KU LEUVEN

Ontology Engineering Group

POLITÉCNICA

# Declarative Construction of Knowledge Graphs: one-year status-update

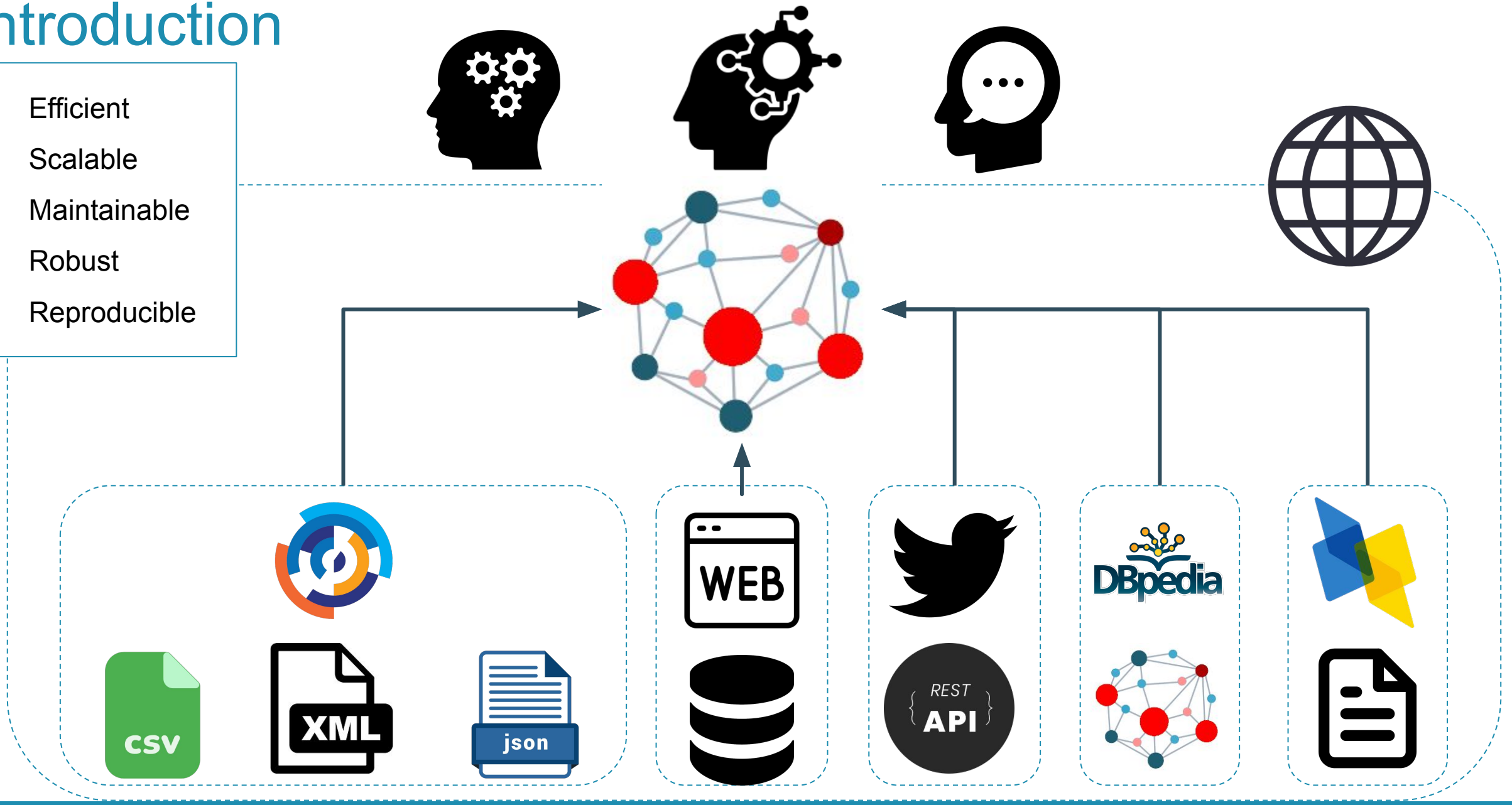
David Chaves-Fraga

david.chaves@kuleuven.be

@dchavesf

# Introduction

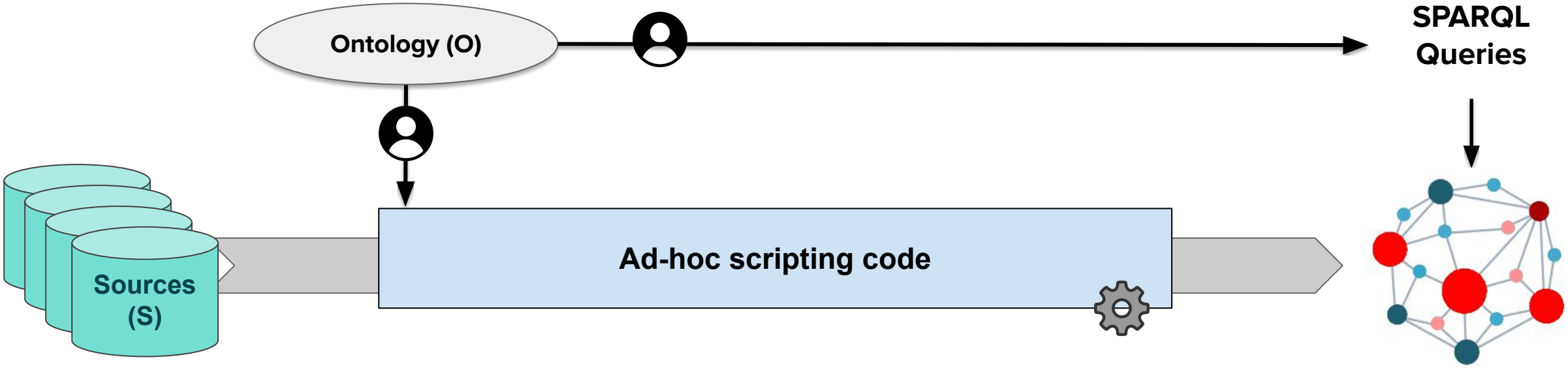
- Efficient
- Scalable
- Maintainable
- Robust
- Reproducible






