KU LEUVEN

Knowledge Graph Construction

From Optimization To Automation

Dr. David Chaves-Fraga

✉ david.chaves@kuleuven.be

@dchavesf

A bit about myself…

PhD in Artificial Intelligence at Universidad Politécnica de Madrid (2021)

“Knowledge Graph Construction from Heterogeneous Data Sources Exploiting Declarative Mapping Rules”

POLITÉCNICA

Co-chair W3C Community Group on Knowledge Graph Construction (2019-now)

Main Coordinator of Open Summer of Code Spain (2018-2021)

Workshop Organizer: Knowledge Graph Construction and Semantics For Transport

KGC Workshop

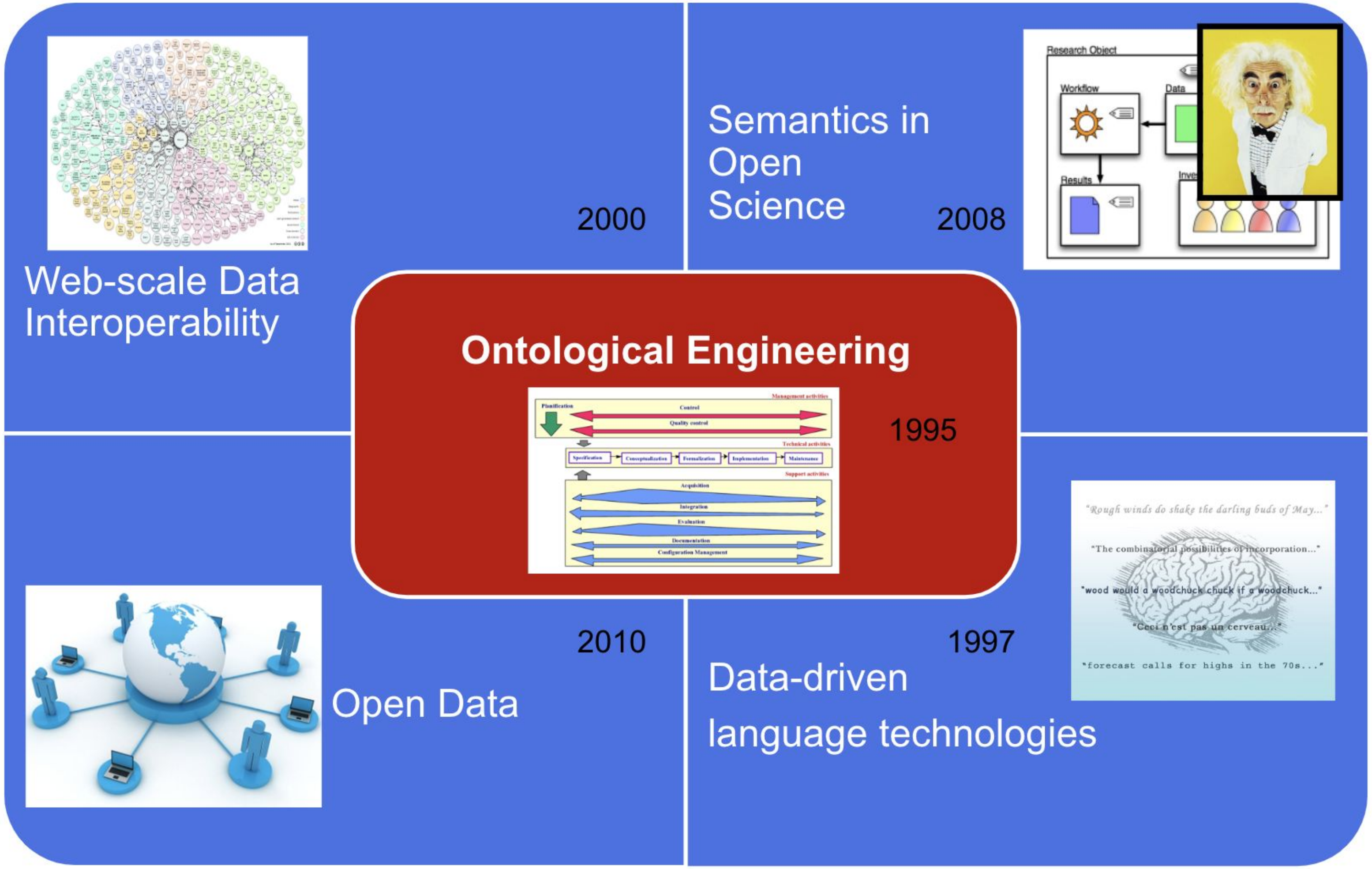
SEM4TRA

Ontology Engineering Group

Universidad Politécnica de Madrid (UPM)
(most important technical university in Spain)

Directors: Asunción Gómez-Pérez, Oscar Corcho

Position: Top-3 in the UPM ranking (>250 groups)



Web-scale Data Interoperability

2000

Semantics in Open Science

2008

Ontological Engineering

1995

Open Data

2010

Data-driven language technologies

1997

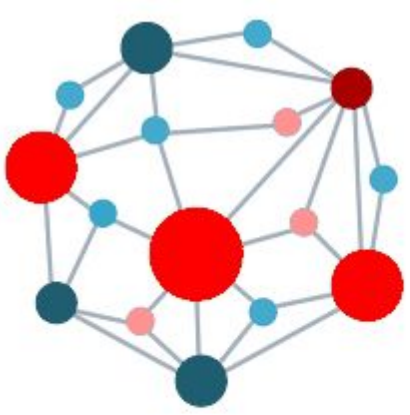
Research Object
Workflow
Data
Results
"Rough winds do shake the darling buds of May..."
"The combinatorial possibilities of incorporation..."
"wood would a woodchuck chuck if a woodchuck..."
"Ceci n'est pas un cerveau..."
"forecast calls for highs in the 70s..."

