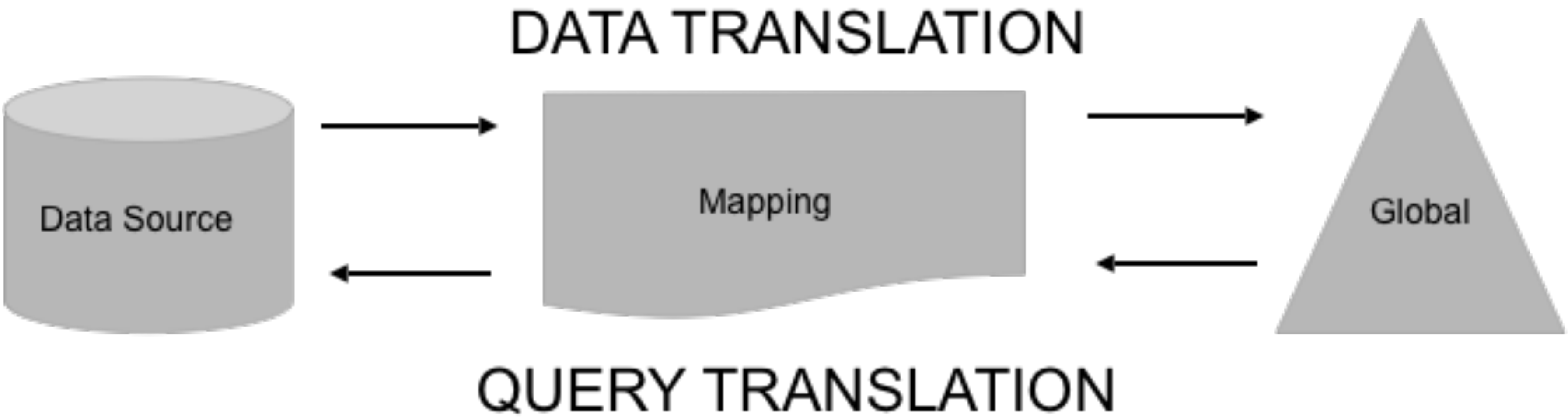
GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
AGENCIA ESTATAL DE INVESTIGACIÓN
Ontology Engineering Group

Knowledge Graph Construction and Access using Declarative Mappings

David Chaves-Fraga, Ontology Engineering Group
Universidad Politécnica de Madrid, Spain
Freddy Priyatna, Ahmad Alobaid, Andrea Cimmino
Ana Iglesias, Jhon Toledo, Daniel Doña, Luis Pozo,
Edna Ruckhaus, Oscar Corcho

✉ dchaves@fi.upm.es
@dchavesf

10/01/2020
Datos 4.0



DATA TRANSLATION

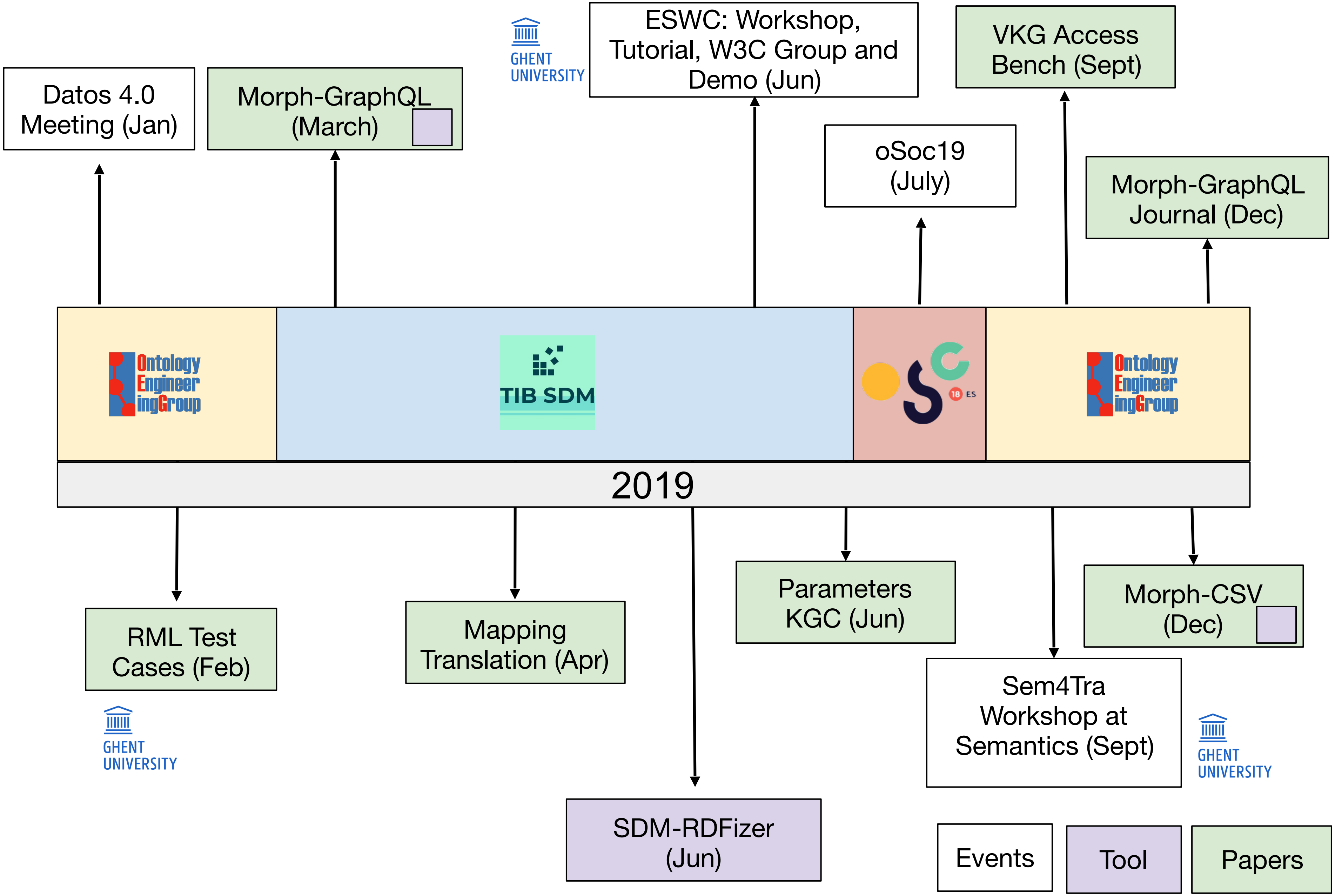
Data Source

Mapping

Global

QUERY TRANSLATION

2019



Datos 4.0 Meeting (Jan)

Morph-GraphQL (March)

GHENT UNIVERSITY

ESWC: Workshop, Tutorial, W3C Group and Demo (Jun)

oSoc19 (July)

VKG Access Bench (Sept)

Morph-GraphQL Journal (Dec)

Ontology Engineering Group

TIB SDM

Ontology Engineering Group

2019

RML Test Cases (Feb)

GHENT UNIVERSITY

Mapping Translation (Apr)

SDM-RDFizer (Jun)

Parameters KGC (Jun)

Sem4Tra Workshop at Semantics (Sept)

Morph-CSV (Dec)

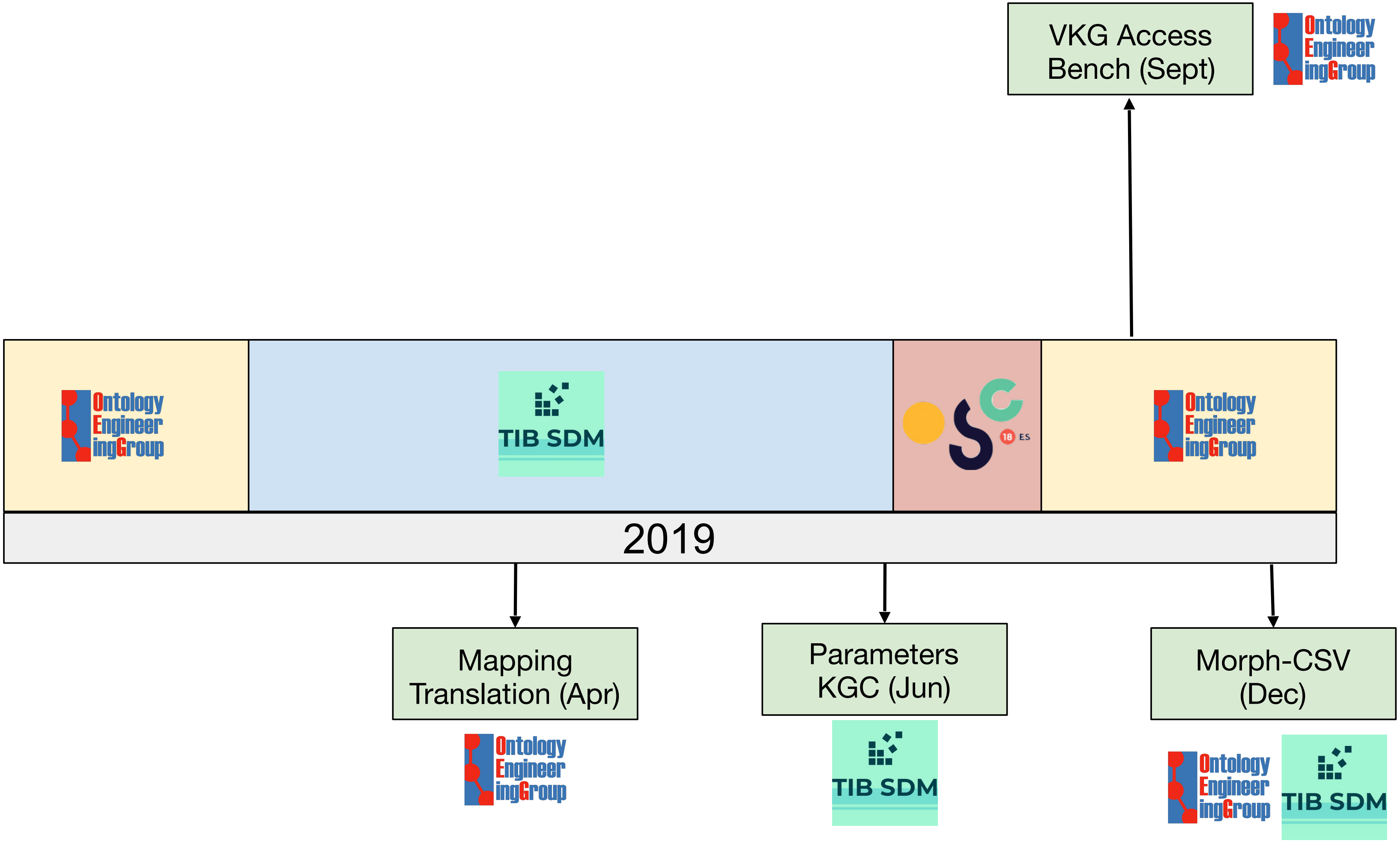
GHENT UNIVERSITY

Events

Tool

Papers

2019



VKG Access Bench (Sept)