KU LEUVEN

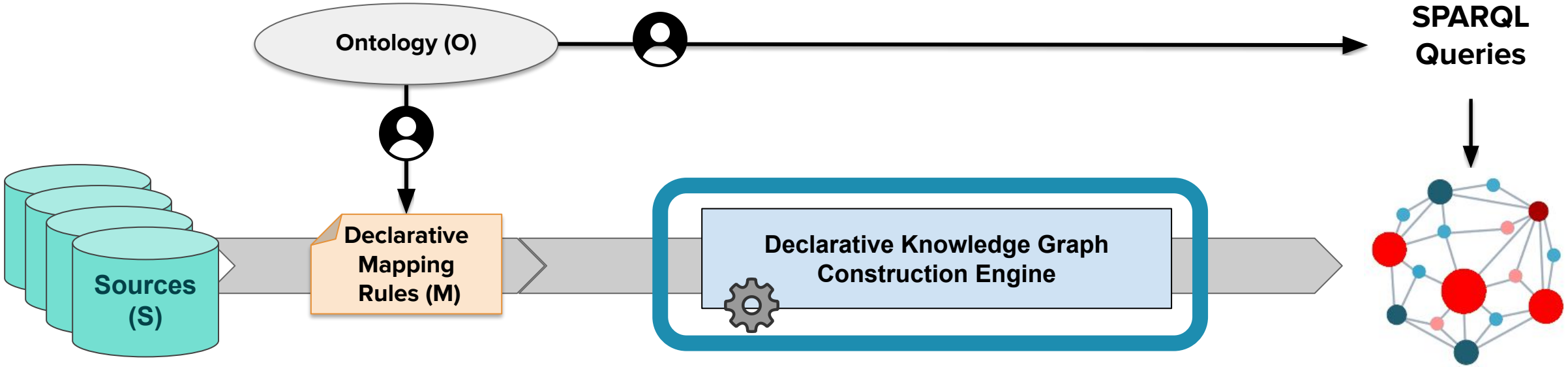
# Knowledge Graph Construction: Scripting-based