Knowledge Graphs

KG Embeddings

Explainable AI

Multilinguality



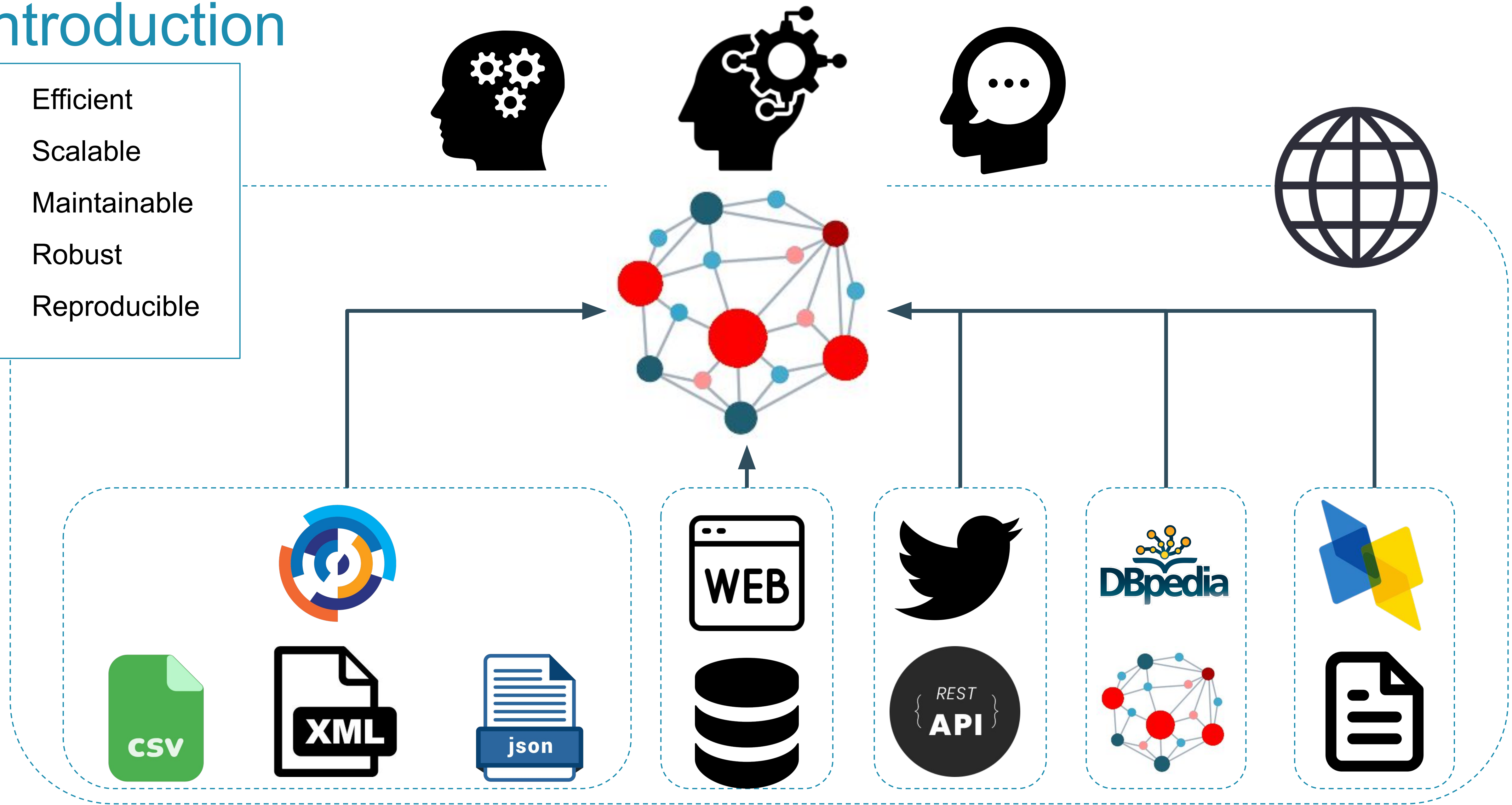
Question Answering Systems

Search Interfaces

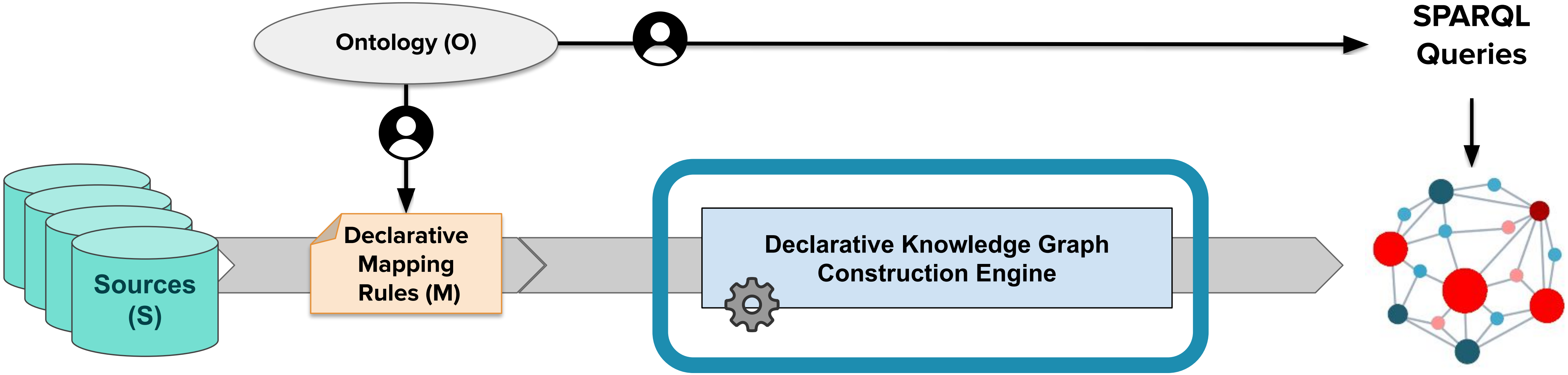
Data Labelling

Introduction

- Efficient
- Scalable
- Maintainable
- Robust
- Reproducible



Knowledge Graph Construction


Ontology (O)
SPARQL Queries
Sources (S)
Declarative Mapping Rules (M)
Declarative Knowledge Graph Construction Engine

Knowledge Graph Construction = Data Integration System (DIS) = <S, M, O>

Main Contributions to the SoA

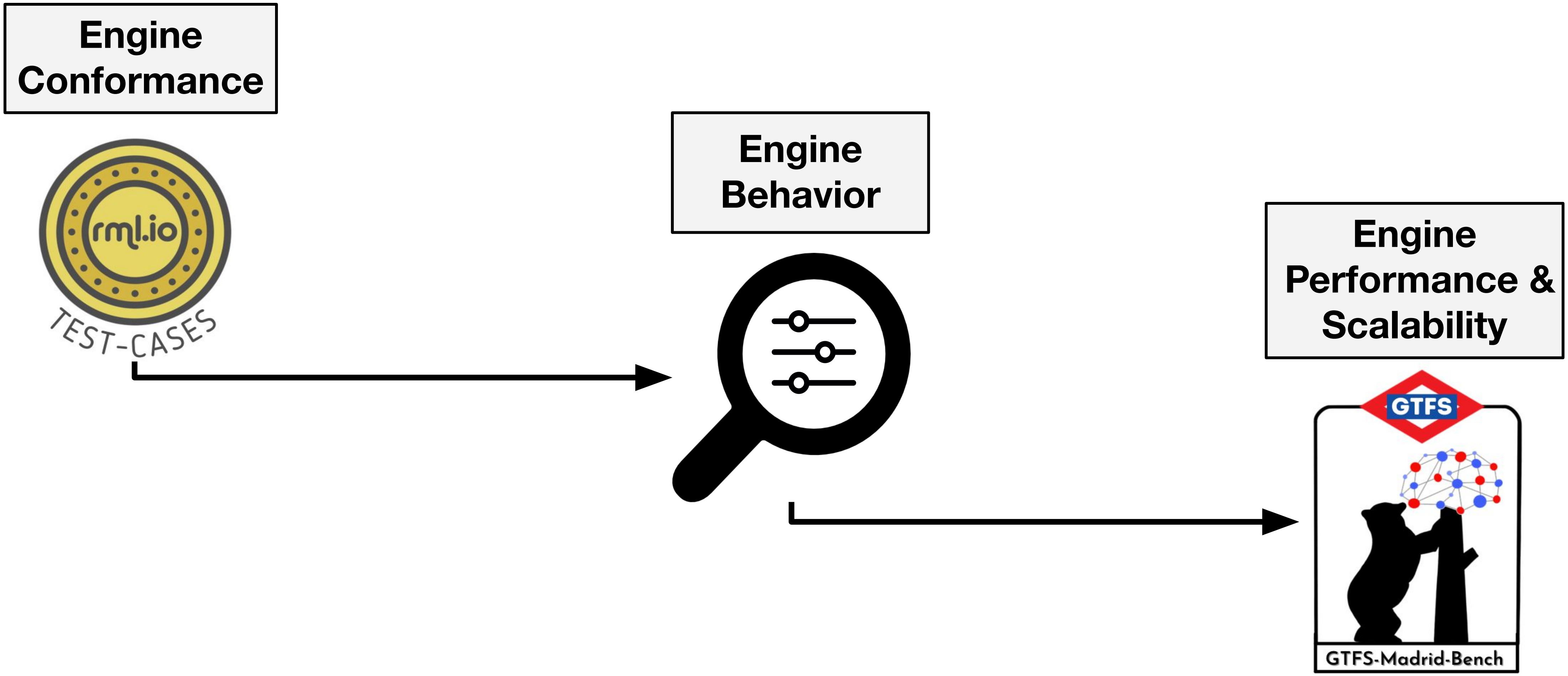
Research:

1) Evaluation framework for KGC engines
2) Optimizations over KGC:
   a) Cleansing KGC
   b) Materialized KGC (data transformation)

Applications:

1) Main author of the Spanish guideline on KGC for Open Data Portals
2) Lead researcher of “Knowledge Graph Construction for Universities” project
3) Worked on several EU and National projects in the Transport Domain

