W3C Community Group Knowledge Graph Construction

http://w3id.org/kg-construct

Chairs presentation

(Last year) PhD Student at OEG-UPM with Oscar Corcho
research: KG Construction from Heterogeneous Data exploiting Mapping Rules
- Morph suite beyond RDB (CSV, GraphQL, Skyline): https://morph.oeg.fi.upm.es/
- Interoperability among Mapping Languages
- Benchmarking (Virtual) KG construction systems
- Optimizations in KG construction (with SDM-TIB)
- Transport and Smart Cities Domain

postdoc/senior researcher at IDLab, Ghent University/imec
research: High Quality KG Construction from Heterogeneous Data RML.io
- UIs for rules definitions for KG construction (with Pieter Heyvaert)
- data transformation with FnO.io (with Ben De Meester)
- privacy and modeling ontologies & shapes (with Sven Lieber)
- KG construction from big & streaming data (with Gerald Haesendonck)
- KG generation & consumption trade-offs (with Dylan Van Assche)
- Query-answering & OBDA (with Thomas Delva)

Let's start from the beginning…

W3C Workshop on
RDF Access to Relational Databases

RdfRDB Workshop · Program · Workshop Report · Meeting Minutes · Call For Participation · Accepted Papers

25-26 October, 2007 — Cambridge, MA, USA

Hosted by NOVARTIS

W3C

The **mission** of the RDB2RDF Working Group, part of the Semantic Web Activity, is to standardize languages f[...]
Language (R2RML).

RDB2RDF Working Group Charter

Also On This Page W3C RDB2RDF Standards and Notes Inputs Schedule Membership Meeting Records

I. RDB2RDF W3C Standards and Notes

- R2RML: RDB to RDF Mapping Language, W3C Recommendation
- A Direct Mapping of Relational Data to RDF, W3C Recommendation
- R2RML and Direct Mapping Test Cases, W3C Editor's Draft
- Use Cases and Requirements for Mapping Relational Databases to RDF, W3C Working Draft

Current and Upcoming Events

- Weekly teleconferences on Tuesdays 12:00EST (16:00 UTC) An agenda is sent to rdb2rdf-wg 24 hours in advance. The minutes follow within a day or two.
- We will be monitoring comments on [...]

W3C

A Direct Mapping of Relational Data to RDF

W3C Recommendation 27 September 2012

W3C

R2RML: RDB to RDF Mapping Language

W3C Recommendation 27 September 2012

And 2 years ago...

Knowledge Graph Building Workshop

Co-located with the Extended Semantic Web Conference 2019

Portorož, Slovenia – 3 June 2019

SEE CALL FOR PAPERS

Alessandro Negro

Call for Participation in Knowledge Graph Construction Community Group

W3C Team | Posted on: January 8, 2019
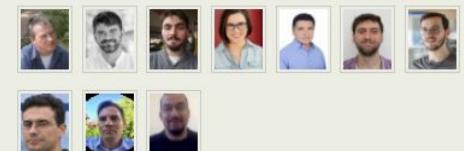
The Knowledge Graph Construction Community Group has been launched:

The overall goal of this community group is to support its participants into developing better 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
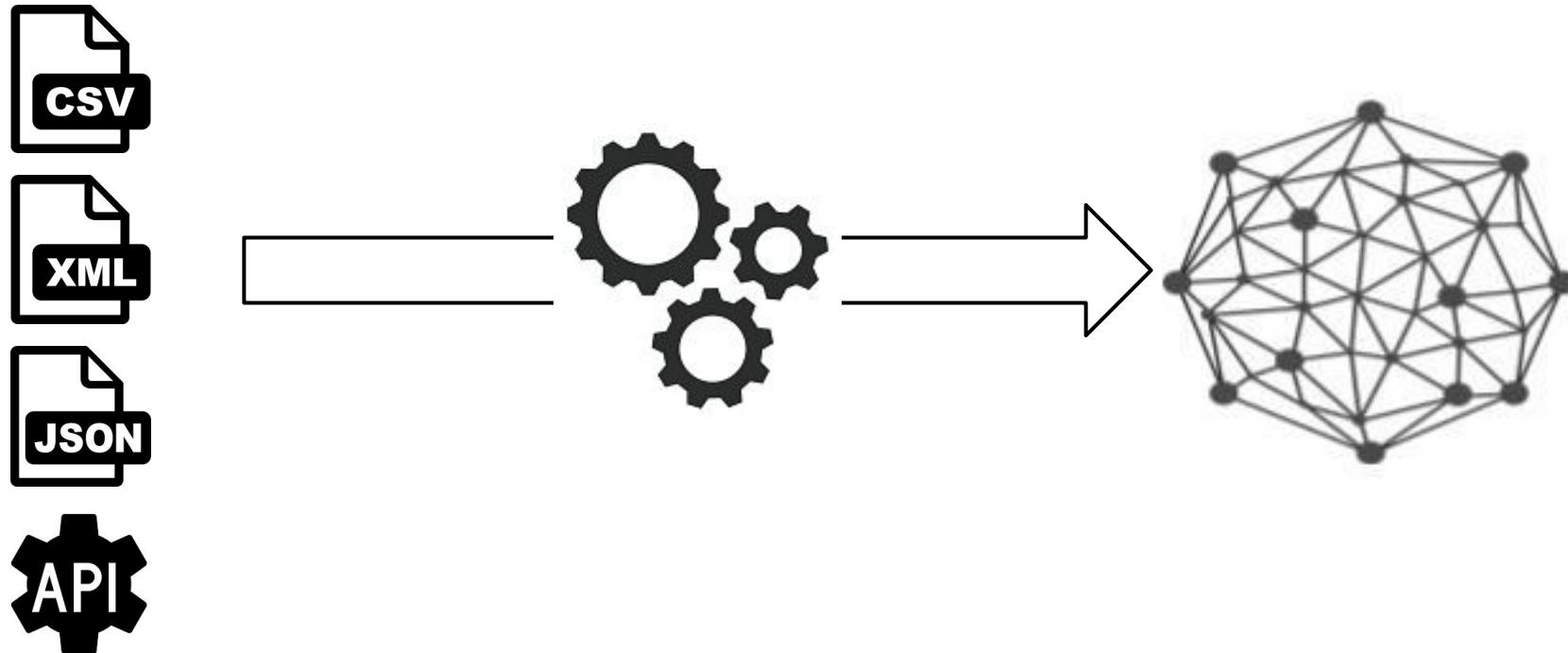
### Participants (85)

View all participants

1st edition ESWC 2019...

KG-construct community group

focus on: semi-structured data (XML, CSV, JSON, etc) → KG (RDF)

Goals

G1: study current methods and implementations

G2: discuss use cases & derive requirements that are not covered

G3: formulate guidelines and best practices

G4: develop methods, resources & tools for evaluations

Goals: G1

G1: study current methods and implementations

- Mature implementations/ideas
- Efficient methods to construct KG
- Virtual KG over heterogeneous sources
- Data streams
- Declarative rules
- Mapping management systems
- …

Goals: G2

G2: discuss use cases & derive requirements that are not covered

- Functions inside or not mapping rules
- Limitations of current languages/tools
- Manual tasks
- …

basic-template.md	Update basic-template.md	5 months ago
betweenourworlds-anime.md	add Between Our Worlds use case	6 months ago
dfki-messy-person-data.md	dfki's use cases	3 months ago
dfki-messy-spreadsheets.md	dfki's use cases	3 months ago
idlab-covid19.md	Add covid 19 use case	6 months ago
idlab-dbpedia.md	correcting names of the files	6 months ago
idlab-facebook.md	Add Facebook use case	6 months ago
idlab-twitter.md	Add Twitter use case	6 months ago
idlab-velopark.md	correcting names of the files	6 months ago
inria-covidontheweb.md	Add use case Covid-on-the-Web	4 months ago
inria-kg-vs-webapis.md	Typos	3 months ago
kadaster-ld.md	add kadaster ld use case	last month
oeg-covid19.md	udpate oeg-covid19	5 months ago
oeg-opencities.md	open cities finished	3 months ago

Goals: G3

G3: formulate guidelines and best practices

- Barriers for adoption of proposed technologies
- Avoid ad-hoc constructions of KG
- Good data source generation (e.g., from NLP process)
- …

Goals: G4

G4: develop methods, resources & tools for evaluations (e.g., implementation reports (current: RML & revived R2RML))

- RML implementation-report: https://rml.io/implementation-report/
    - Generalization of test-cases for any mapping language
- Benchmarks for performance and scalability
- Data quality over KGs or over data sources?