Efficient 🤷 Scalable 🤷 Maintainable 🙆 Robust 🙆 Reproducible 🙆

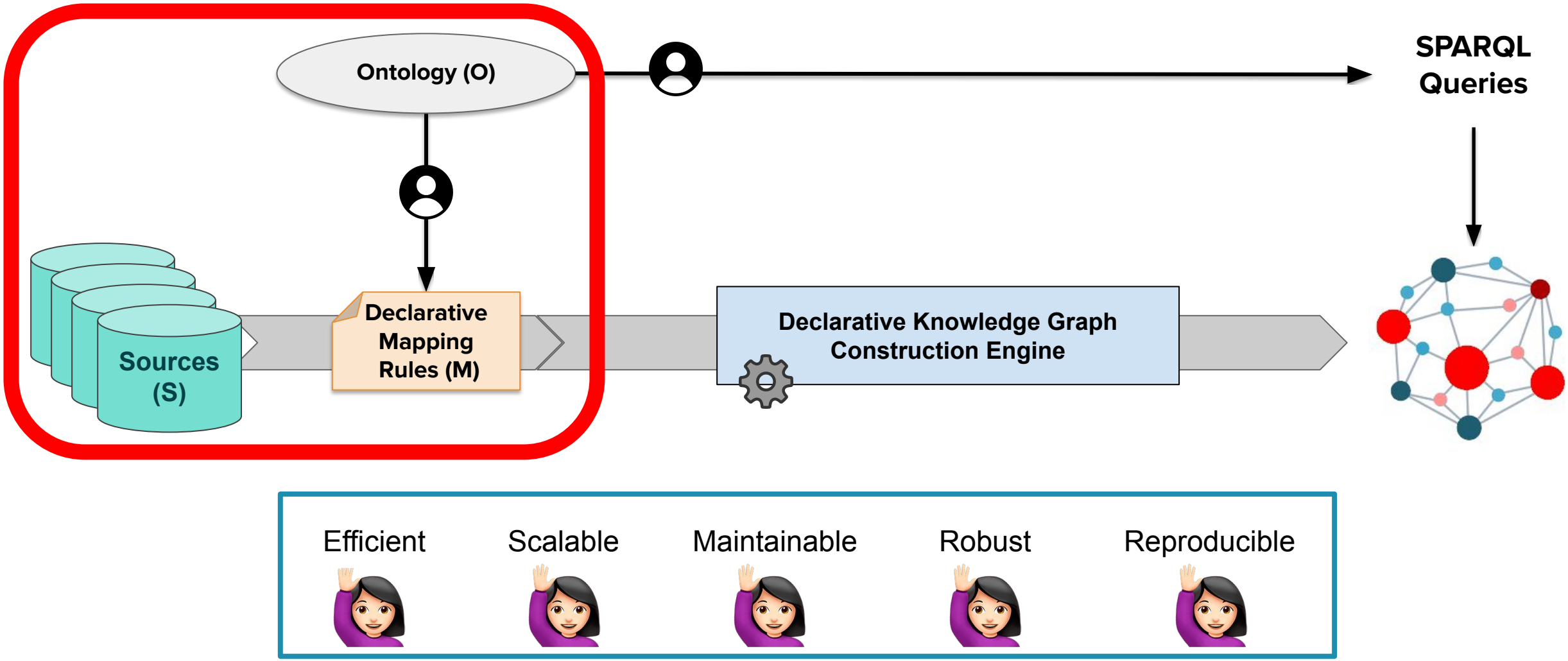
# Knowledge Graph Construction

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




Efficient Scalable Maintainable Robust Reproducible

# KG Construction with Mapping Rules



# Knowledge Graph Construction: Scenarios

Partner/company wants to create a KG with its own data…

**Scenario 1**

No vocab/ontology
Domain expert w/o technical skills

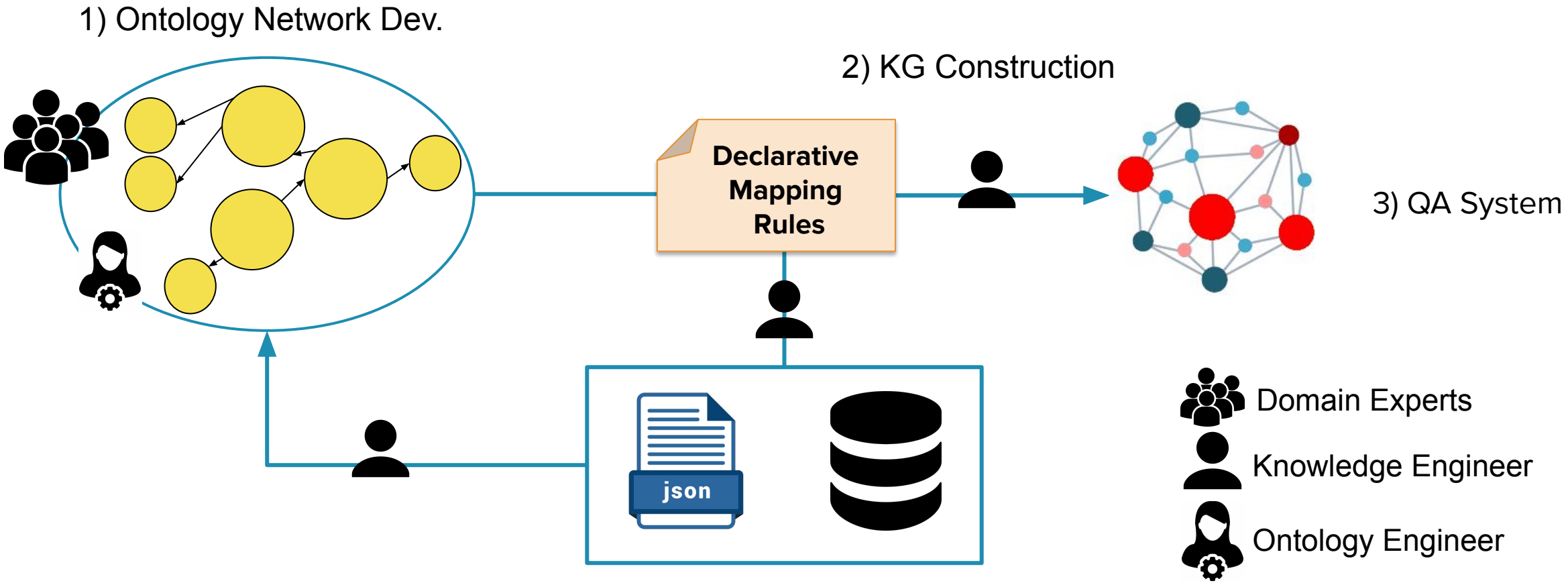
**Scenario 2**

Vocab/ontology defined (e.g., standard)
Domain expert with technical skills

**Scenario 3**

Target KG (e.g., DBpedia)
Reduce domain experts effort

# S1: Everything under your control (more or less)




Corcho, O., Chaves-Fraga, D., et al. (2021). A High-Level Ontology Network for ICT Infrastructures. In *International Semantic Web Conference* (pp. 446-462)

# S1: Everything under your control (more or less)

Outcomes:

- Development of the ontology network (~ 6 months)
- Mapping templates with OWL2YARRRML (automatic)
- Mapping rule creation in YARRRML (~1 month)
- Complex environment for testing/development of a KGC engine
  - Morph-KGC* (https://github.com/oeg-upm/morph-kgc)

Lessons Learned:

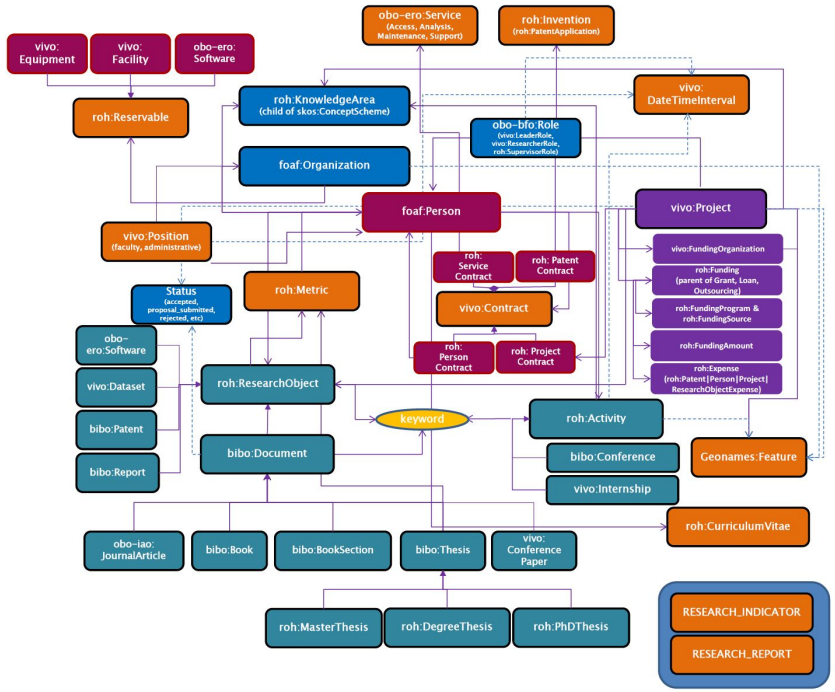
- Simple but useful support tools (OWL2YARRML)
- Domain experts w/o technical skills → ontology conceptualization
- Independent maintainability difficult to guarantee
- Ontology is stable → mapping rules key resource for the KGC

* Arenas-Guerrero, J., Chaves-Fraga, D., Toledo, J., Pérez, M. S., & Corcho, O. (2022). Morph-KGC: Scalable knowledge graph materialization with mapping partitions. *Semantic Web Journal (accepted)*.