Evaluation Framework I



Engine Conformance

Engine Behavior

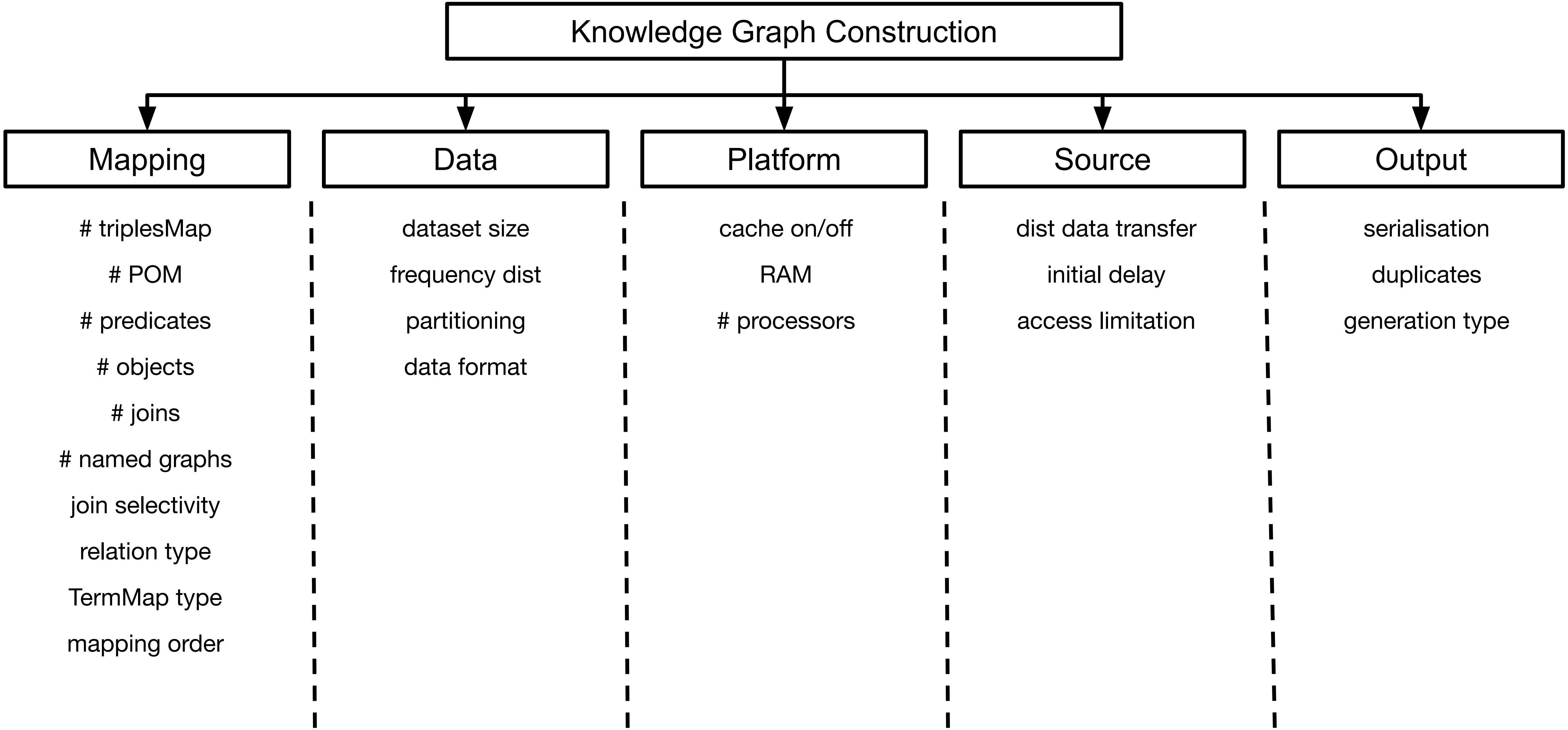
Engine Performance & Scalability

Evaluation Framework for Declarative Knowledge Graph Construction Engines from Heterogeneous Data Sources

Analysis of parameters that affect the construction of Knowledge Graphs

Chaves-Fraga, D., Endris, K. M., Iglesias, E., Corcho, O., & Vidal, M. E. (2019). What are the Parameters that Affect the Construction of a Knowledge Graph?. In *ODBASE.* This contribution is one of the result of joint collaboration with the Scientific Data Management Group from German National Library of Science and Technology (TIB), as a result of a research stay in the institution

Independent Variables



Knowledge Graph Construction

Mapping

# triplesMap

# POM

# predicates

# objects

# joins

# named graphs

join selectivity

relation type

TermMap type

mapping order

Data

dataset size

frequency dist

partitioning

data format

Platform

cache on/off

RAM

# processors

Source

dist data transfer

initial delay

access limitation

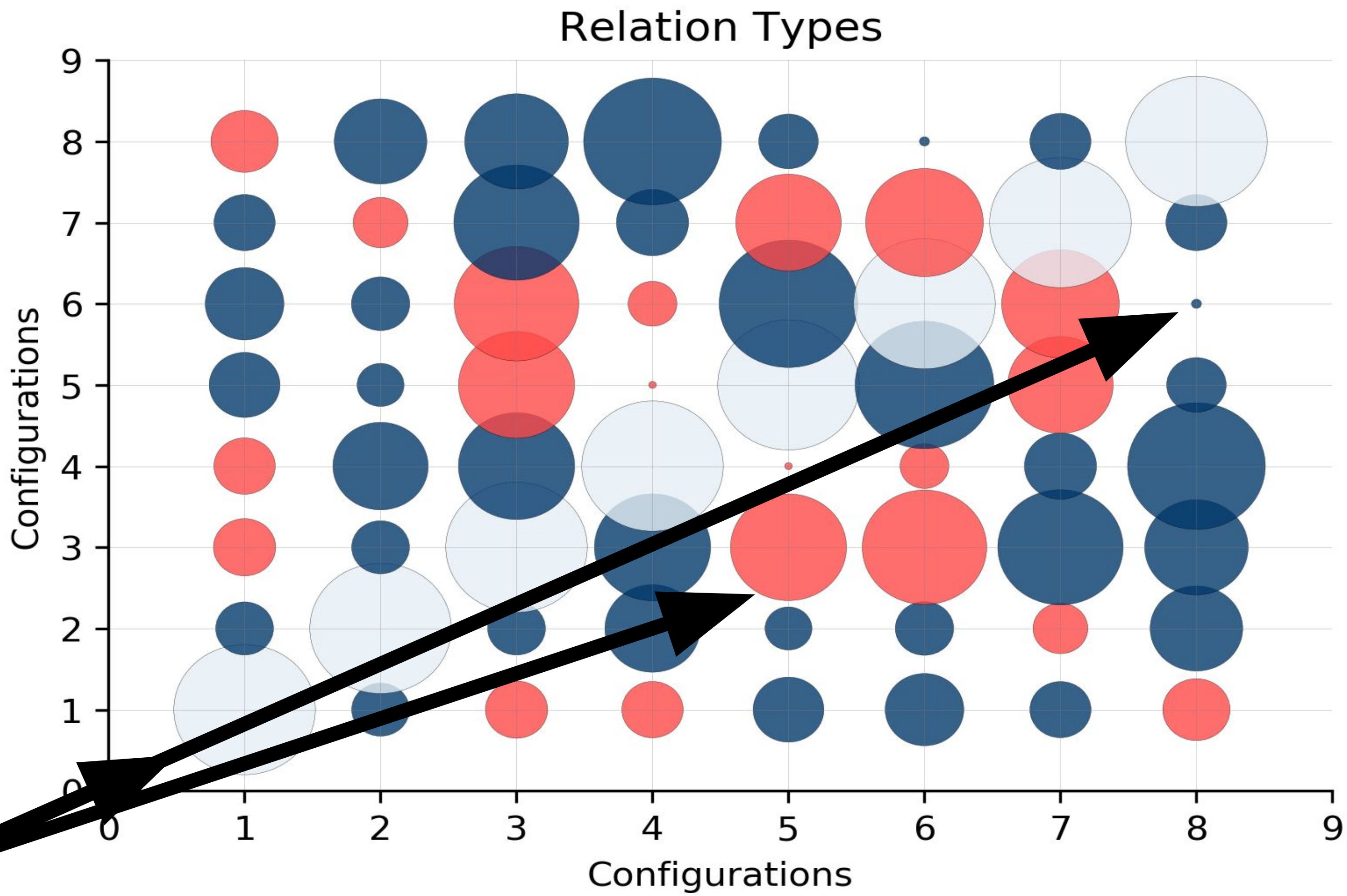
Output

serialisation

duplicates

generation type

Results for Relation Types



Comparing same configuration. Correlation ≃ 1.0

Weak positive correlation ≃ 0.1

Configurations 1-4: SDM-RDFizer on 1-N, N-1, N-M and combination

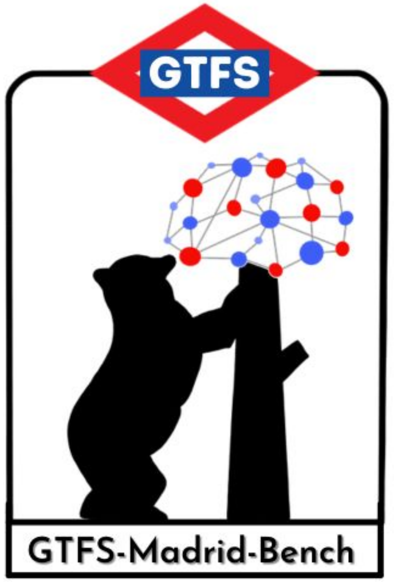
Configurations 5-8: RMLMapper on 1-N, N-1, N-M and combination