2019

Mapping Translation (Apr)

Parameters KGC (Jun)

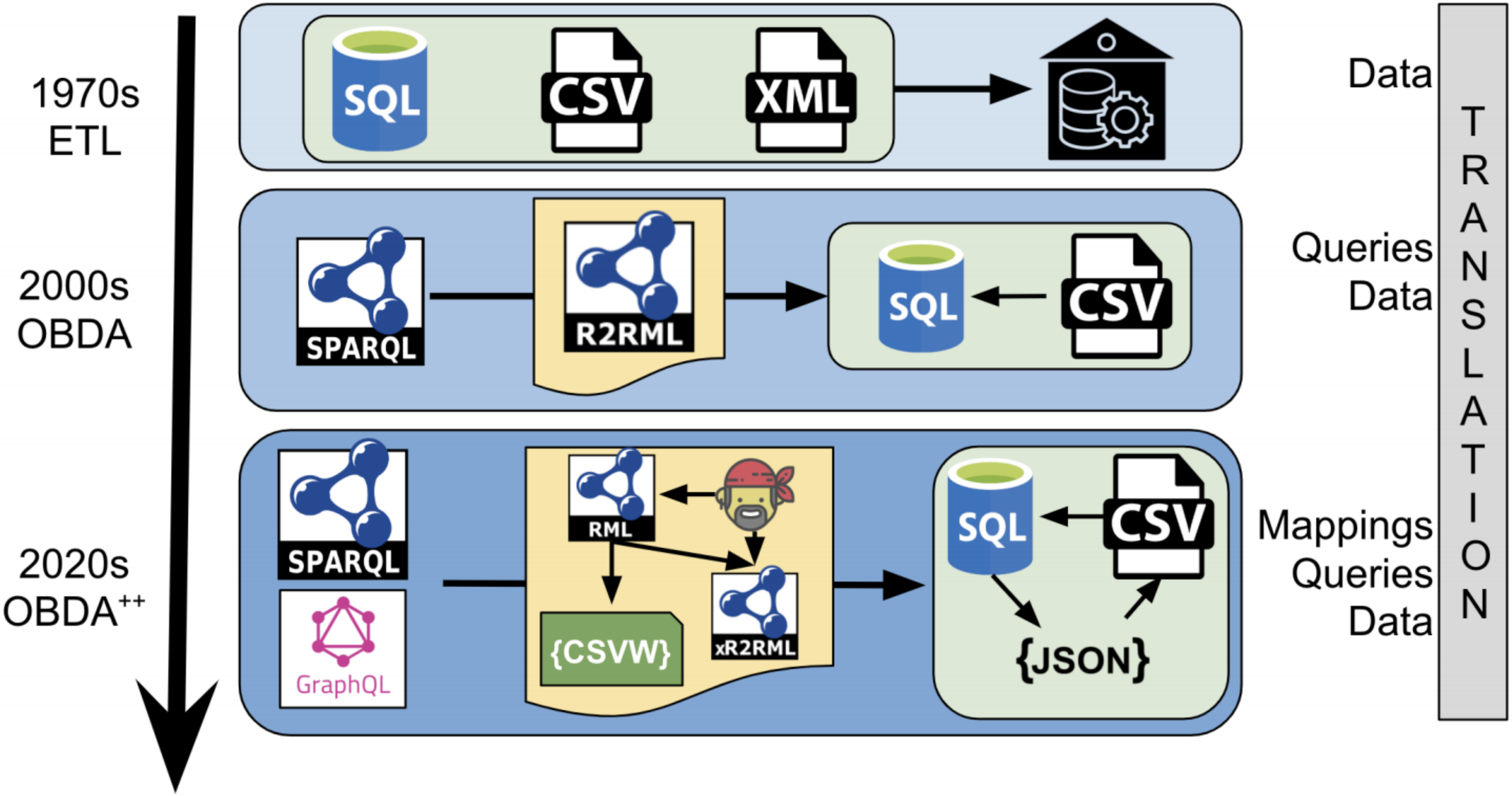
Morph-CSV (Dec)

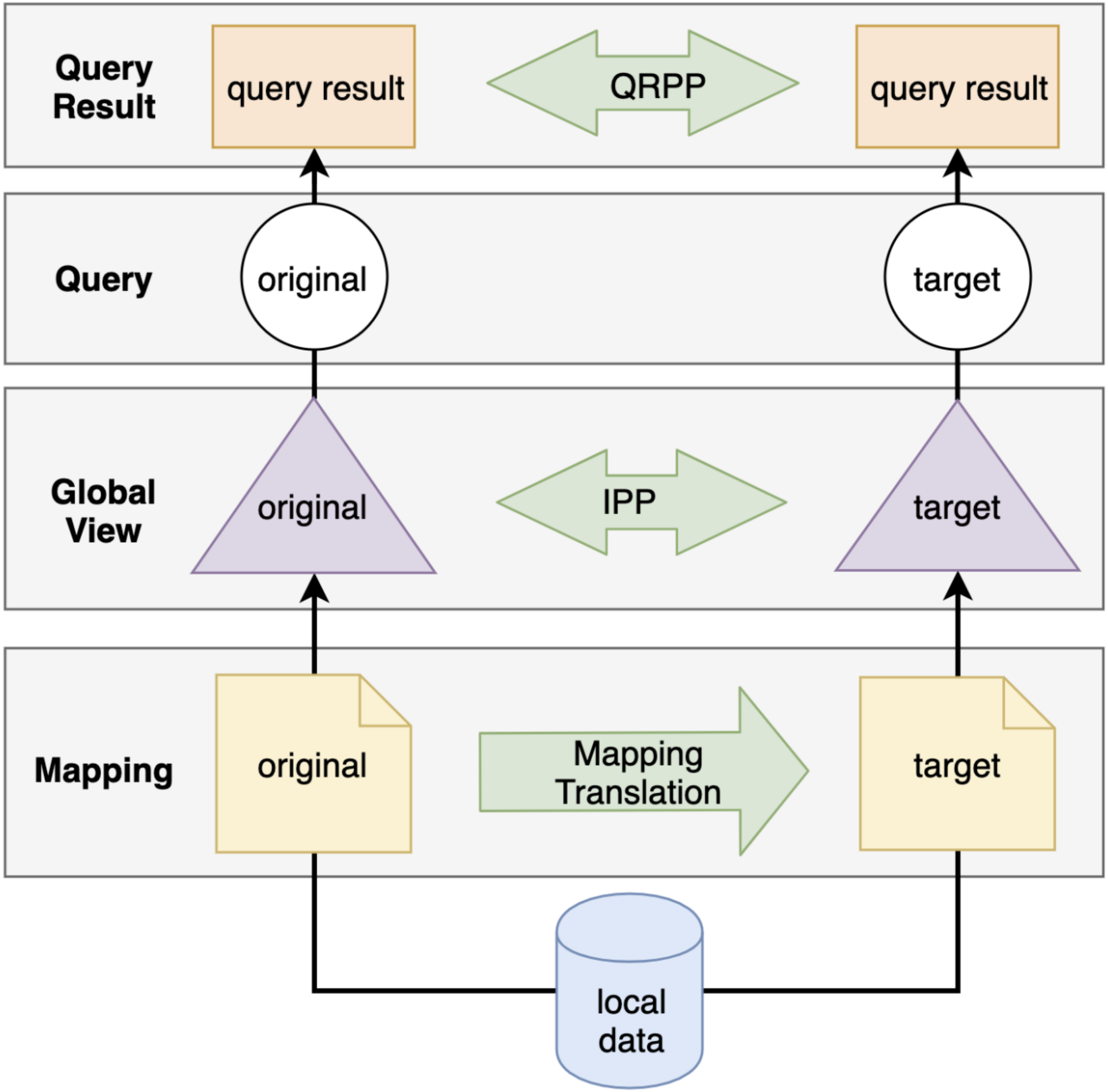
- Multiple use cases on KG Construction from Heterogeneous data sources (not same as RDB)
- Emergence of ad-hoc mapping languages to solve ad-hoc problems
- 1 mapping language → 1 tool

Mapping Translation

- Multiple use cases on KG Construction 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)


1970s
ETL
SQL
CSV
XML
Data
2000s
OBDA
SPARQL
R2RML
SQL
CSV
Queries
Data
2020s
OBDA++
SPARQL
GraphQL
RML
{CSVW}
xR2RML
SQL
CSV
{JSON}
Mappings
Queries
Data
TRANSLATION


Query Result
query result
QRPP
query result
Query
original
target
Global View
original
IPP
target
Mapping
original
Mapping Translation
target
local data

- 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

Parameters affecting KGC

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

Parameters affecting KGC

- Emergence of tools that process mapping rules for knowledge graph construction
- 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).