# S2: Nothing under your control

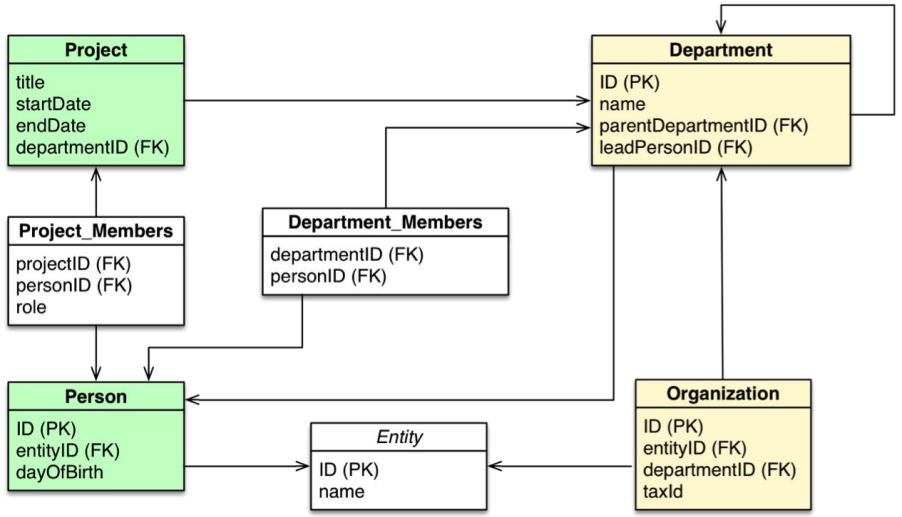
HERCULES
Semántica de Datos de Investigación de Universidades

crue Universidades Españolas




- Ontology v0.3 (will change)
- Not standard documentation (PDF file)

ORACLE



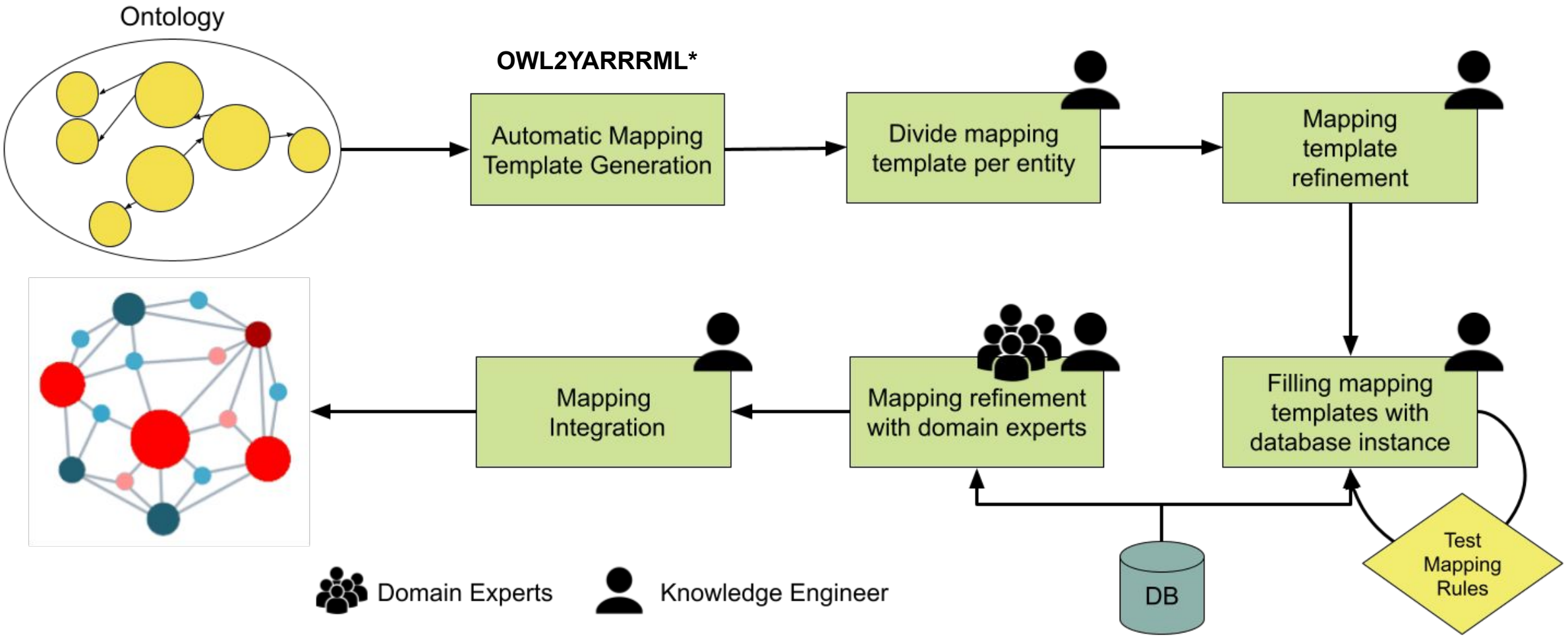

- More than 1800 tables
- Database very well documented
- Oracle supports R2RML mappings

Chaves-Fraga, D., Corcho, O., et al. (2022). Systematic Construction of Knowledge Graphs for Research Performing Organizations in Spain (Under Review)