GTFS-Madrid-Bench: A Benchmark for Knowledge Graph Construction Engines

Chaves-Fraga, D., Priyatna, F., Cimmino, A., Toledo, J., Ruckhaus, E., & Corcho, O. (2020). GTFS-Madrid-Bench: A benchmark for virtual knowledge graph access in the transport domain. *Journal of Web Semantics*.

Arenas-Guerrero, J., Scrocca, M., Iglesias-Molina, A., Toledo, J., Pozo-Gilo, L., Dona, D., Corcho, O., & **Chaves-Fraga, D.** (2021). Knowledge Graph Construction with R2RML and RML: An ETL System-based Overview. *In Proceedings of the 2nd International Workshop on Knowledge Graph Construction* (ESWC).

Highlights



A comprehensive benchmark for (virtual) knowledge graph access

- Transport Domain (GTFS)
- Unified evaluation framework for heterogeneous KGC engines
- Tested over 5 virtual tools and 10 materializers tools
- Highly influenced by BSBM (queries) and NPD (data generation)

Mappings and Queries Features

Dataset:

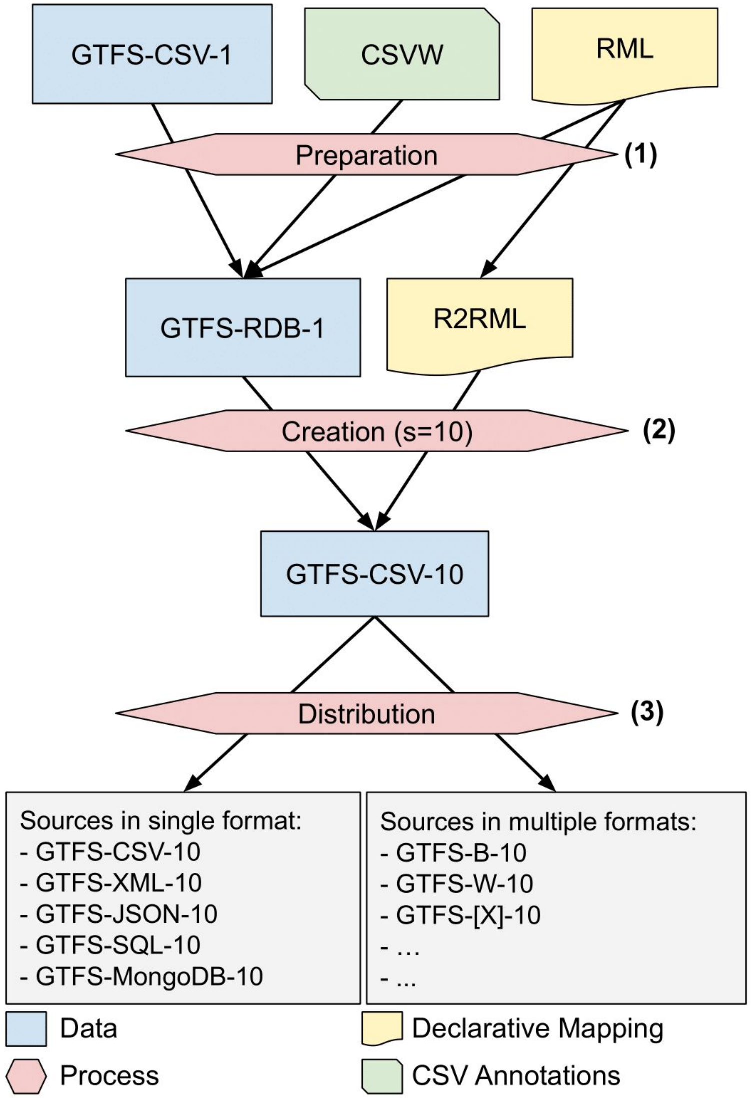
- Morph-CSV to generate GTFS-RDB
- VIG to scale-up
- Distribution based on user preferences

Queries:

- 18 queries covering different configurations and SPARQL operators
- Aligned with user stories in Madrid's transport domain
- Triple patterns: from 3 to 15; Sources: 1 to 5
- Single and chain star-shaped groups

Mappings:

- 10 Sources, 12 TriplesMap (12 Classes), 71 POM (70 P), 60 SOM, 11 ROM
- 1 R2RML, 5 RML (YARRRML serialization), 1 xR2RML, 1 CSVW annotations + RML-Mapping generator



Evaluation

TO (TimeOut), W (wrong nº results), E (error executing the query)

Dataset	Processor		Query																	
	Cache	Name	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	q11	q12	q13	q14	q15	q16	q17	q18
GTFS-SQL-1	Warm	Morph-RDB	5.85	02.07	E	1.82	W	1.86	1.97	E	26.02	1.80	E	1.81	2.06	W	1.89	E	2.11	E
	Cold	Ontario	18.02	E	TO	E	E	E	E	W	E	E	E	E	E	W	E	E	E	E
		Morph-RDB	7.14	2.65	E	2.42	W	2.36	2.43	E	28.65	2.38	E	2.41	2.69	W	2.58	E	2.68	E
		Ontop	8.37	05.04	5.18	E	W	E	W	E	16.56	E	E	E	05.06	W	5.10	W	5.00	W
GTFS-MongoDB-1	Warm	Morph-xR2RML	W	W	W	W	W	W	W	W	W	W	W	W	W	28.67	W	W	6.52	W
	Cold	Morph-xR2RML	W	W	W	W	W	W	W	W	W	W	W	W	W	28.17	W	W	6.96	W
GTFS-CSV-1	Cold	Morph-RDB	6.94	03.04	E	2.78	E	2.78	TO	E	TO	2.97	E	6.23	3.97	E	E	E	3.14	W
		Morph-CSV	15.11	10.88	E	10.72	E	9.95	10.84	E	40.90	10.70	E	11.60	11.82	E	E	E	11.48	W
		Ontario	W	E	17.34	E	E	E	E	W	E	E	E	E	E	W	E	E	E	E
GTFS-XML-1	Cold	Ontario	E	E	E	E	E	E	E	E	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	E	E	E	W	E	E	E	E
GTFS-MINEXT-1	Cold	Ontario	W	E	E	E	E	E	E	W	E	E	E	E	E	W	E	E	E	E
GTFS-MAXEXT-1	Cold	Ontario	W	E	17.14	E	E	E	E	W	E	E	E	E	E	W	E	E	E	E