Evaluation Systems for Knowledge Graph Construction

Evaluation System X (Template):

- Name:
- Description:
- Repository/Website:
- Main Features (e.g., parameters that tests):
- Supported data sources and formats:
    - Data format:
    - Sizes or Generator:
- Purpose(Virtual KG /Materialized KG/Both):
- Supported mapping language(s):
- Target and source models:
- Contact point:
- DOI:
- License:

And we are right now...

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
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

Anastasia Dimou   *Chairs*

David Chaves-Fraga

Participants (124)

Ongoing work:

- New R2RML Implementation report:
  http://w3id.org/kg-construct/r2rml-implementation-report

- Supporting report generation [R2]RML
  https://github.com/kg-construct/r2rml-test-cases-support
  https://github.com/kg-construct/rml-test-cases-support

- Mapping challenges (and solutions):
  http://w3id.org/kg-construct/workshop/challenges
  http://github.com/kg-construct/mapping-challenges/issues

- New and conceptual test cases

- RML+FnO specification

- Bi-weekly meetings (~20 people)

- Monthly blog post

http://github.com/kg-construct   http://w3id.org/kg-construct

# Reviving the R2RML implementation report
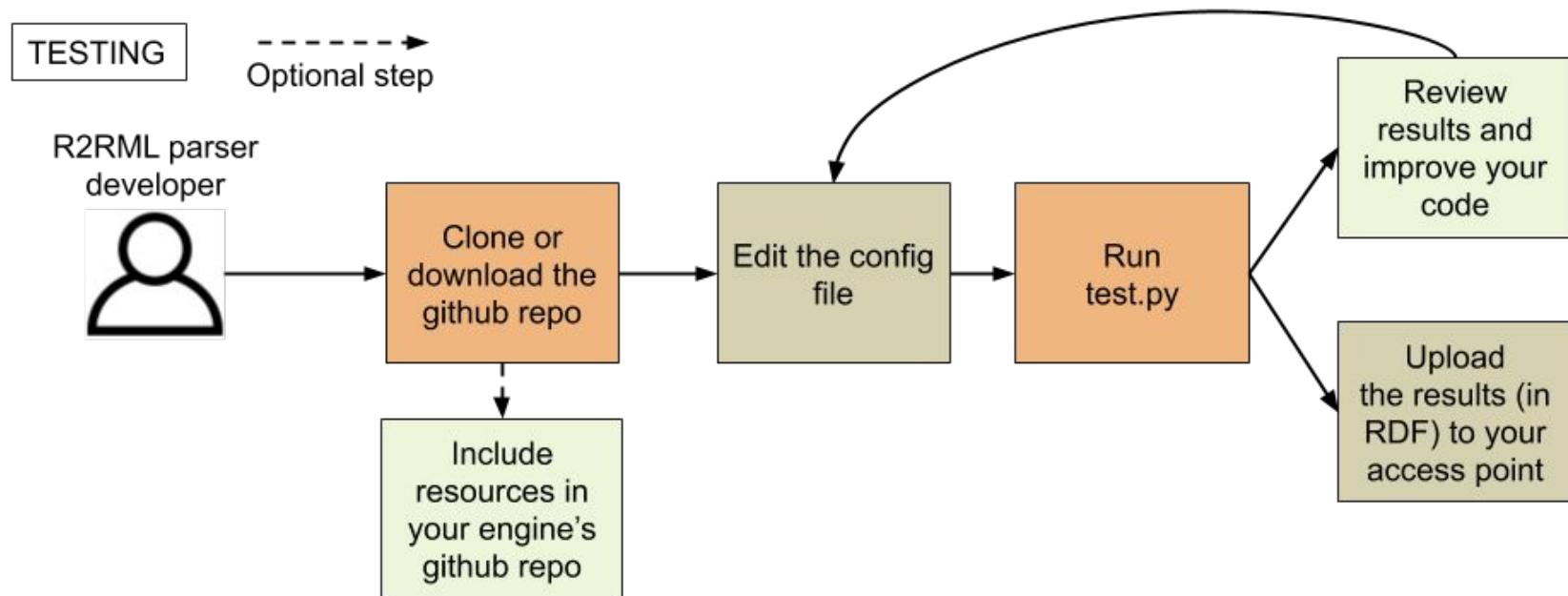
§ 3.1 MySQL

The following table lists the results of the test cases when using an MySQL database.
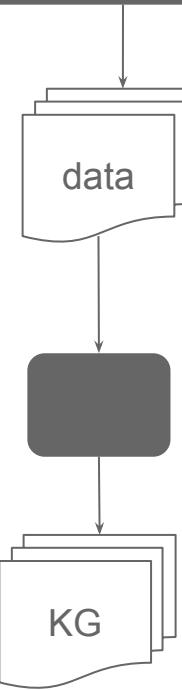
Test Case	Morph-RDB	Db2triples	R2RML-F	Ontop	RMLMapper
R2RMLTC0000	passed	passed	passed	passed	passed
R2RMLTC0001a	passed	failed	failed	passed	passed
R2RMLTC0001b	passed	failed	failed	passed	passed
R2RMLTC0002a	passed	failed	failed	passed	passed
R2RMLTC0002b	passed	failed	failed	passed	passed
R2RMLTC0002c	failed	passed	failed	passed	passed
R2RMLTC0002d	passed	failed	failed	failed	passed
R2RMLTC0002e	failed	passed	failed	passed	passed
R2RMLTC0002f	failed	passed	failed	failed	failed
R2RMLTC0002g	failed	passed	failed	passed	passed
R2RMLTC0002h	failed	passed	failed	passed	failed
R2RMLTC0002i	passed	failed	failed	passed	passed

- **Decentralized approach**
- Using SW technologies
- 1 section per DBMS
- Results are on engines repos
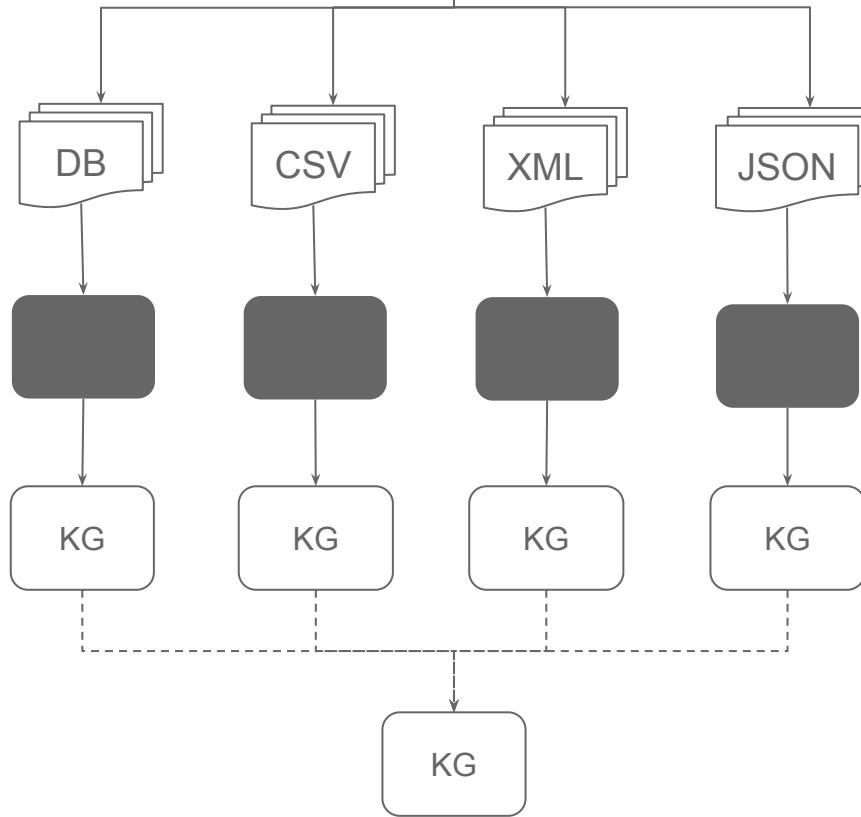
Supporting the generation of EARL reports