KU LEUVEN

# S2: Nothing under your control




* https://github.com/oeg-upm/owl2yarrrml

# S2: Nothing under your control

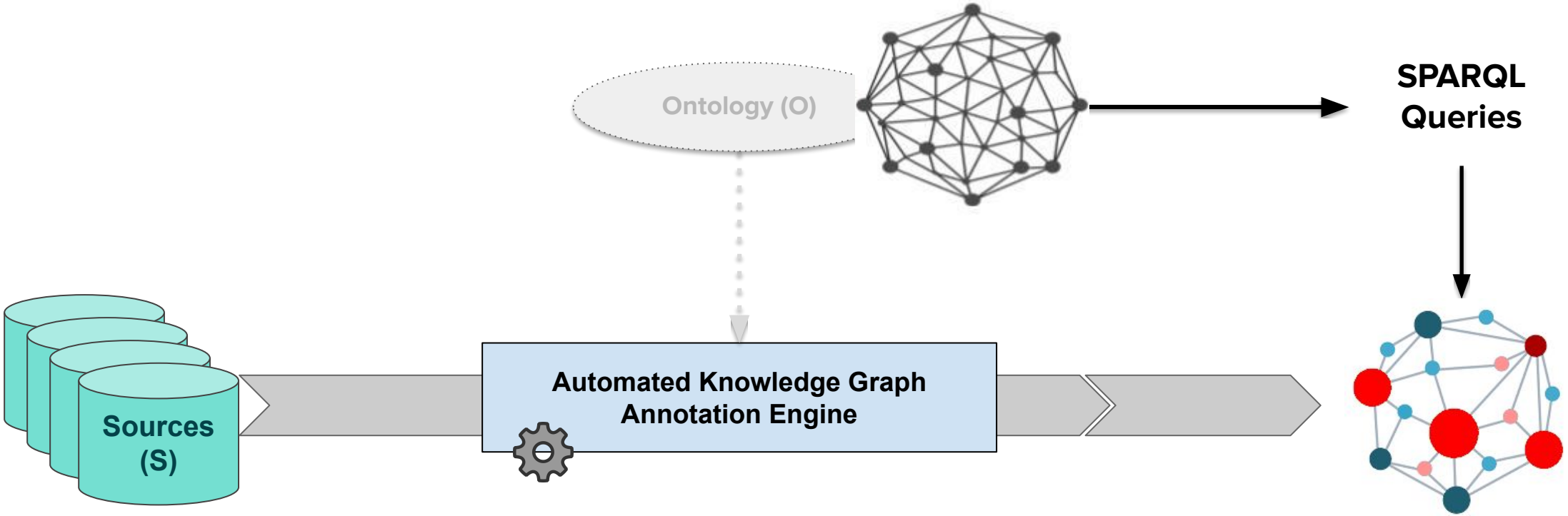
Outcomes:

- Total time: 7 months for mapping creation
- More than 5K rules in R2RML (N-Triples syntax)
- Virtual KG over each entity
- Materialized KG to feed a central repository

Lessons Learned:

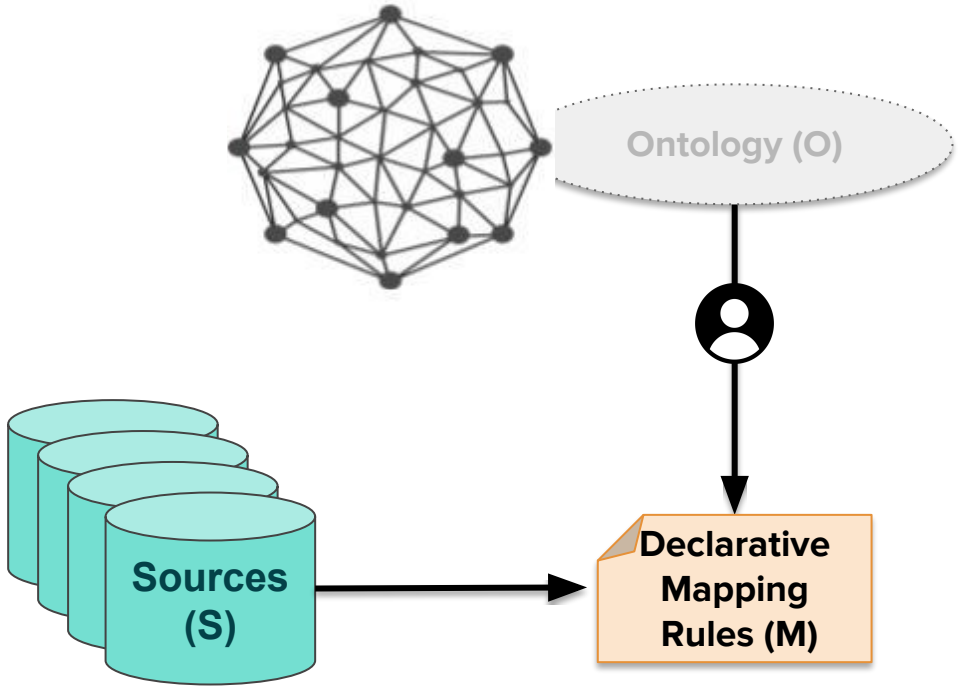
- Simple but useful support tools (OWL2YARRML)
- Domain experts with technical knowledge in the loop
- Divide and Conquer in complex scenarios
- Delegate complex tasks to the DBMS

# S3: Automation KG Construction




Chaves-Fraga, D. & Dimou, A. (2022). Declarative Description of Knowledge Graphs Construction Automation: Status & Challenges. In *Third International Workshop on Knowledge Graph Construction co-located with ESWC2022.*