- Only the SPARQL-to-SQL engines provide an acceptable support for SPARQL operators
- Virtual KGC proposals beyond relational databases are not mature enough and more research is needed
- The problem of translating SPARQL queries for querying raw data (CSV, JSON, XML) should not be understood as a technical case

morph

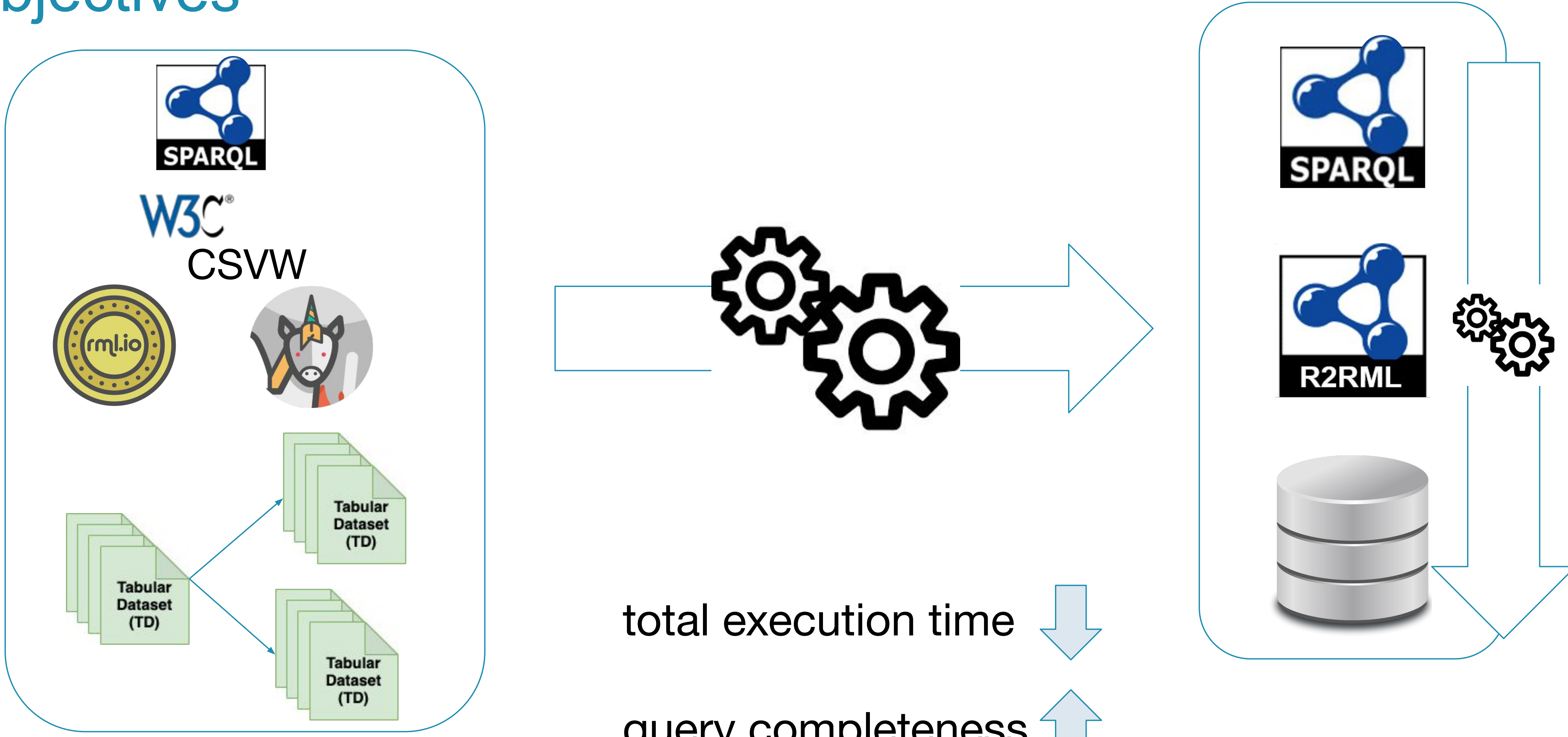
> *"The terms data cleaning and data pre-processing should be removed"*
>
> An Anonymous Data Engineer

Virtual Knowledge Graph Construction over Tabular Data

Chaves-Fraga, D., Ruckhaus, E., Priyatna, F., Vidal, M. E., & Corcho, O. (2021). Enhancing Virtual Ontology Based Access over Tabular Data with Morph-CSV. *Semantic Web*.

Chaves-Fraga, D., Pozo-Gilo, L., Toledo, J., Ruckhaus, E., & Corcho, O. (2020). Morph-CSV: Virtual Knowledge Graph Access for Tabular Data. In *International Semantic Web Conference (P&D)*.