Size	SDM-RDFizer	RMLMapper
Two POM	1.72	0.92
Five POM	1.85	1.84
Ten POM	1.98	3.46

Motivating Example

Size	SDM-RDFizer	RMLMapper
Two POM	1.72	0.92
Five POM	1.85	1.84
Ten POM	1.98	3.46

↑ RMLMapper

↑ SDM-RDFizer

Motivating Example

Size	SDM-RDFizer	RMLMapper
Two POM	1.72	0.92
Five POM	1.85	1.84
Ten POM	1.98	3.46

↑ RMLMapper

↑ SDM-RDFizer

Join Selectivity	SDM-RDFizer	RMLMapper
High	2.16	38.6
Medium	2.20	40.43
Low	2.19	46.06

Motivating Example

Size	SDM-RDFizer	RMLMapper
Two POM	1.72	0.92
Five POM	1.85	1.84
Ten POM	1.98	3.46

⬆ RMLMapper

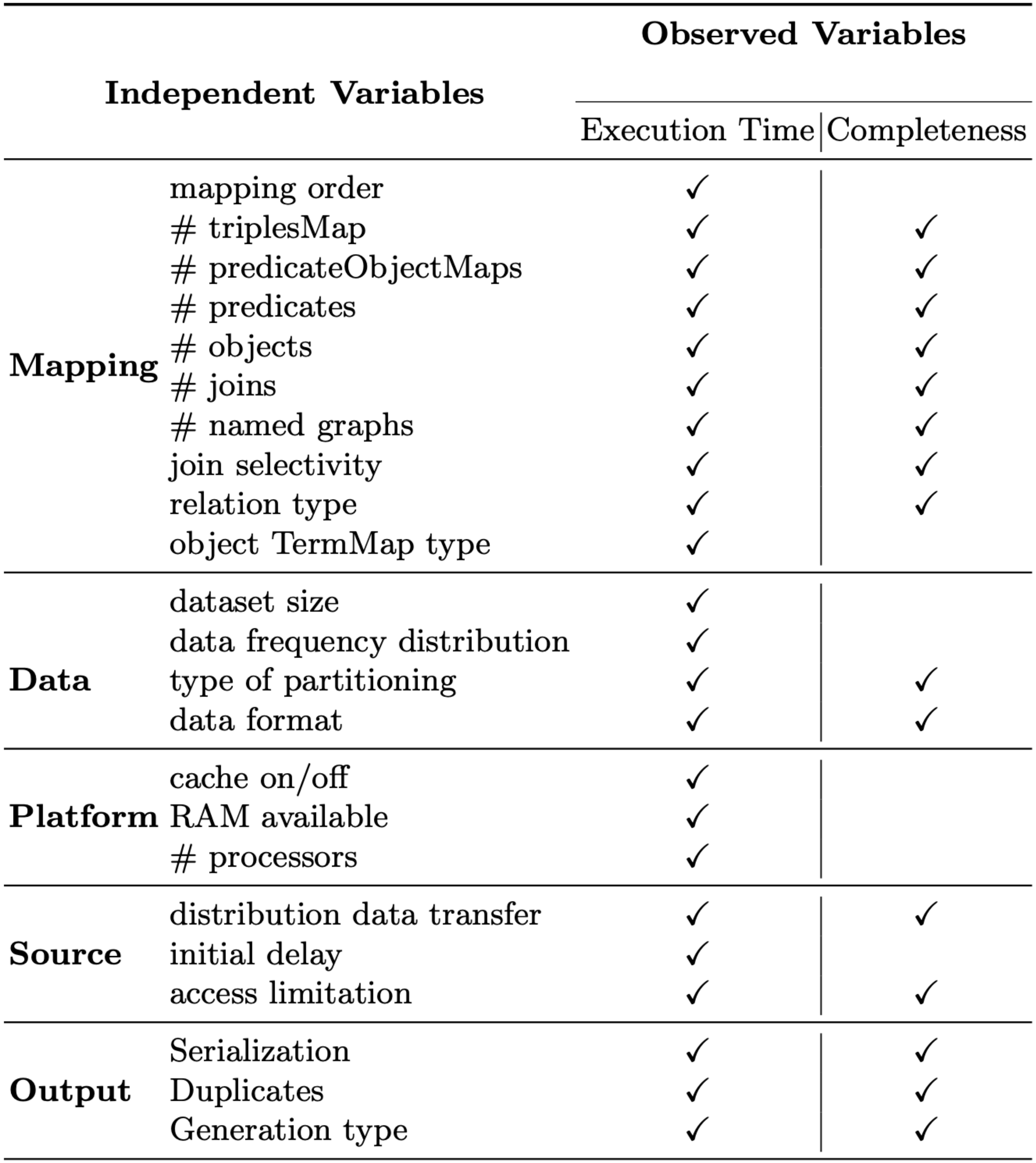
⬆ SDM-RDFizer

Join Selectivity	SDM-RDFizer	RMLMapper
High	2.16	38.6
Medium	2.20	40.43
Low	2.19	46.06