# S3: Automation KG Construction




**Rules**

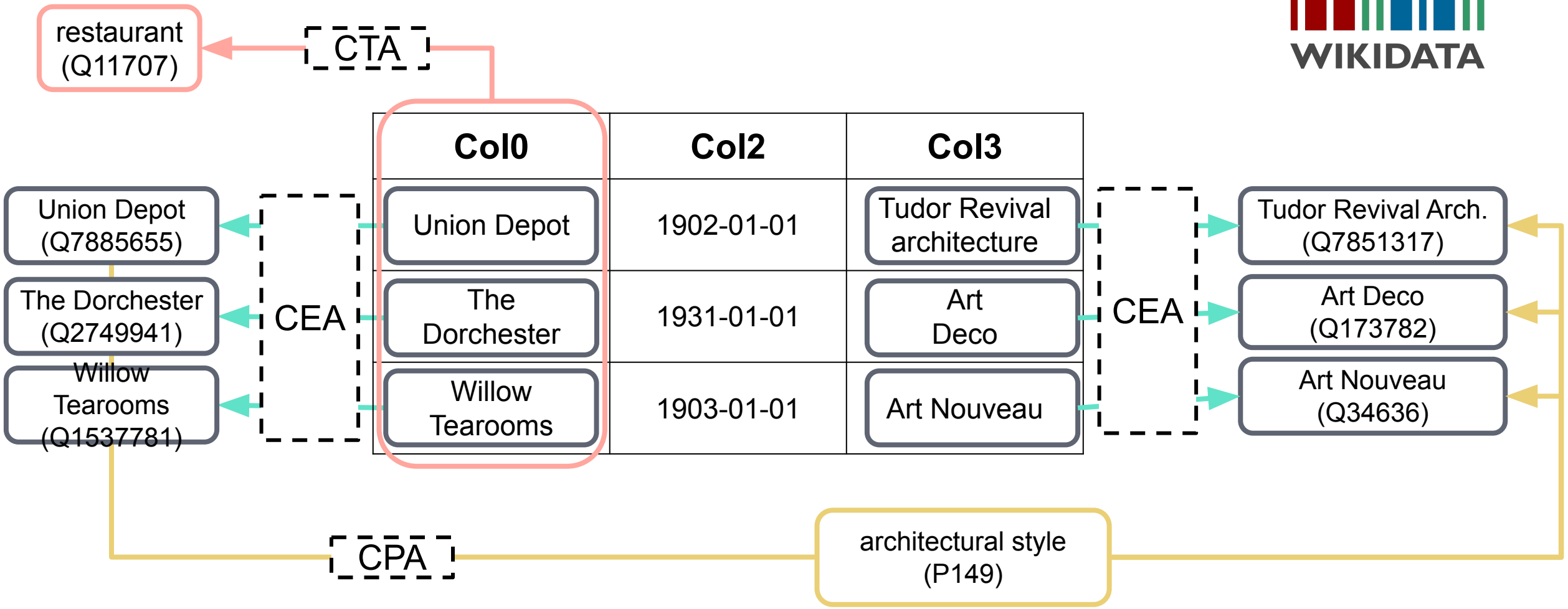
- Declarative approach
- Understanding the domain
- Target KG
- Linear iteration
- Time consuming
- High quality KG
- Non-reproducible task
- Explainable

**Automation**

- No manual work
- No knowledge about the domain
- Target Annotation
- Multiple iterations
- Faster
- Quality can be compromised
- Reproducible tasks?
- Non-explainable

RQ1) Are hybrid approaches feasible to explain and optimize a knowledge graph construction process?
RQ2) Can we describe a knowledge graph construction automation process using declarative rules?

# Automation KGC: SemTab Challenge



# Analysis of current SemTab tools

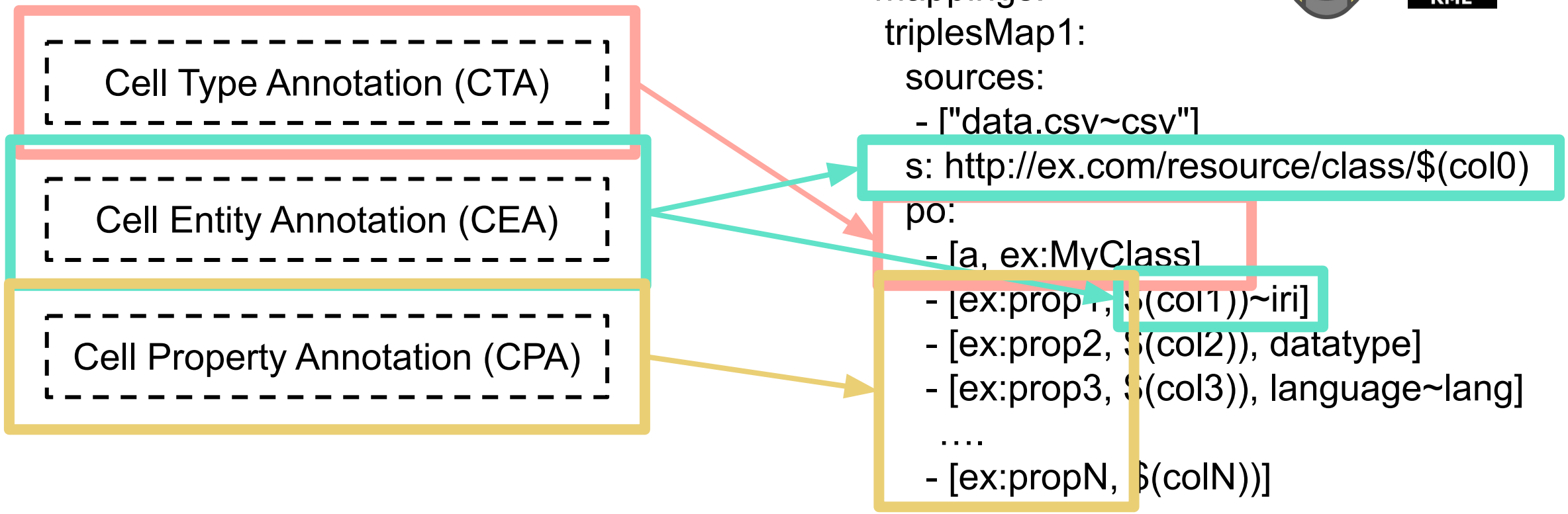
**We tried to compare the SemTab annotators…**

**Open Source Tools:** JenTab, MTab and Mantis V

**Outcomes:**

- Similar steps (e.g., KGs lookup, preprocessing, datatype prediction)
- Common procedures (e.g., majority vote/levenshtein distance)
- Blackboxes/Not explainable
- Iterative process

# SemTab and RML



# More questions than answers

1) What happened to the **RDB2RDF automation** approaches (e.g., MIRROR, AutoMap4OBDA)? **adapt/extend them to this new generation**?