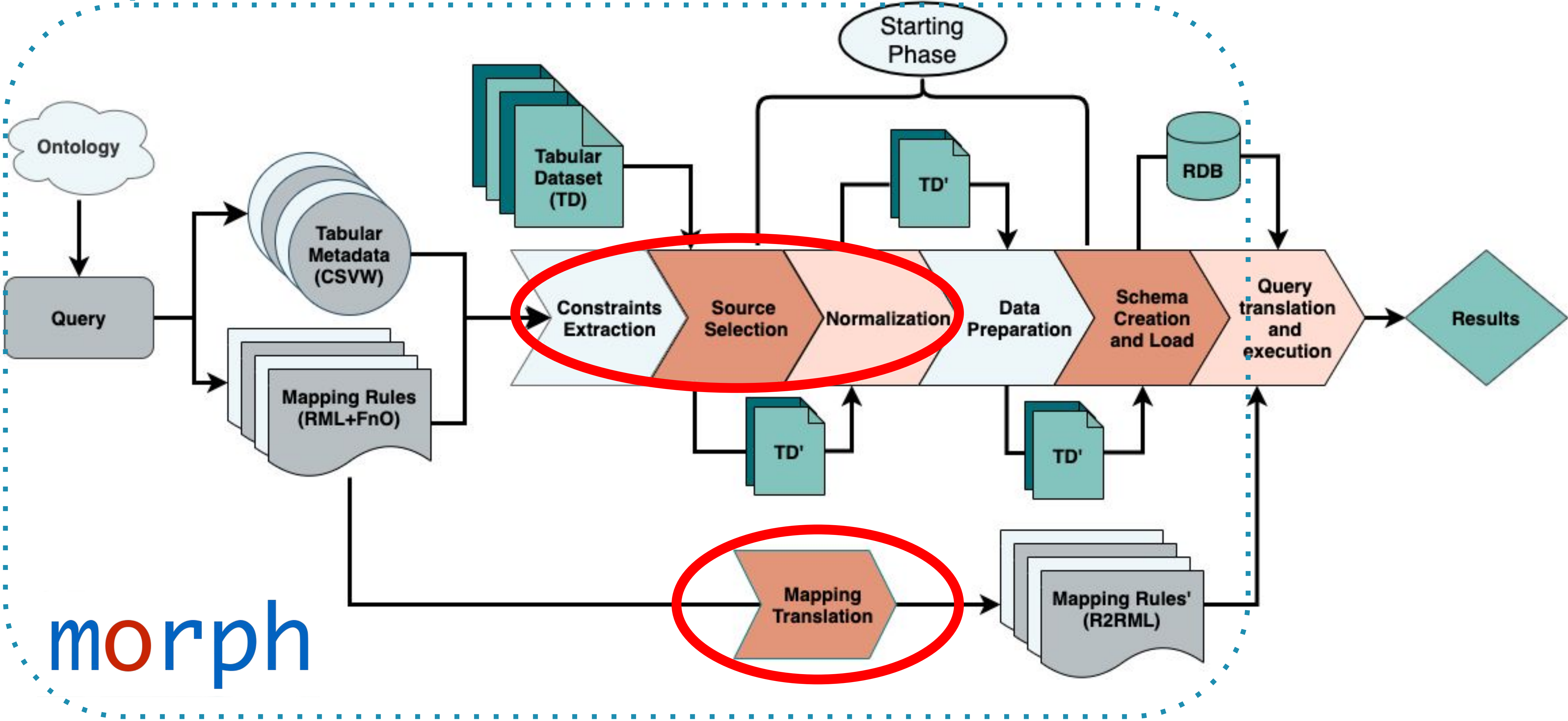
Objectives



total execution time ⇩

query completeness ⇧

Morph-CSV



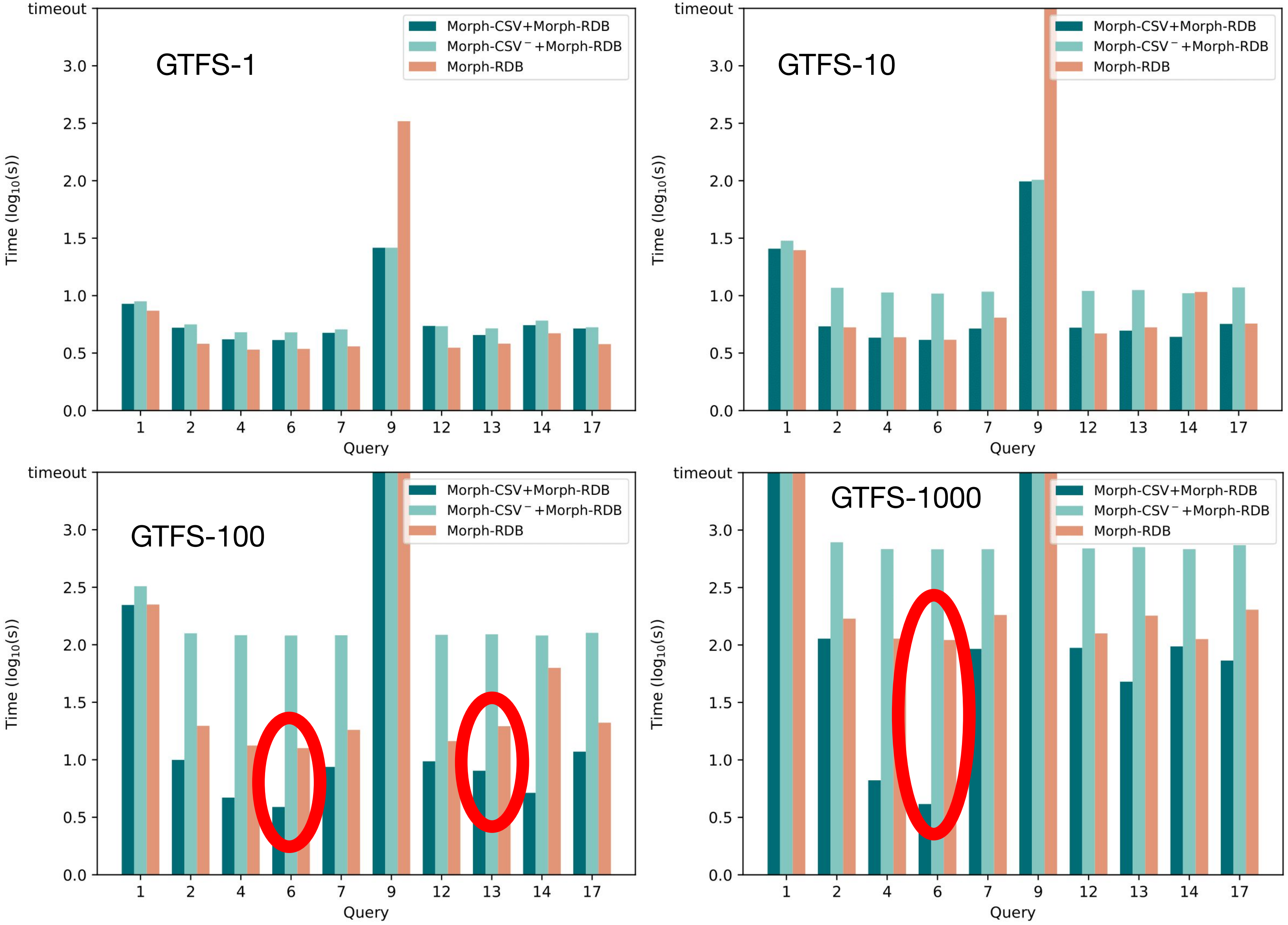
KU LEUVEN

GTFS-Madrid-Bench

Baseline:
Morph-RDB

Approach 1 (complete RDB):
Morph-CSV^{-} + Morph-RDB

Approach 2 (source selection):
Morph-CSV + Morph-RDB

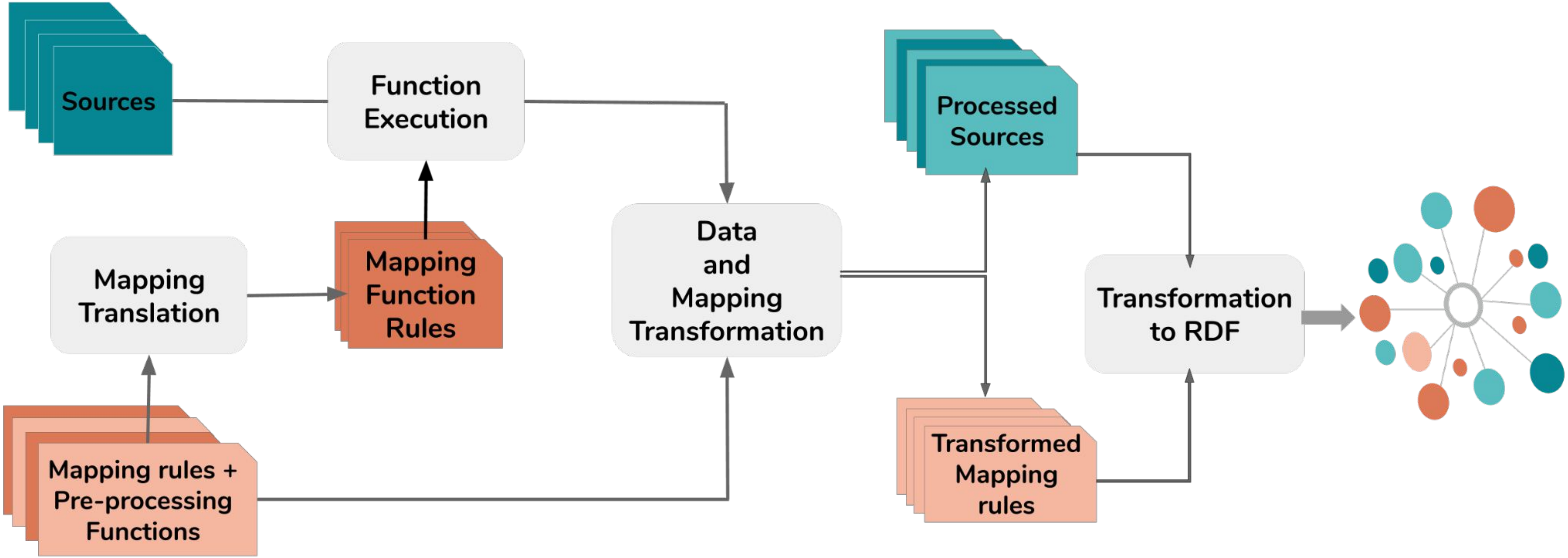


FunMap

FunMap: Efficient Execution of Functional Mappings for Scaled-Up Knowledge Graph Creation

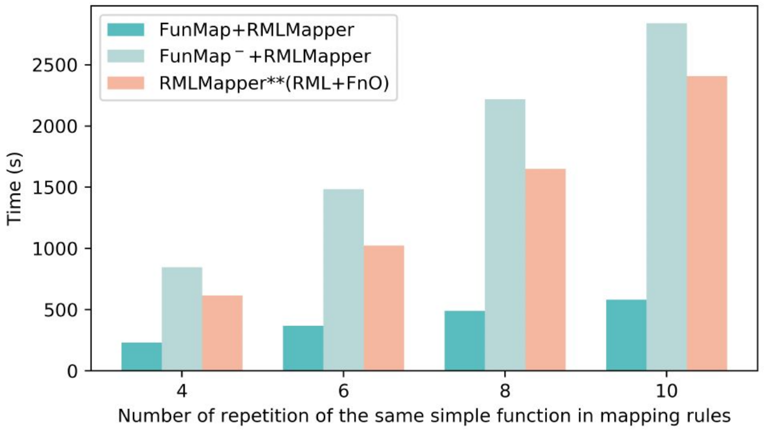
Jozashoori, S., **Chaves-Fraga, D**., Iglesias, E., Vidal, M. E., & Corcho, O. (2020, November). FunMap: Efficient Execution of Functional Mappings for Knowledge Graph Creation. In *International Semantic Web Conference* (Core A). First two authors contributed equally to the research. Fully reproduced paper