⬆ RMLMapper

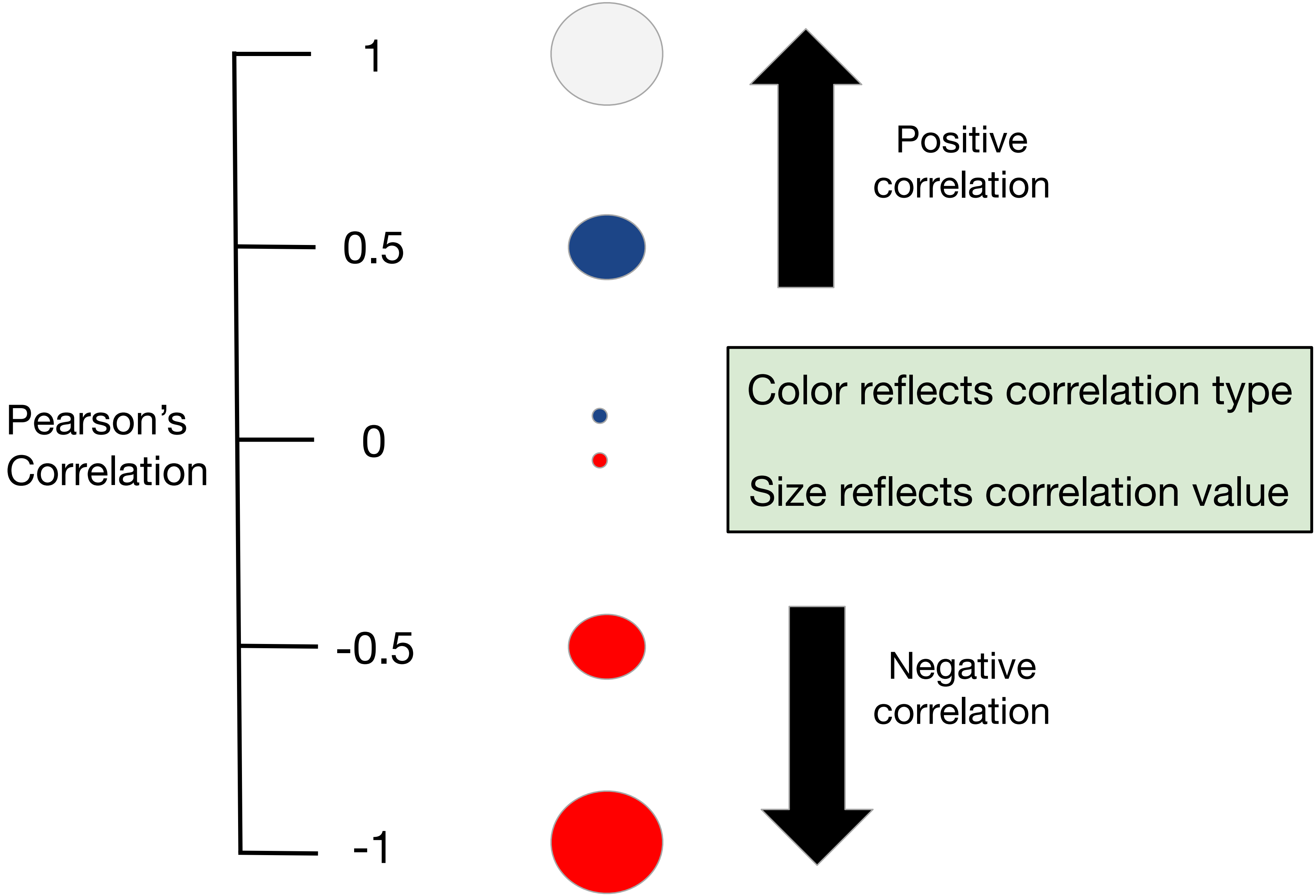
= SDM-RDFizer

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	✓	✓

Experimental setup



Pearson’s Correlation

1

0.5

0

-0.5

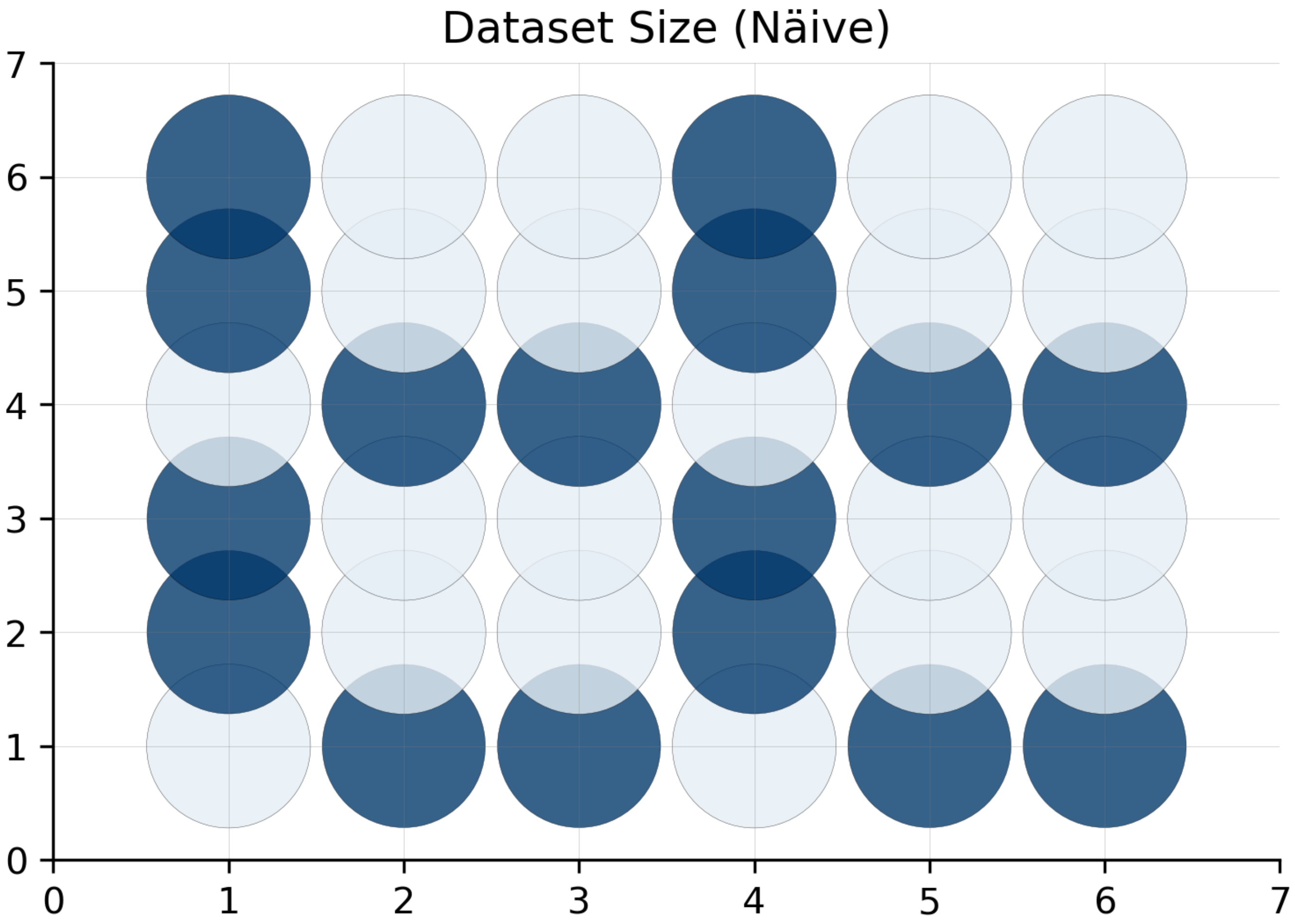
-1

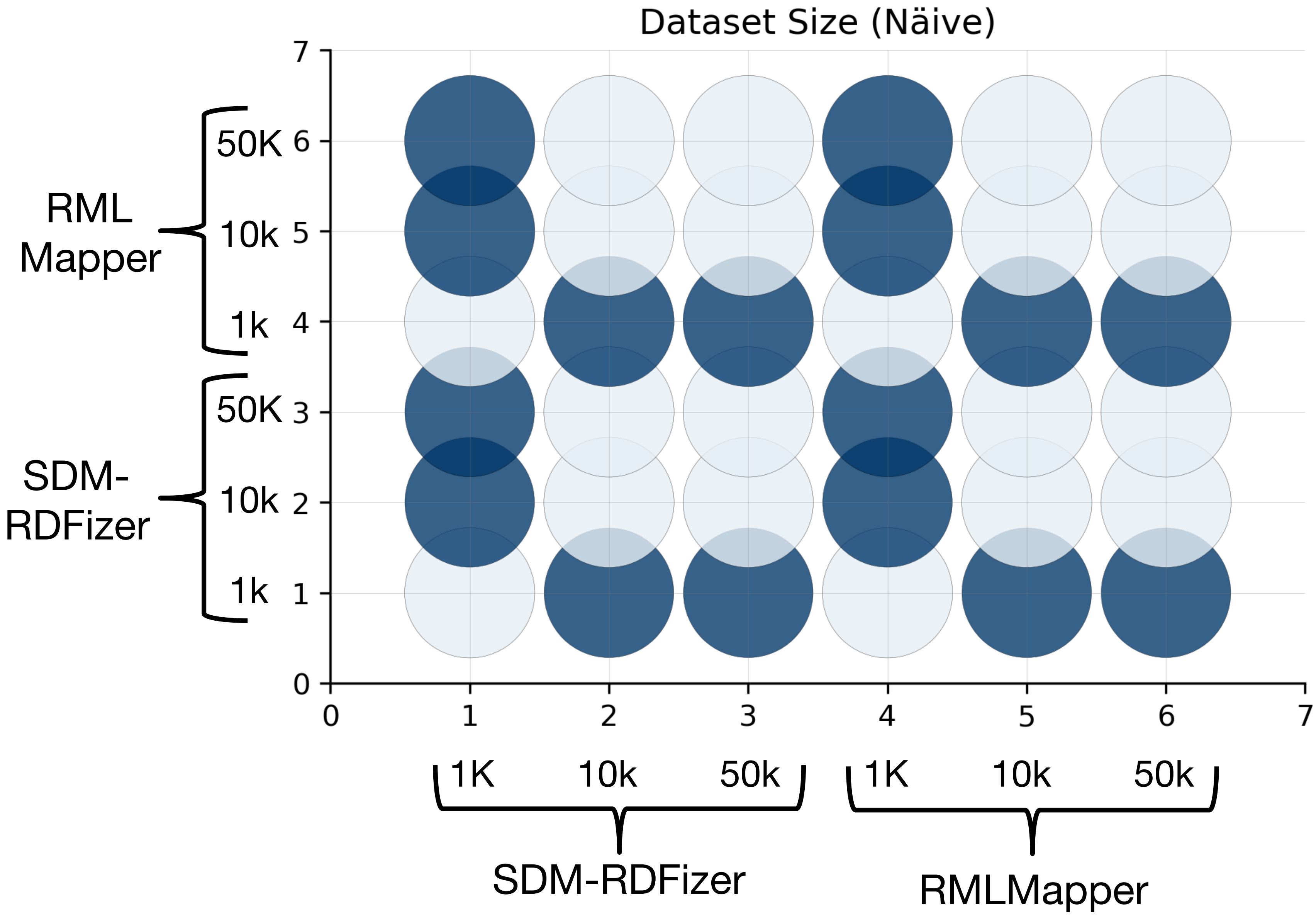
Positive correlation

Negative correlation