2) Should we extend current mapping languages to describe more complex tasks beyond triples generation?

3) Should we use declarative description of functions to enhance the transparency & explainability of current SemTab solutions?

4) Are declarative mapping languages the ideal way of representing automation despite the difference among paradigms?

# Conclusions and Final Remarks

- Mapping rules (in any form) are the central resource of KG generation
- Background of domain experts / users have to be considered
- Adaptability means successful
- Trade-offs: Automation VS Data quality
- Governance of Data Integration Systems (Sources, Mappings, Ontology)

Do you want to know more/get involved? http://w3id.org/kg-construct

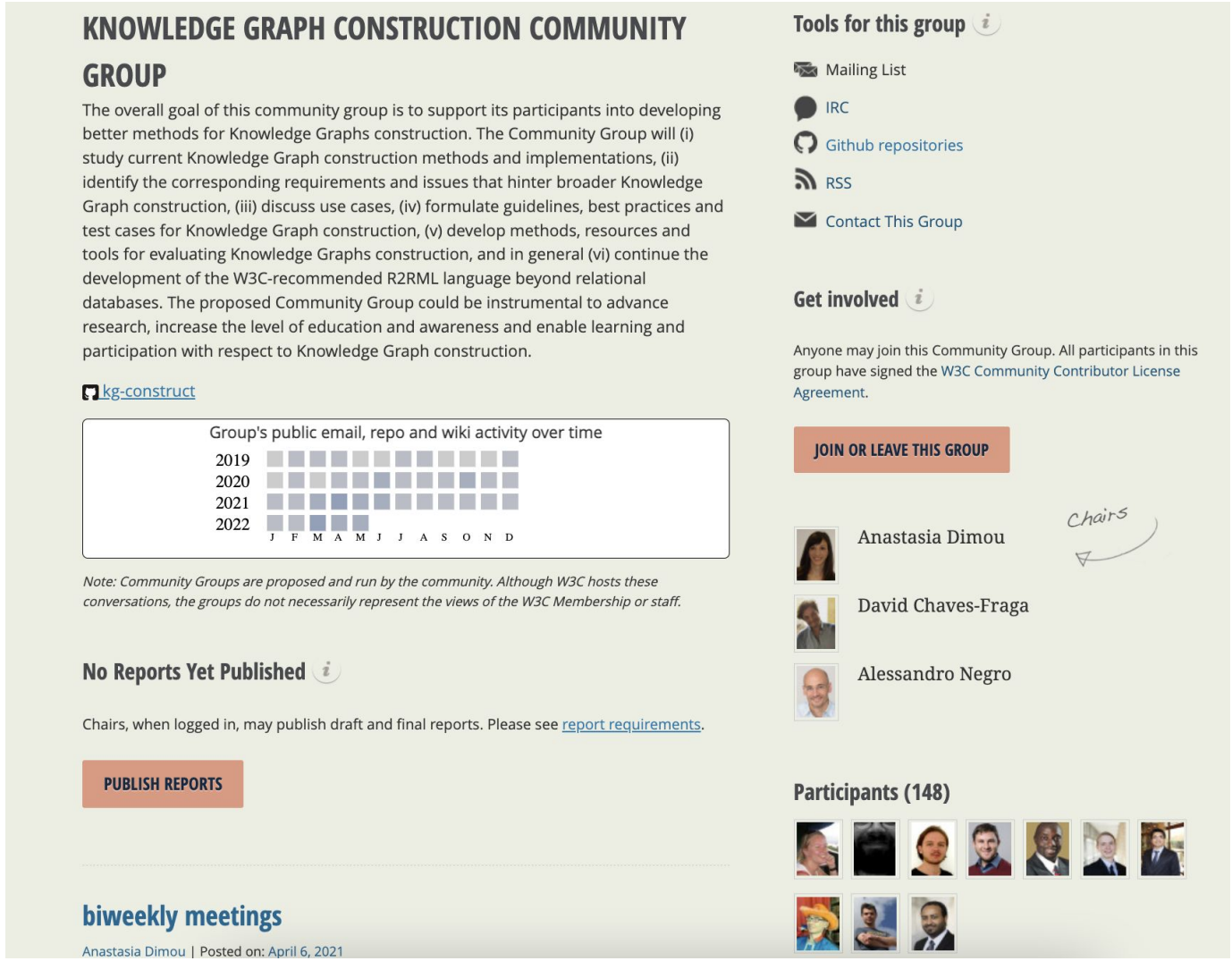
Awesome KGC tools: https://github.com/kg-construct/awesome-kgc-tools

# W3C Community Group Knowledge Graph Construction

http://w3id.org/kg-construct

KU LEUVEN

# W3C Community Group - Knowledge Graph Construction



## 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.

kg-construct

Group's public email, repo and wiki activity over time

2019
2020
2021
2022

J F M A M J J A S O N D

*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

### biweekly meetings

Anastasia Dimou | Posted on: April 6, 2021

### Tools for this group

Mailing List
IRC
Github repositories
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
David Chaves-Fraga
Alessandro Negro

### Participants (148)

148 participants (~25 active)