FunMap

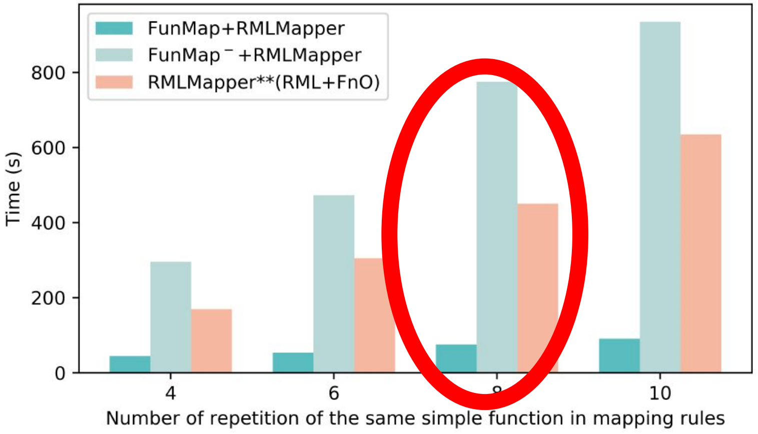

Sources
Function Execution
Processed Sources
Mapping Translation
Mapping Function Rules
Data and Mapping Transformation
Transformation to RDF
Mapping rules + Pre-processing Functions
Transformed Mapping rules

Experimental results

Simple functions (lower, upper)

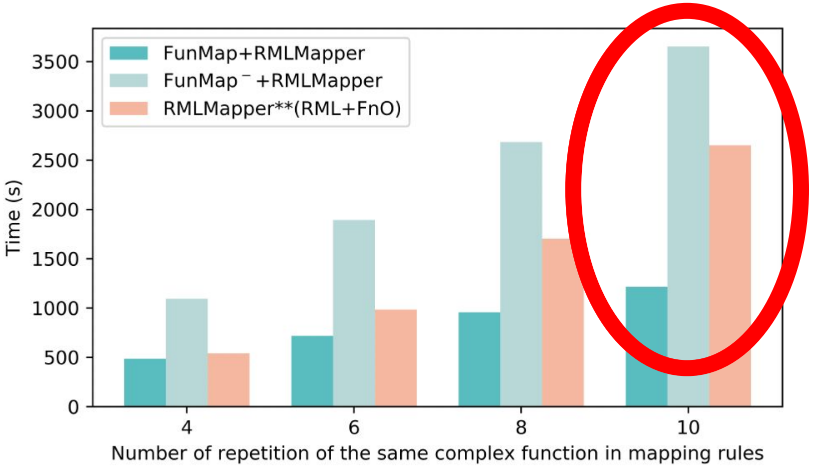


(c) RMLMapper - 25% of duplicates

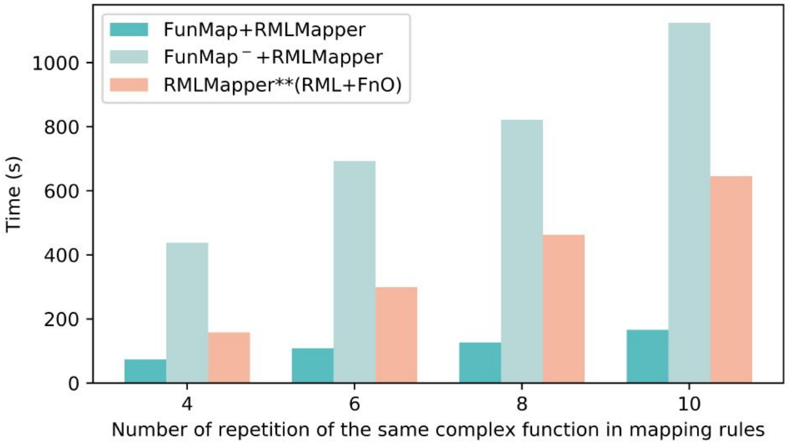


(d) RMLMapper - 75% of duplicates

Complex functions (if, replace, multiple columns)



(c) RMLMapper - 25% of duplicates



(d) RMLMapper - 75% of duplicates

Other relevant contributions…

Journal

Corcho, O., Priyatna, F., & **Chaves-Fraga, D.** (2020). Towards a new generation of ontology based data access. In *Semantic Web Journal*

Chaves-Fraga, D., Priyatna, F., Alobaid, A., & Corcho, O. (2020). Exploiting Declarative Mapping Rules for Generating GraphQL Servers with Morph-GraphQL. *IJSEKE*

Goncalves, M., **Chaves-Fraga, D.**, & Corcho, O. (2021). Handling Qualitative Preferences in SPARQL over Virtual Ontology-Based Data Access. In *Semantic Web Journal.*

Arenas-Guerrero, J., **Chaves-Fraga, D.**, Toledo, J., Pérez, M. S., & Corcho, O. (2022). Morph-KGC: Scalable Knowledge Graph Materialization with Mapping Partitions. In *Semantic Web Journal (Under Review)*.

Iglesias-Molina A., Cimmino A., Ruckhaus E., **Chaves-Fraga D.**, García-Castro R., & Corcho O. (2022). Towards an Ontological Approach for Integrating Declarative Mapping Languages. In *Semantic Web Journal (Under Review)*.

Conference

Iglesias, E., Jozashoori, S., **Chaves-Fraga, D.**, Collarana, D., & Vidal, M. E. (2020). SDM-RDFizer: An RML interpreter for the efficient creation of RDF knowledge graphs. In *Proceedings of the 29th ACM CIKM*

Corcho, O., **Chaves-Fraga, D.**, et al (2021). A High-Level Ontology Network for ICT Infrastructures. In *International Semantic Web Conference (Resource Track).*

Heyvaert, P., **Chaves-Fraga, D.**, Priyatna, F., Corcho, O., Mannens, E., Verborgh, R., & Dimou, A. (2019). Conformance test cases for the RDF mapping language (RML). In *Iberoamerican KGSWC*

Goncalves, M., **Chaves-Fraga, D.**, & Corcho, O. (2020). Morph-Skyline: Virtual Ontology-Based Data Access for Skyline Queries. In *International Joint Conference On Web Intelligence And Intelligent Agent Technology (WI-IAT'20)*

Iglesias-Molina, A., **Chaves-Fraga, D**., Priyatna, F., & Corcho, O. (2019). 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*.

Others

Rojas, J., **Chaves-Fraga, D.**, Colpaert, P., Verborgh, R., & Mannens, E. (2017). Providing Reliable Access to Real-Time and Historic Public Transport Data Using Linked Connections. In *International Semantic Web Conference (Posters, Demos & Industry Tracks)*.

Chaves-Fraga, D., Antón, A., Toledo, J., & Corcho, O. (2019). ONETT: Systematic Knowledge Graph Generation for National Access Points. In *1st International Workshop on Semantics for Transport (SEMANTICS)*.

Chaves-Fraga, D., Rojas, J., Vandenberghe, P. J., Colpaert, P., & Corcho, O. (2017). The tripscore Linked Data client: calculating specific summaries over large time series. *In Proceedings of the 1st International Workshop on Decentralizing the Semantic Web (ISWC).*

Iglesias-Molina, A., **Chaves-Fraga, D.**, Priyatna, F., & Corcho, O. (2019). 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 (K-CAP 2019)*.