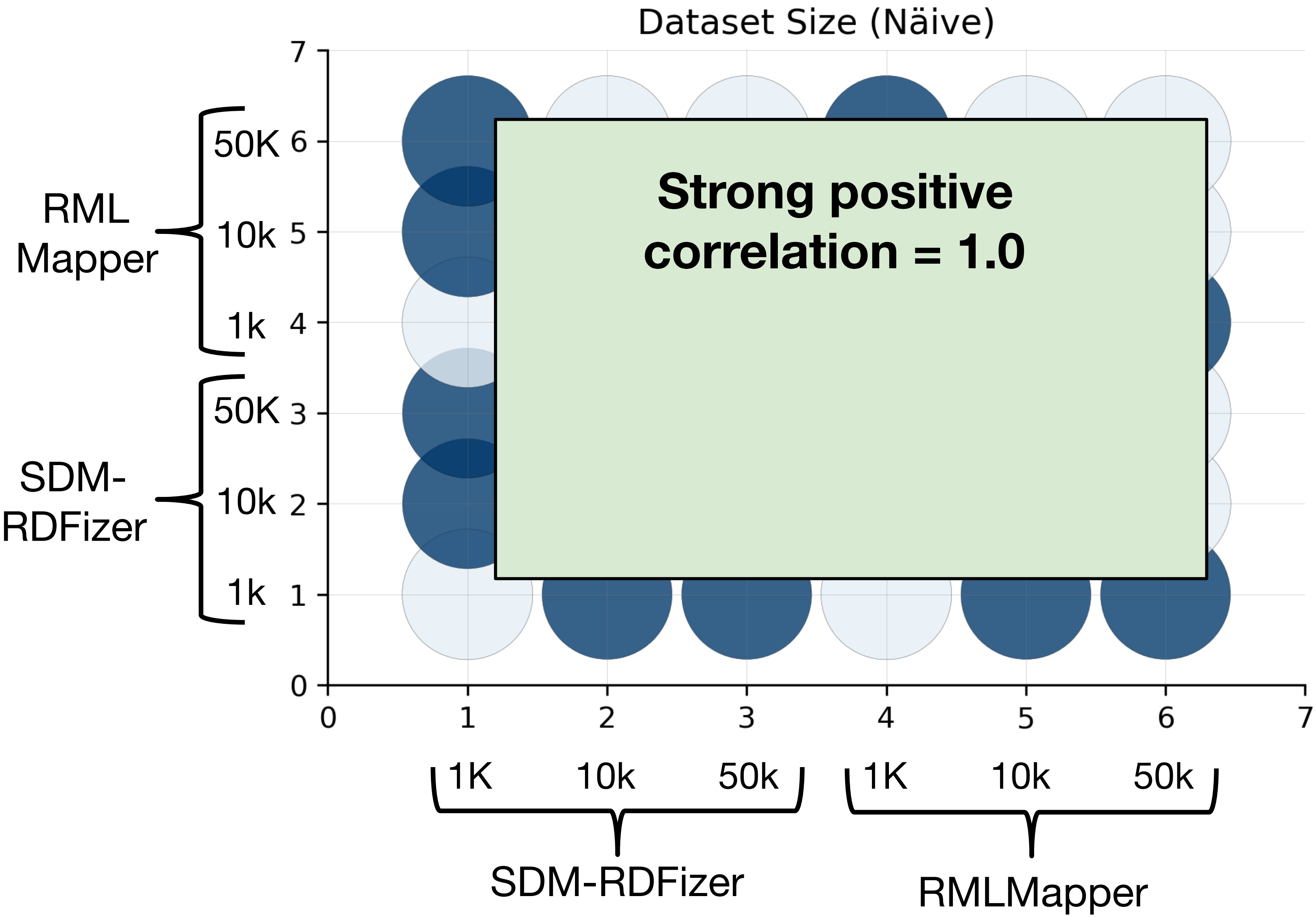
Color reflects correlation type

Size reflects correlation value

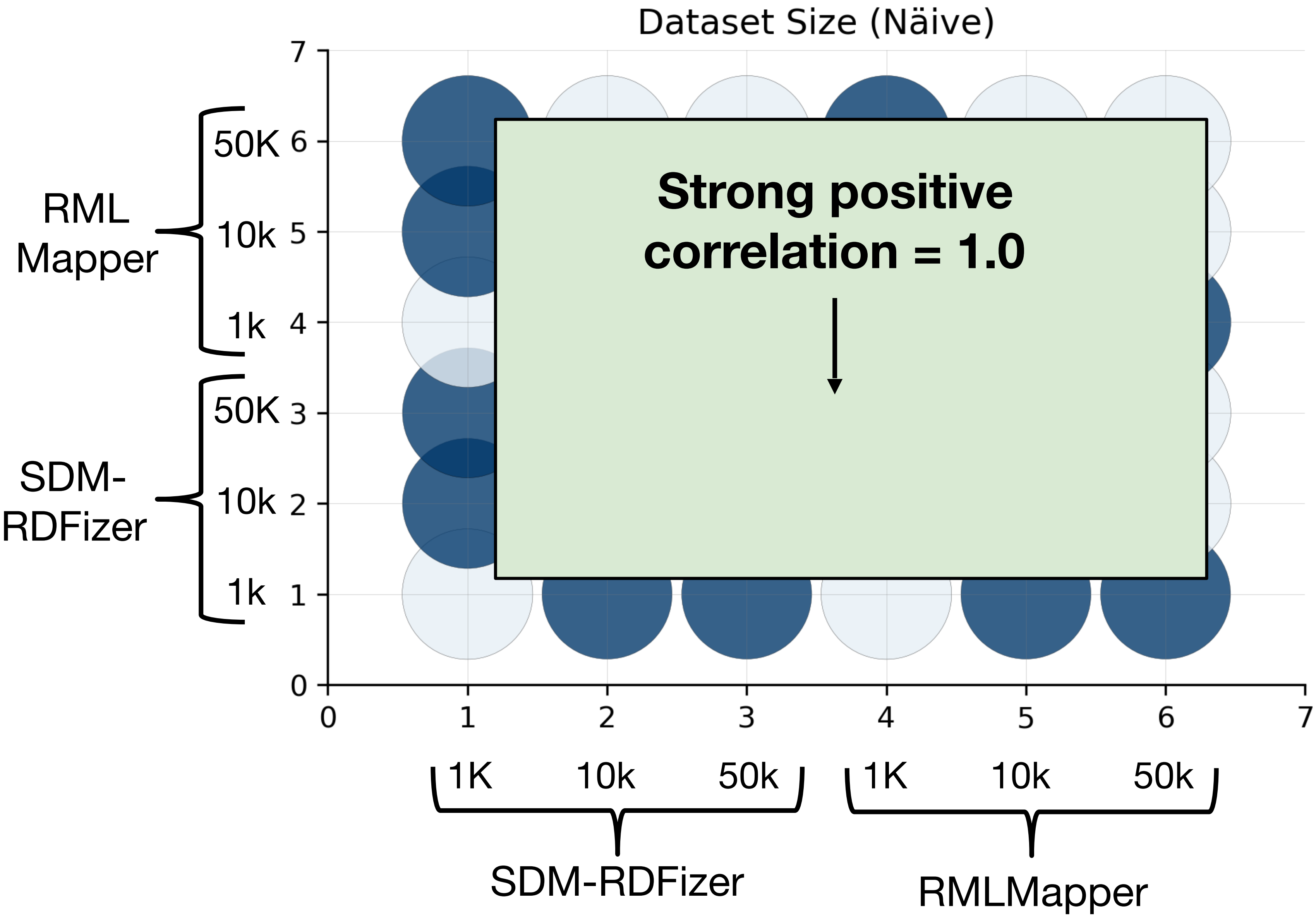

Dataset Size (Näive)
0
1
2
3
4
5
6
7
0
1
2
3
4
5
6
7


Dataset Size (Näive)
RML
Mapper
50K
10k
1k
SDM-
RDFizer
50K
10k
1k
0
1
2
3
4
5
6
7
0
1
2
3
4
5
6
7
1K
10k
50k
1K
10k
50k
SDM-RDFizer
RMLMapper

