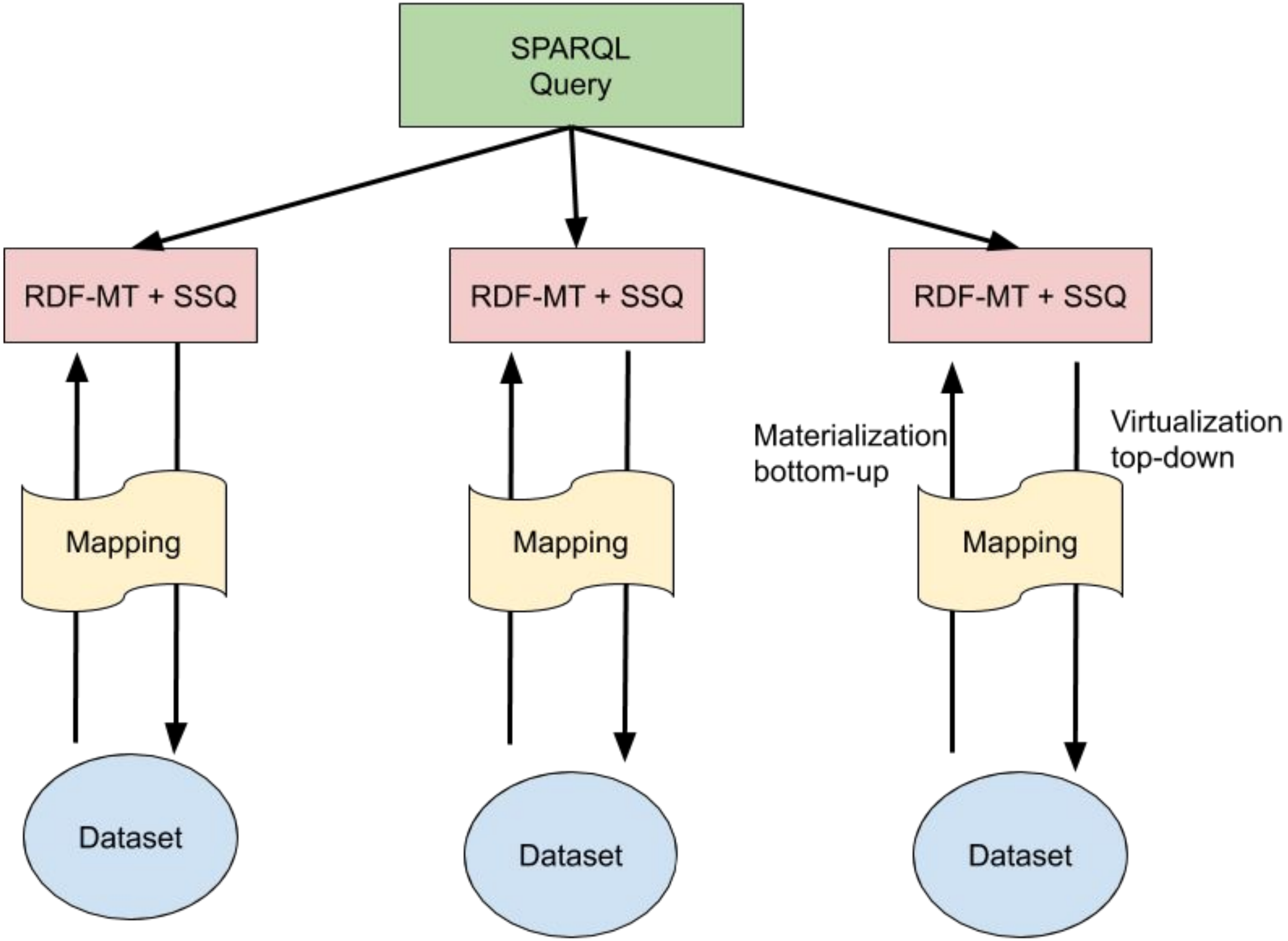
Fig. 3: Comparison of Knowledge Graph Creation Tools on Different Types of Relations. The first four (4) configurations, i.e. 1-4 in both x-axis and y-axis, represent results of SDM-RDFizer on *1-N*, *N-1*, *N-M*, and combination of all relations types, respectively. The later configurations, 5-8 both in x-axis and y-axis, shows results of RMLMapper on *1-N*, *N-1*, *N-M*, and combination of all relations types, respectively. Grey bubbles correspond to correlation value of 1.0; blue bubbles show a positive correlation while red bubbles show a negative correlation. The plots reveal that both type of relations and size of the dataset need to be taken into account to uncover patterns in the behaviour of the engines.

On Going Work: Benchmarks

1) GTFS-Bench: The OBDI Benchmark (virtual KG)
   a) Query translation over heterogeneous data sources
   b) Transport Domain (for SPRINT project)
   c) Nothing similar in SoA
   d) Target: SI on Benchmarking of JoWS

2) Evaluation of KG Creation Engines (materialization)
   a) Multiple use cases: Bio, Healthcare, Transport
   b) Based on RML engines and tabular data
   c) Involving the parameters from ODBASE paper
   d) Target: SWJ (Nov-Dec)

Virtual VS Materialized KG



Lessons Learned at TIB

- We are researches not engineerings
- One paper one idea (not like this presentation ;-))
- Writing a paper is not an art
- A paper starts from the experimentation
- Meetings are important
- Research is not only work
- What? and Why? before How?

Knowledge Graph Construction

David Chaves-Fraga, Ontology Engineering Group
Universidad Politécnica de Madrid, Spain
Data Integration Team

✉dchaves@fi.upm.es
@dchavesf

12/09/2019
OEG