Badenes-Olmedo, C., **Chaves-Fraga, D.**, et al. (2020). Drugs4Covid: Drug-driven Knowledge Exploitation based on Scientific Publications. *arXiv preprint arXiv:2012.01953*

Next (or Current) Steps at KU Leuven

Competitive Margarita Salas Fellowship (granted by UPM)

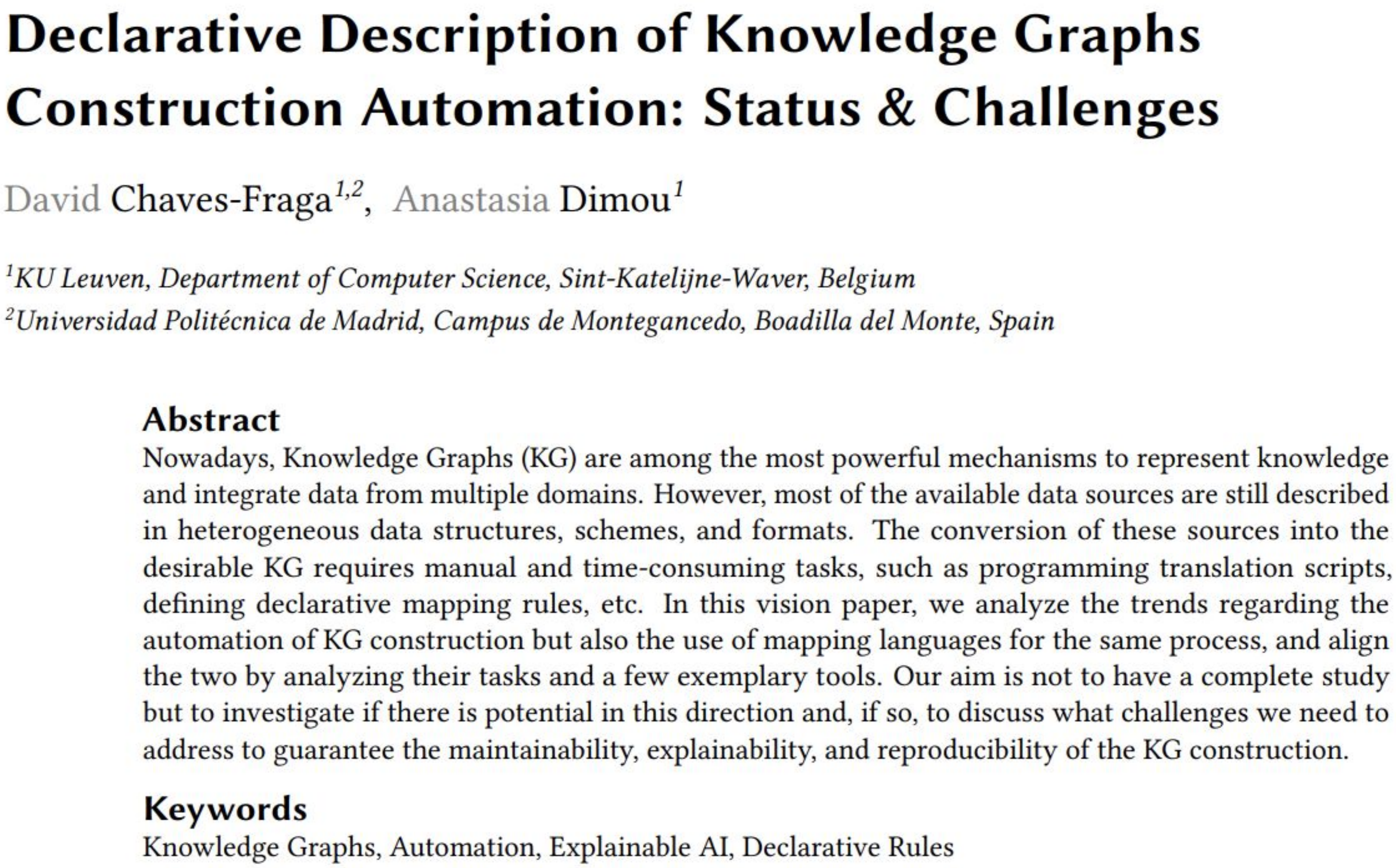
- 2 years at KU Leuven (2022-2023) + 1 year coming back to OEG (2024)

DecentralizedData4All:

- Rejected as Marie-Curie proposal :-(, but good reviews (82.4/100)
- **Next generation of KGC engines**
  a) Automation of rules generation (semantic tabular annotation)
  b) Automation of data cleansing pipelines (learning + efficient)
  c) Explainable and reproducible workflows for data transformation
  d) RML standardization through W3C
  e) RDF-star generation from RML-star mappings

Any other potential collaboration? :-) :-) → **david.chaves@kuleuven.be**

First contribution at KGCW2022 (ESWC)



Declarative Description of Knowledge Graphs Construction Automation: Status & Challenges

David Chaves-Fraga[1,2], Anastasia Dimou[1]

[1] *KU Leuven, Department of Computer Science, Sint-Katelijne-Waver, Belgium*

[2] *Universidad Politécnica de Madrid, Campus de Montegancedo, Boadilla del Monte, Spain*

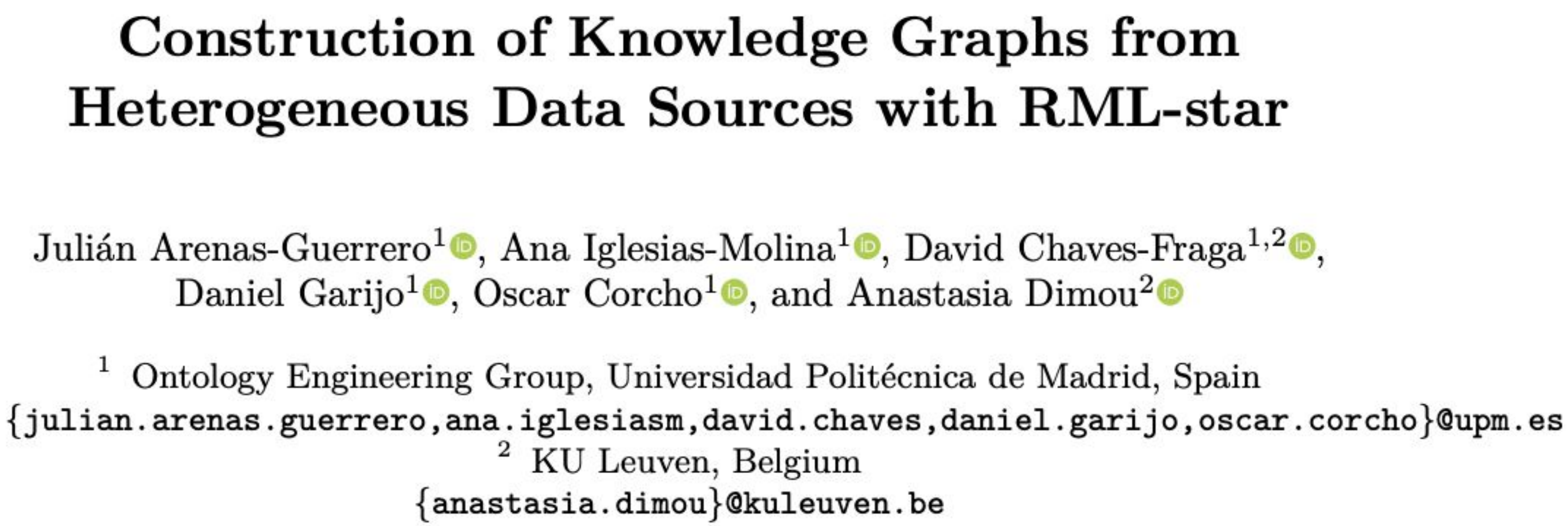
Abstract

Nowadays, Knowledge Graphs (KG) are among the most powerful mechanisms to represent knowledge and integrate data from multiple domains. However, most of the available data sources are still described in heterogeneous data structures, schemes, and formats. The conversion of these sources into the desirable KG requires manual and time-consuming tasks, such as programming translation scripts, defining declarative mapping rules, etc. In this vision paper, we analyze the trends regarding the automation of KG construction but also the use of mapping languages for the same process, and align the two by analyzing their tasks and a few exemplary tools. Our aim is not to have a complete study but to investigate if there is potential in this direction and, if so, to discuss what challenges we need to address to guarantee the maintainability, explainability, and reproducibility of the KG construction.

Keywords

Knowledge Graphs, Automation, Explainable AI, Declarative Rules

Second contribution at ISWC (to be submitted)



Construction of Knowledge Graphs from Heterogeneous Data Sources with RML-star

Julián Arenas-Guerrero[1], Ana Iglesias-Molina[1], David Chaves-Fraga[1,2], Daniel Garijo[1], Oscar Corcho[1], and Anastasia Dimou[2]

[1] Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
{julian.arenas.guerrero,ana.iglesiasm,david.chaves,daniel.garijo,oscar.corcho}@upm.es
[2] KU Leuven, Belgium
{anastasia.dimou}@kuleuven.be

KU LEUVEN

Thanks! More info at: http://davidchavesfraga.com/

Dr. David Chaves-Fraga

✉ david.chaves@kuleuven.be

@dchavesf