Bi-weekly meetings

RML standardization

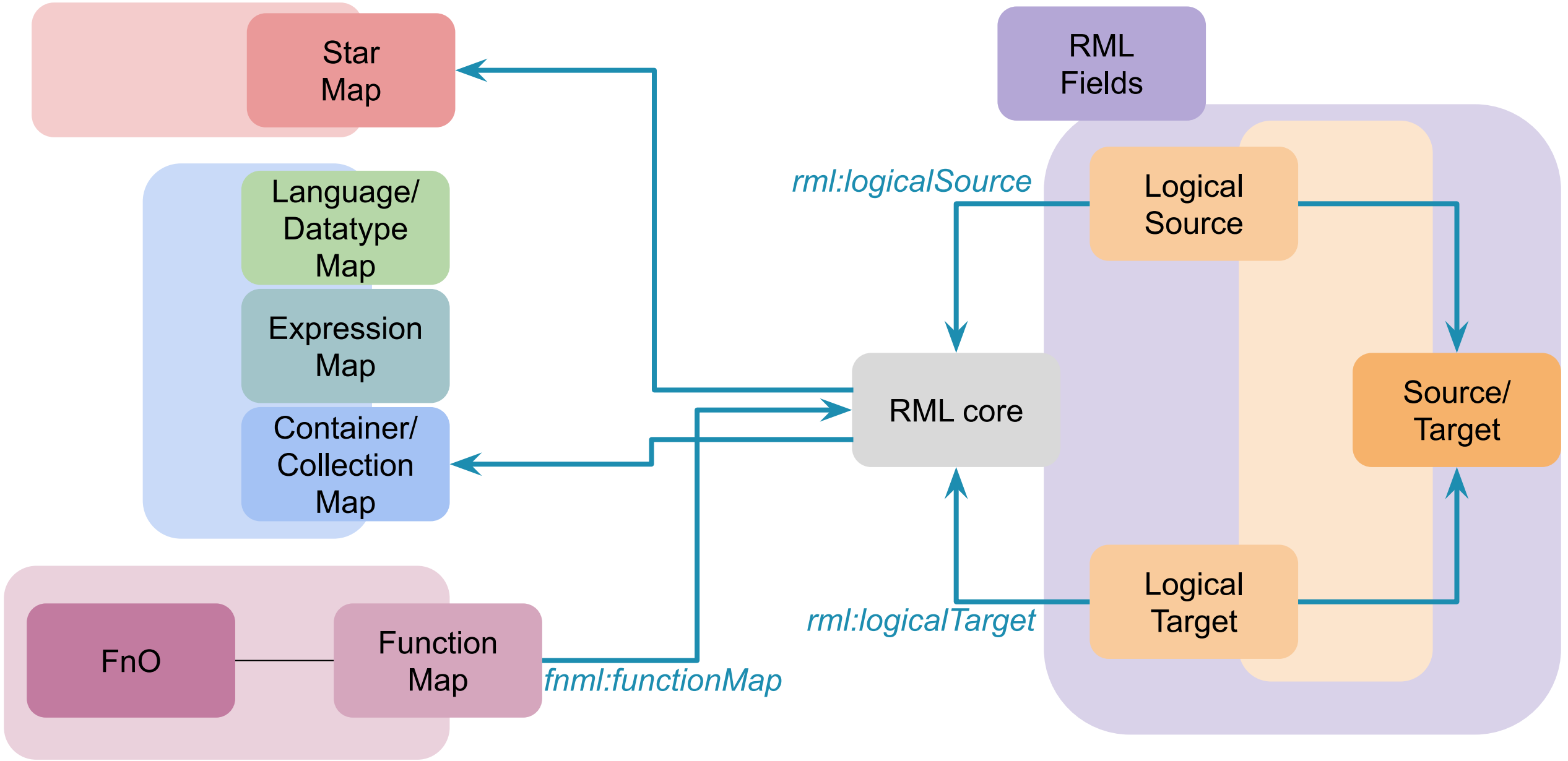
6 on-going specs:

RML-star, RML-functions, RML-fields, RML-lists, RML-IO, RML-expressions

https://w3id.org/kg-construct

http://github.com/kg-construct

KU LEUVEN


Star Map
Language/ Datatype Map
Expression Map
Container/ Collection Map
FnO
Function Map
fnml:functionMap
RML Fields
rml:logicalSource
Logical Source
RML core
Source/ Target
rml:logicalTarget
Logical Target

# RML-star and Morph-KGC$^{star}$

## RML-star

Draft Community Group Report 05 May 2022

**Latest editor's draft:**
https://w3id.org/kg-construct/rml-star

**Editors:**
Ana Iglesias-Molina (Ontology Engineering Group – Universidad Polit
Julián Arenas-Guerrero (Ontology Engineering Group – Universidad I
Thomas Delva (Ghent University – imec – IDLab)
Anastasia Dimou (KU Leuven)
David Chaves-Fraga (Ontology Engineering Group – Universidad Pol

**This Version**
https://kg-construct.github.io/rml-star-spec/20220504/

**Previous Version**
https://kg-construct.github.io/rml-star-spec/20220222/

**Website**
https://rml.io/



## Morph-KGC$^{star}$: Declarative Generation of RDF-star Datasets from Heterogeneous Data

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,daniel.garijo,oscar.corcho}@upm.es
[2] KU Leuven, Dept. Computer Science, Leuven.AI & Flanders Make, Belgium
{david.chaves,anastasia.dimou}@kuleuven.be

**Abstract.** RDF-star has been proposed as an extension of RDF to annotate statements with triples. Libraries and graph stores have started adopting RDF-star, but the generation of RDF-star data remains largely unexplored. In order to allow generating RDF-star from heterogeneous data, RML-star has been proposed as an extension of RML. However, no implementation has been developed so far. In this work, we present Morph-KGC$^{star}$, which extends the Morph-KGC materialization engine to generate RDF-star datasets. We validate Morph-KGC$^{star}$ by running test cases derived from the N-Triples-star syntax tests and we apply it to two real-world use cases from the biomedical and open science domains.

**Keywords:** Knowledge Graphs · RDF-star · RML-star



KU LEUVEN

KU LEUVEN

Ontology Engineering Group

POLITÉCNICA

# Declarative Construction of Knowledge Graphs: one-year status-update

David Chaves-Fraga

david.chaves@kuleuven.be

@dchavesf