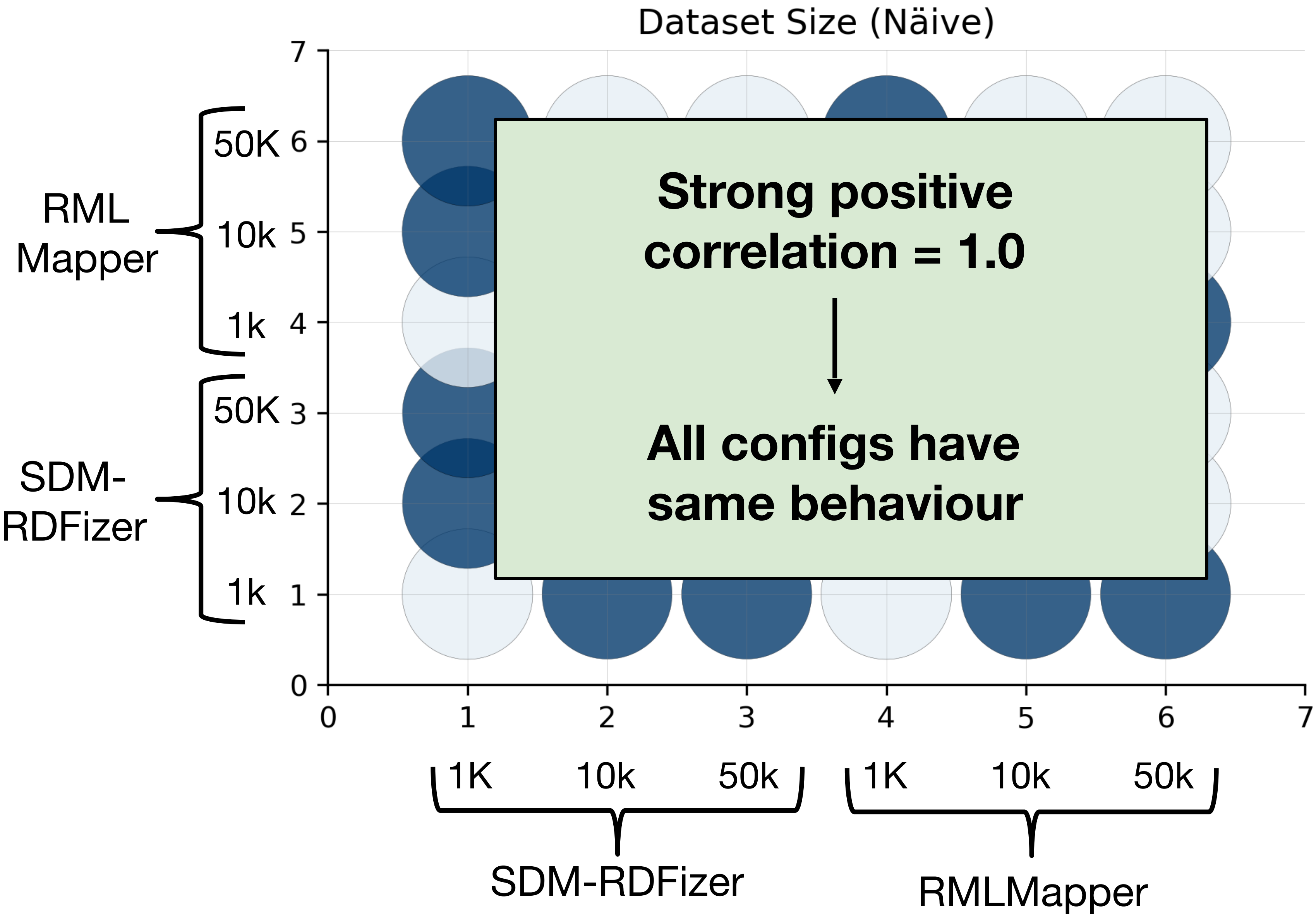
Results I: Naïve



Strong positive correlation = 1.0



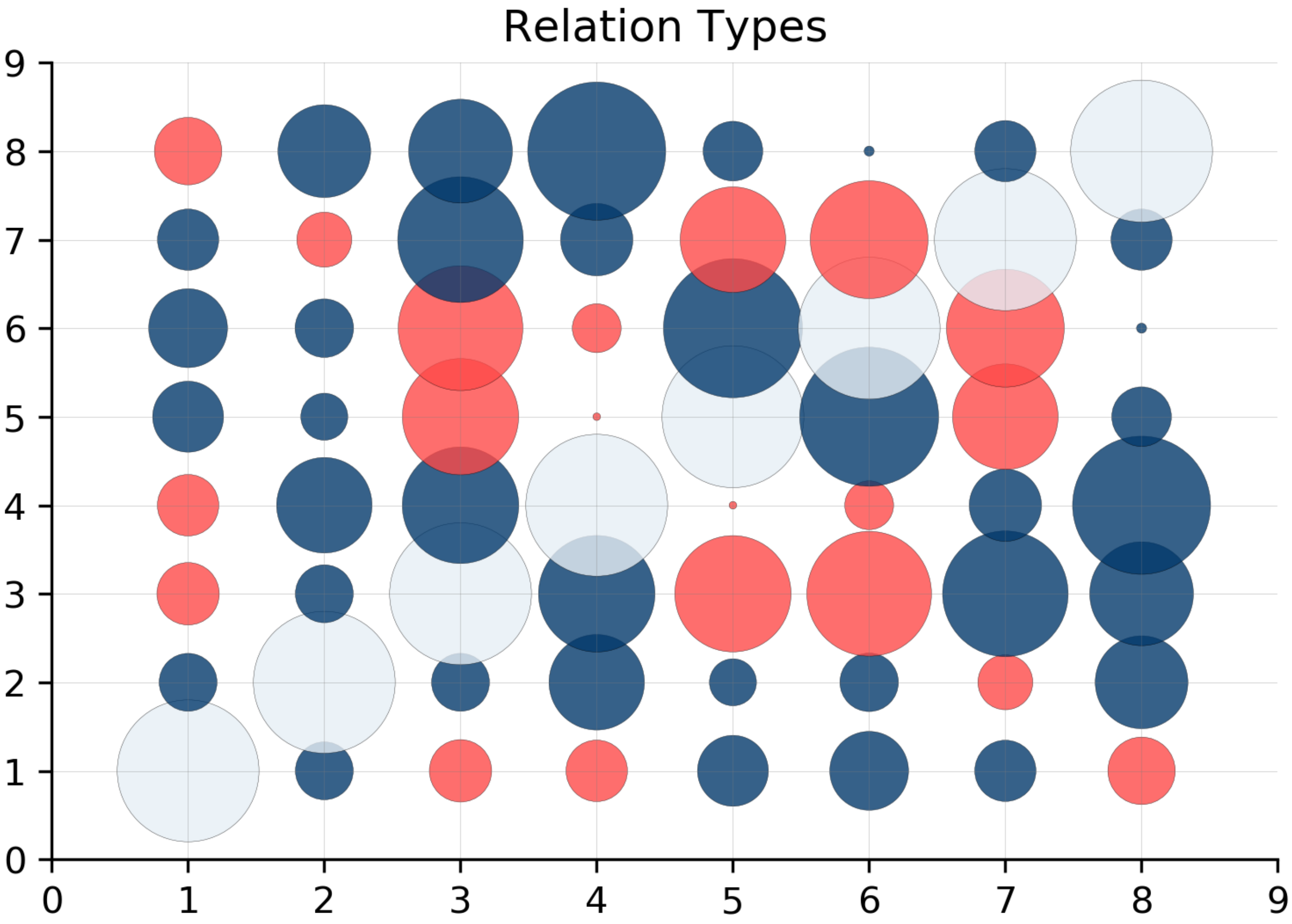
Strong positive correlation = 1.0

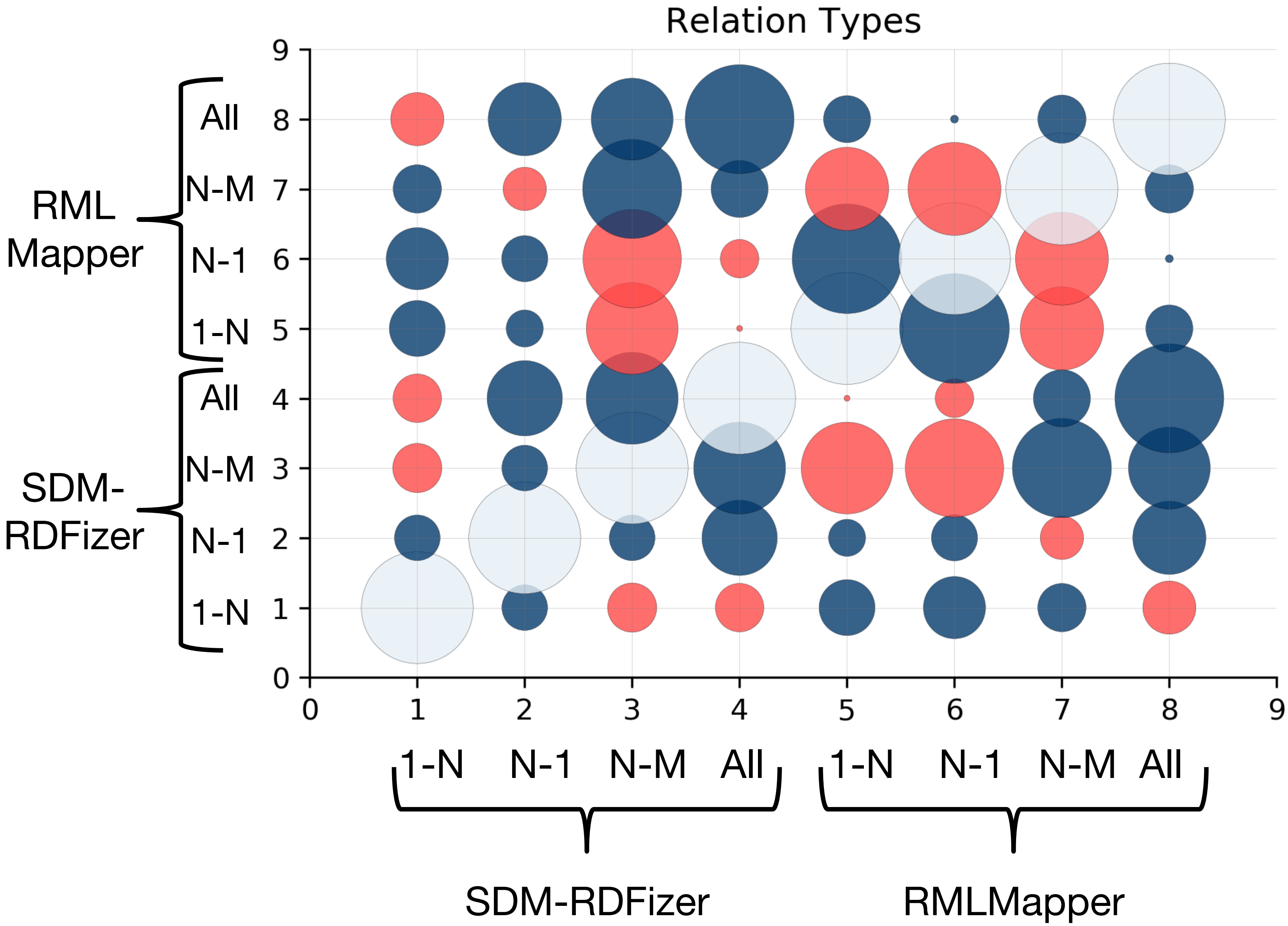


Strong positive correlation = 1.0

↓

All configs have same behaviour


Relation Types
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9


Relation Types
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
RML Mapper
All
N-M
N-1
1-N
SDM-RDFizer
All
N-M
N-1
1-N
1-N
N-1
N-M
All
1-N
N-1
N-M
All
SDM-RDFizer
RMLMapper

GTFS-Madrid-Bench: A VKG Benchmark

A comprehensive benchmark for virtual knowledge graph access, which considers multiple data formats and different data scales:

- Query translation over heterogeneous data sources
- Transport Domain (GTFS)
- OBDA/OBDI
- Tested over 5 tools from the state of the art

Paper (Under Review - JoWS): David Chaves-Fraga, Freddy Priyatna, Andrea Cimmino, Jhon Toledo, Edna Ruckhaus, Oscar Corcho. **GTFS-Madrid-Bench: A Benchmark for Virtual Knowledge Graph Access in the Transport Domain**

GTFS-Madrid-Bench: A VKG Benchmark

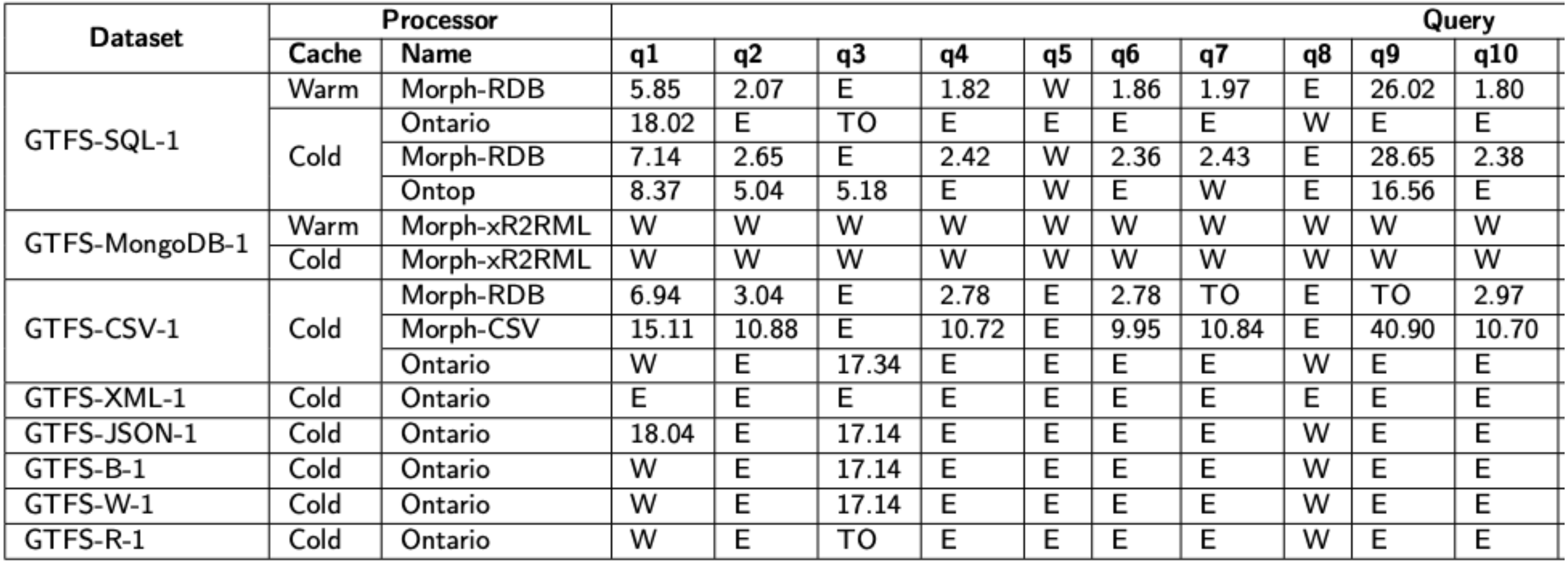
A comprehensive benchmark for virtual knowledge graph access, which considers multiple data formats and different data scales:

- Query translation over heterogeneous data sources
- Transport Domain (GTFS)
- OBDA/OBDI
- Tested over 5 tools from the state of the art

Paper (Under Review - JoWS): David Chaves-Fraga, Freddy Priyatna, Andrea Cimmino, Jhon Toledo, Edna Ruckhaus, Oscar Corcho. **GTFS-Madrid-Bench: A Benchmark for Virtual Knowledge Graph Access in the Transport Domain**

- **Data:** we have generated from several datasets (GTFS-[1,5,10,50,100,500]) in multiple formats (CSV, JSON, XML, SQL, MongoDB). The preparation script will download all these datasets and generate a docker-image for each dataset which is contained in a database (MySQL and MongoDB)
- **Generation:** If any practitioner or developer want to create datasets with other scale values all the resources are available.
- **Queries:** 18 queries increasing in terms of complexity.
- **Mappings:** 1 R2RML mapping document, 7 RML mapping document, 1 xR2RML mapping document, 1 YARRRML mapping and 1 CSVW annotations
- **Engines:** docker-compose with all the tested engines and running scripts

Some results….



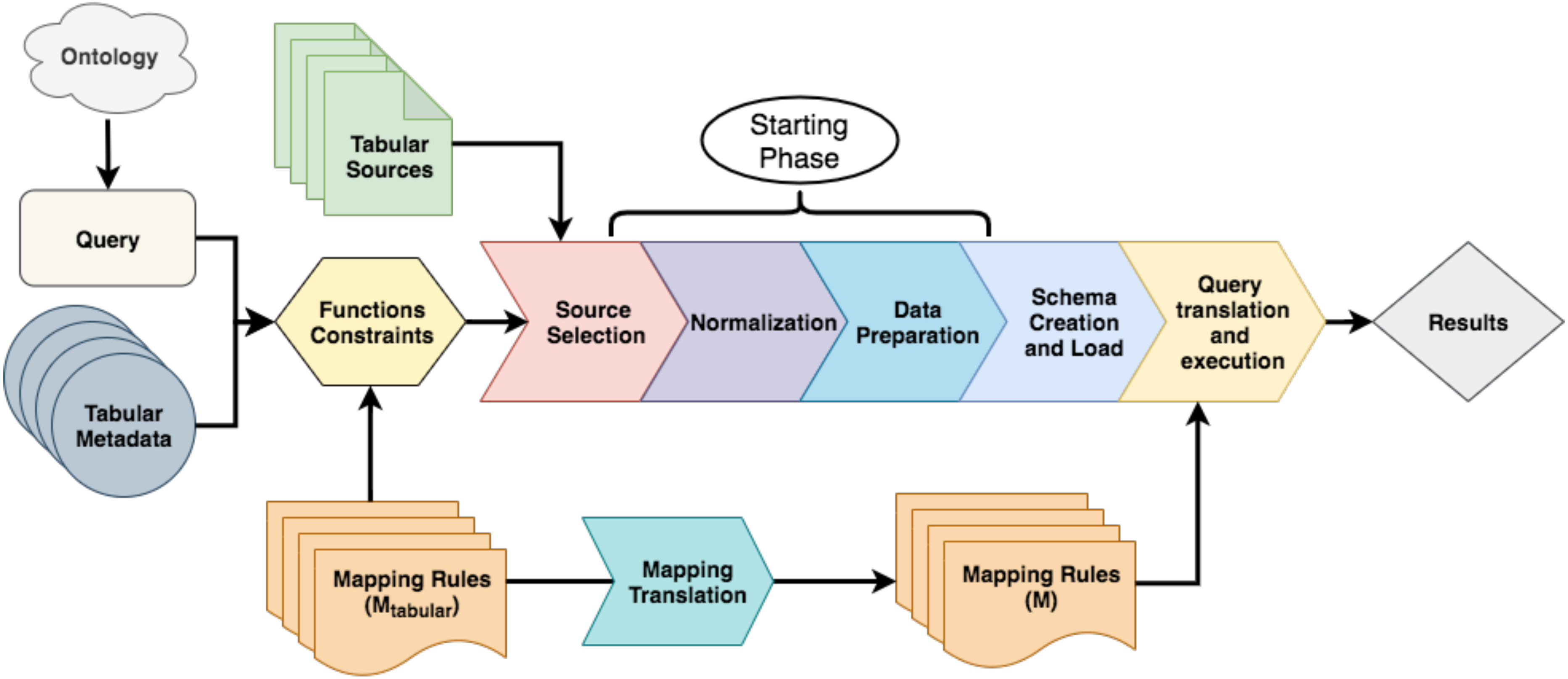
Dataset	Processor		Query									
	Cache	Name	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10
GTFS-SQL-1	Warm	Morph-RDB	5.85	2.07	E	1.82	W	1.86	1.97	E	26.02	1.80
	Cold	Ontario	18.02	E	TO	E	E	E	E	W	E	E
		Morph-RDB	7.14	2.65	E	2.42	W	2.36	2.43	E	28.65	2.38
		Ontop	8.37	5.04	5.18	E	W	E	W	E	16.56	E
GTFS-MongoDB-1	Warm	Morph-xR2RML	W	W	W	W	W	W	W	W	W	W
	Cold	Morph-xR2RML	W	W	W	W	W	W	W	W	W	W
GTFS-CSV-1	Cold	Morph-RDB	6.94	3.04	E	2.78	E	2.78	TO	E	TO	2.97
		Morph-CSV	15.11	10.88	E	10.72	E	9.95	10.84	E	40.90	10.70
		Ontario	W	E	17.34	E	E	E	E	W	E	E
GTFS-XML-1	Cold	Ontario	E	E	E	E	E	E	E	E	E	E
GTFS-JSON-1	Cold	Ontario	18.04	E	17.14	E	E	E	E	W	E	E
GTFS-B-1	Cold	Ontario	W	E	17.14	E	E	E	E	W	E	E
GTFS-W-1	Cold	Ontario	W	E	17.14	E	E	E	E	W	E	E
GTFS-R-1	Cold	Ontario	W	E	TO	E	E	E	E	W	E	E

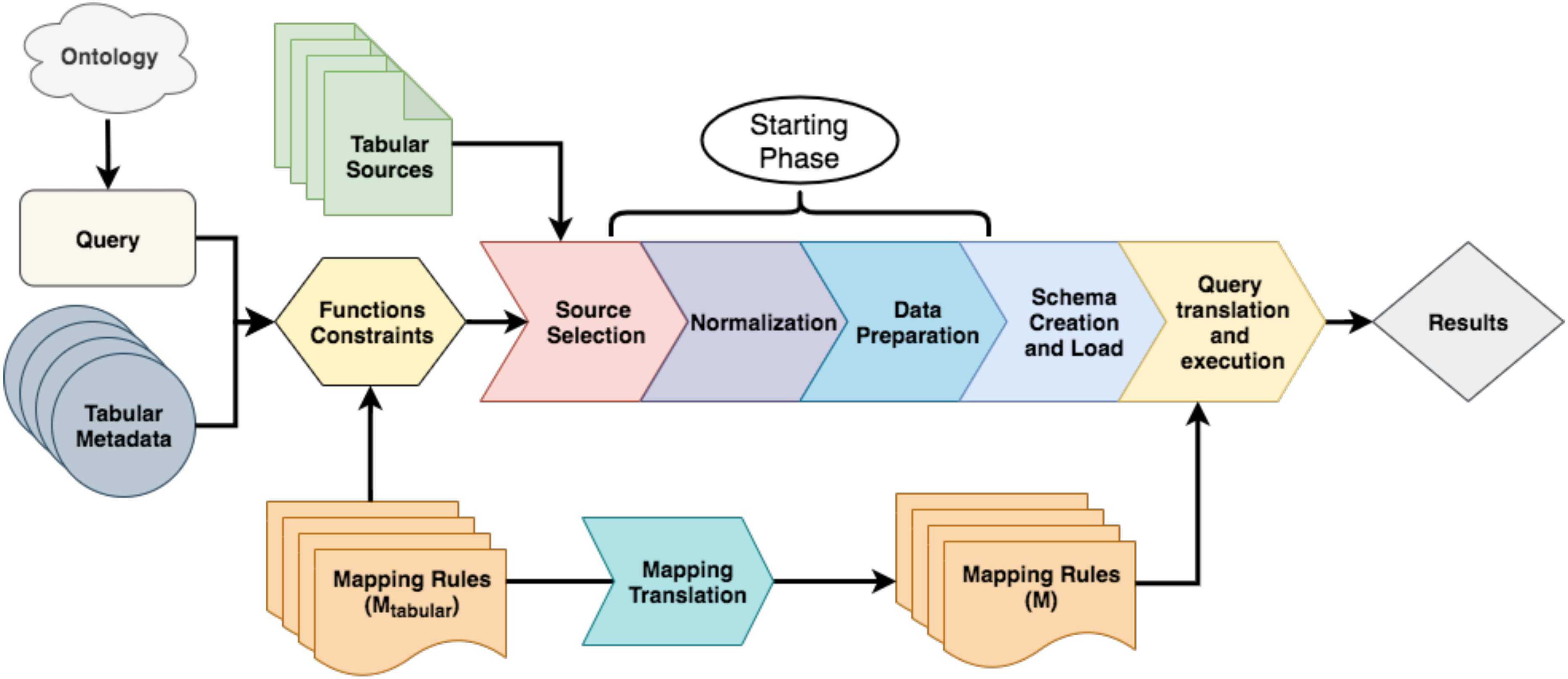
Enhancing OBDA query translation* over Tabular Data

Exploit query/mapping/annotations to enforcing implicit constraints during OBDA query translation:

- Source selection
- Data normalization + data preparation
- Schema creation and loading
- Mapping translation process (to RML/R2RML)
- Can be embedded in the top of any OBDA engine

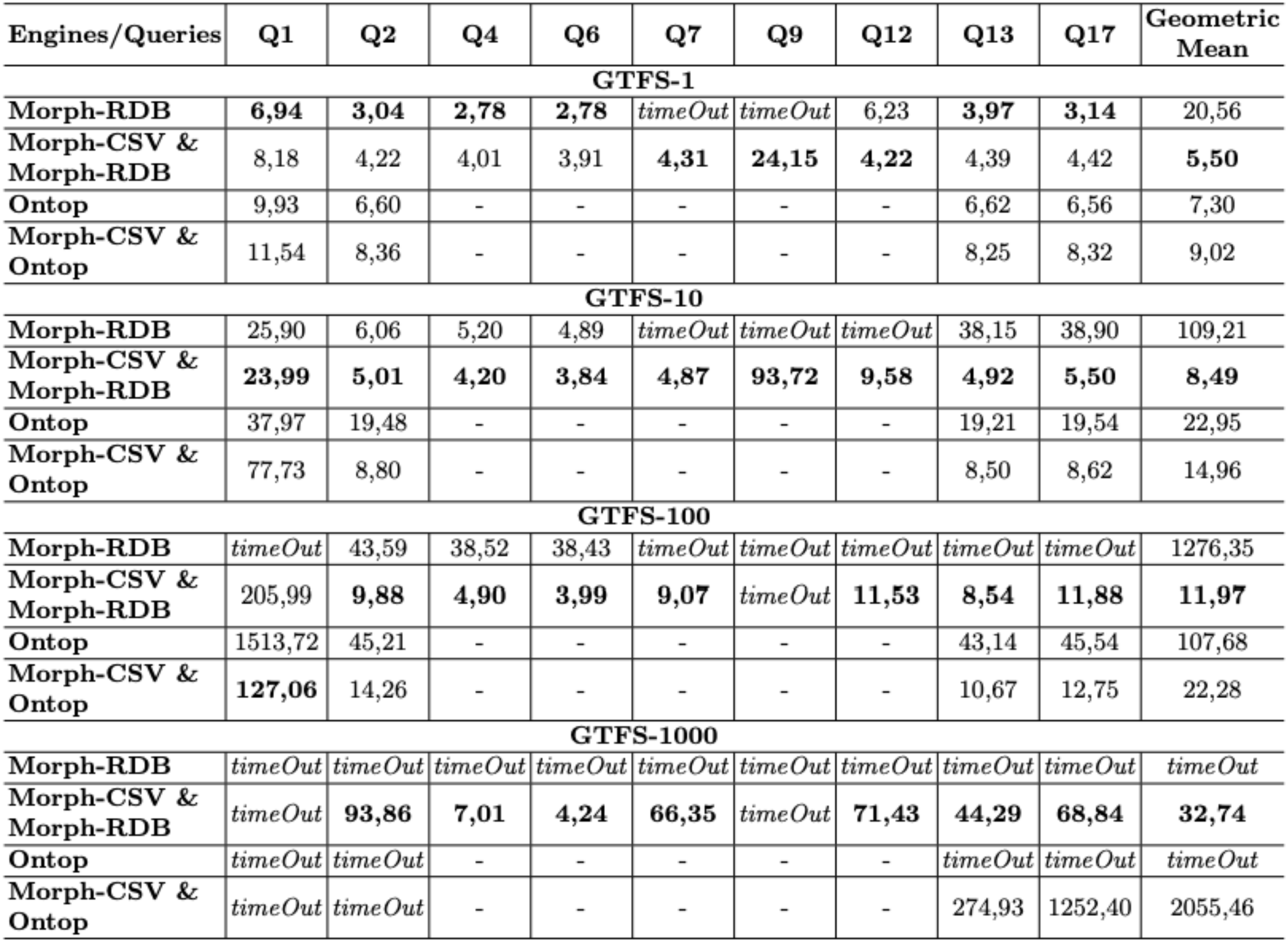
Paper (Under Review - ESWC 2020): David Chaves-Fraga, Edna Ruckhaus,Freddy Priyatna, Maria-Esther Vidal and Oscar Corcho. **Enhancing OBDA query translation over Tabular Data with Morph-CSV**.


Ontology
Tabular Sources
Starting Phase
Query
Functions Constraints
Source Selection
Normalization
Data Preparation
Schema Creation and Load
Query translation and execution
Results
Tabular Metadata
Mapping Rules ($M_{tabular}$)
Mapping Translation
Mapping Rules (M)



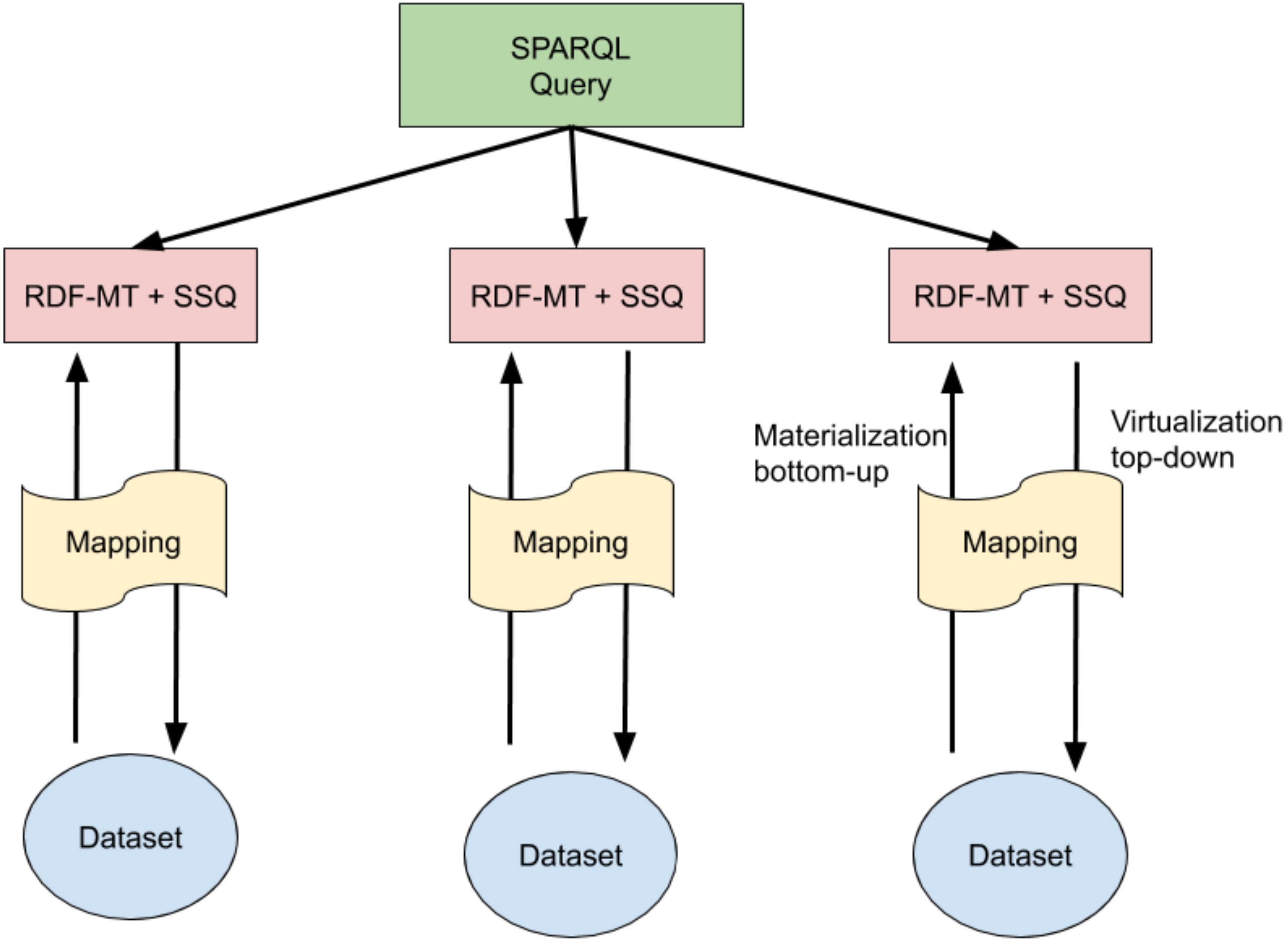
- Push down the application of the steps before query execution.
- Vertical Partitioning
- Complete solution (Future work): Horizontal Partitioning

Performance



Engines/Queries	Q1	Q2	Q4	Q6	Q7	Q9	Q12	Q13	Q17	Geometric Mean
GTFS-1										
Morph-RDB	**6,94**	**3,04**	**2,78**	**2,78**	*timeOut*	*timeOut*	6,23	**3,97**	**3,14**	20,56
Morph-CSV & Morph-RDB	8,18	4,22	4,01	3,91	**4,31**	**24,15**	**4,22**	4,39	4,42	**5,50**
Ontop	9,93	6,60	-	-	-	-	-	6,62	6,56	7,30
Morph-CSV & Ontop	11,54	8,36	-	-	-	-	-	8,25	8,32	9,02
GTFS-10										
Morph-RDB	25,90	6,06	5,20	4,89	*timeOut*	*timeOut*	*timeOut*	38,15	38,90	109,21
Morph-CSV & Morph-RDB	**23,99**	**5,01**	**4,20**	**3,84**	**4,87**	**93,72**	**9,58**	**4,92**	**5,50**	**8,49**
Ontop	37,97	19,48	-	-	-	-	-	19,21	19,54	22,95
Morph-CSV & Ontop	77,73	8,80	-	-	-	-	-	8,50	8,62	14,96
GTFS-100										
Morph-RDB	*timeOut*	43,59	38,52	38,43	*timeOut*	*timeOut*	*timeOut*	*timeOut*	*timeOut*	1276,35
Morph-CSV & Morph-RDB	205,99	**9,88**	**4,90**	**3,99**	**9,07**	*timeOut*	**11,53**	**8,54**	**11,88**	**11,97**
Ontop	1513,72	45,21	-	-	-	-	-	43,14	45,54	107,68
Morph-CSV & Ontop	**127,06**	14,26	-	-	-	-	-	10,67	12,75	22,28
GTFS-1000										
Morph-RDB	*timeOut*	*timeOut*	*timeOut*	*timeOut*	*timeOut*	*timeOut*	*timeOut*	*timeOut*	*timeOut*	*timeOut*
Morph-CSV & Morph-RDB	*timeOut*	**93,86**	**7,01**	**4,24**	**66,35**	*timeOut*	**71,43**	**44,29**	**68,84**	**32,74**
Ontop	*timeOut*	*timeOut*	-	-	-	-	-	*timeOut*	*timeOut*	*timeOut*
Morph-CSV & Ontop	*timeOut*	*timeOut*	-	-	-	-	-	274,93	1252,40	2055,46

Virtual VS Materialized KG



Accepted:

- A. Iglesias-Molina, D. Chaves-Fraga, F. Priyatna, and O. Corcho. **Enhancing the Maintainability of the Bio2RDF Project Using Declarative Mappings.** In *Proceedings of the 12th International Conference on Semantic Web Applications and Tools for Healthcare and Life Sciences*, 2019.
- A. Iglesias-Molina, D. Chaves-Fraga, F. Priyatna and O. Corcho: **Towards the definition of a language-independent mapping template for knowledge graph creation**. In *Proceedings of the Third International Workshop on Capturing Scientific Knowledge co-located with the 10th International Conference on Knowledge Capture*, 2019

Future Work:

- Knowledge Graph Construction in the Biomedical Domain

Knowledge Graph Construction and Access

David Chaves-Fraga, Ontology Engineering Group
Universidad Politécnica de Madrid, Spain
Freddy Priyatna, Ahmad Alobaid, Andrea Cimmino
Ana Iglesias, Jhon Toledo, Edna Ruckhaus, Oscar Corcho

✉dchaves@fi.upm.es
@dchavesf

📅10/01/2020
📍 Datos 4.0