RDB

shema
rules

schema
transformation

R D F

FILE DB RDB Web API

source

shema rules

source access

schema transformation

R D F
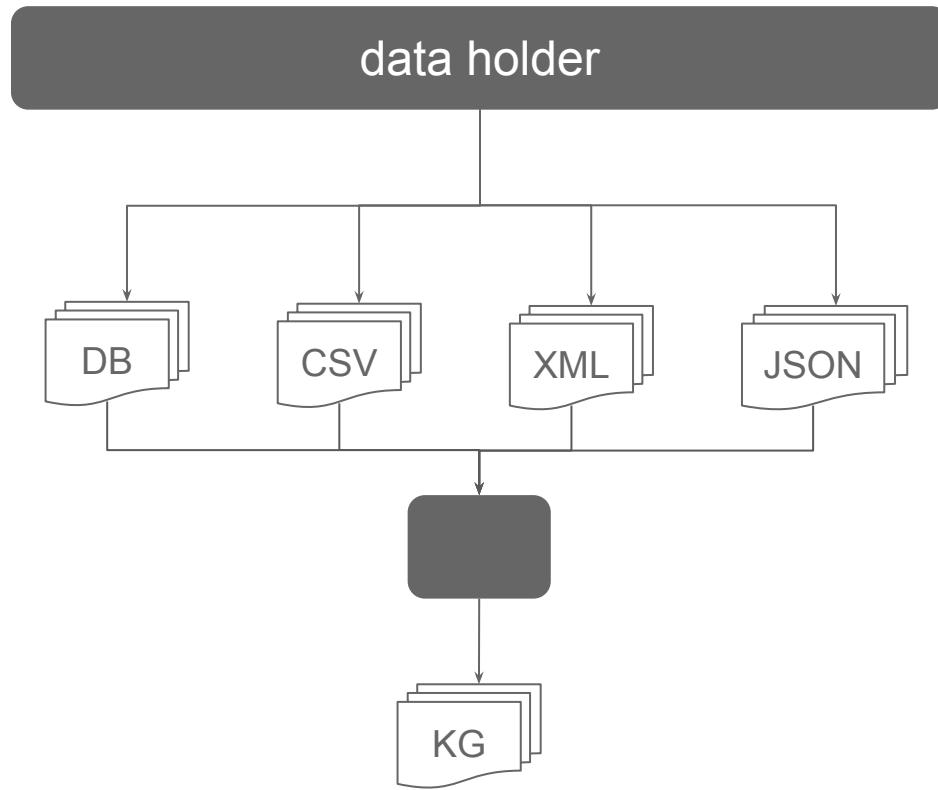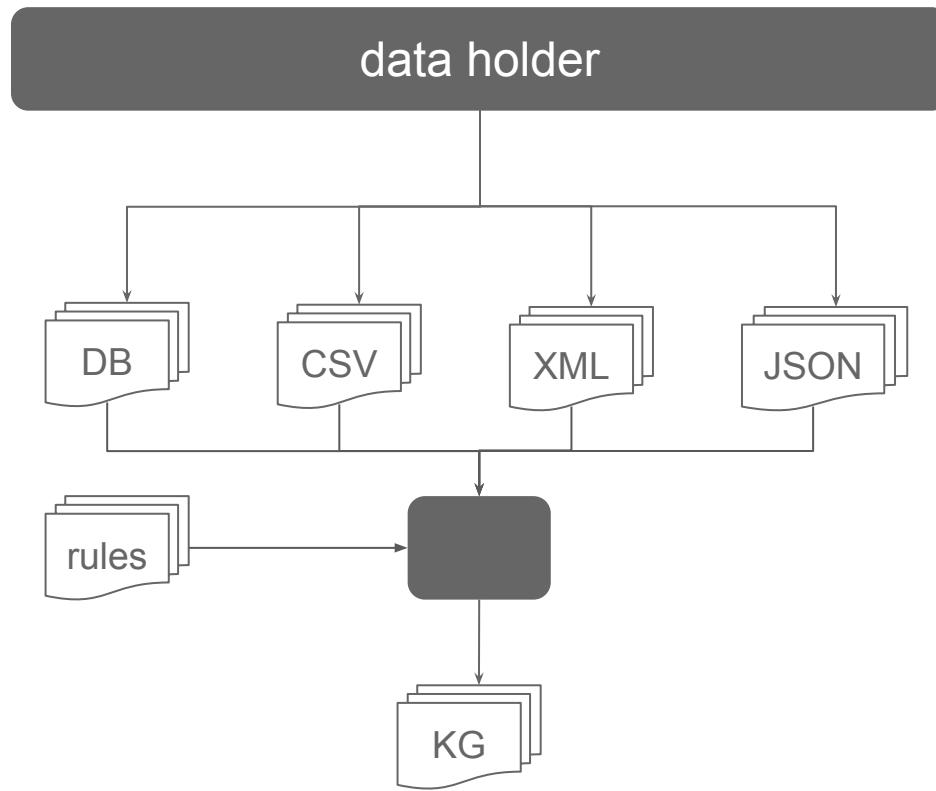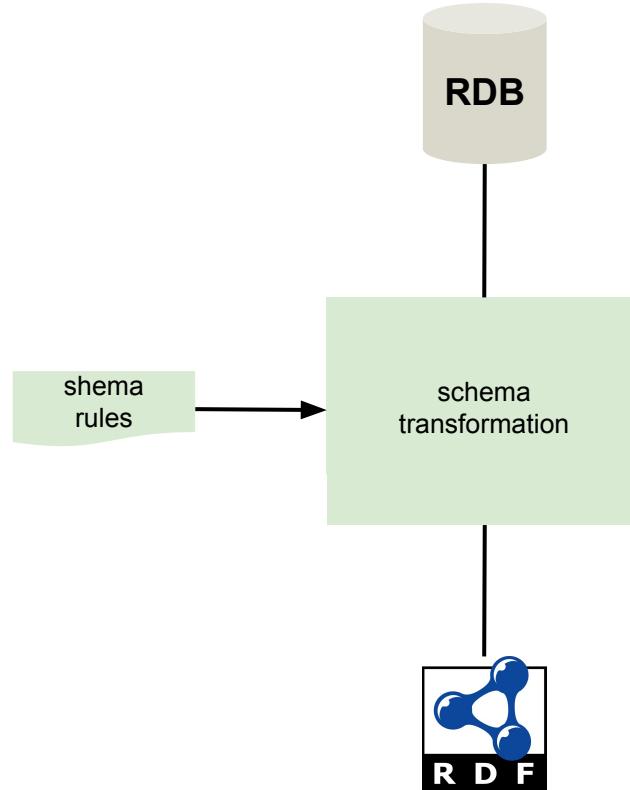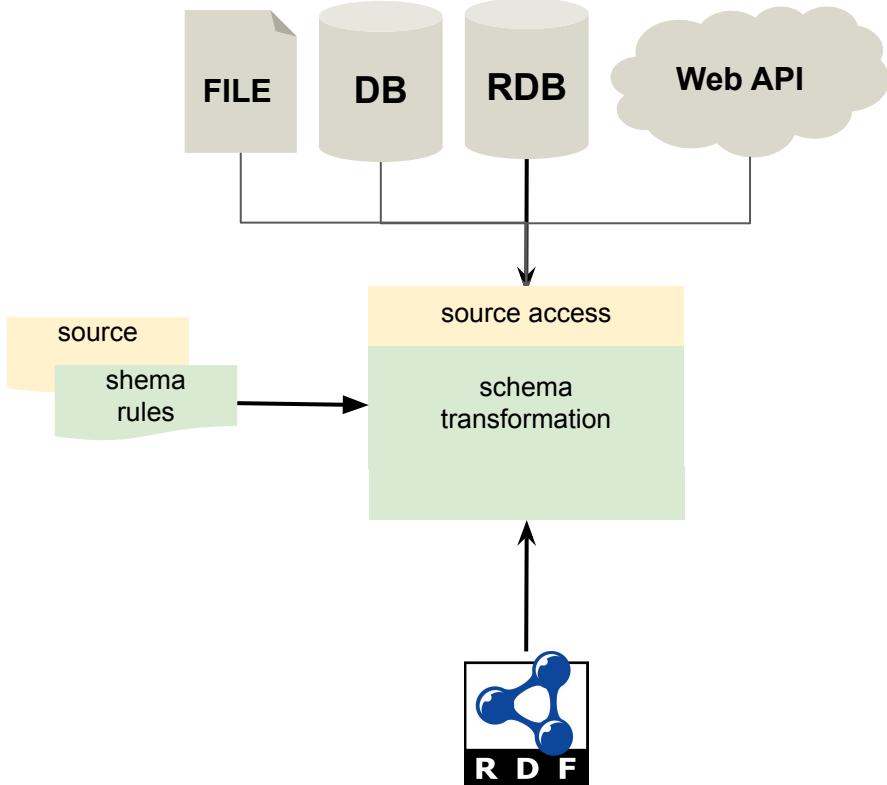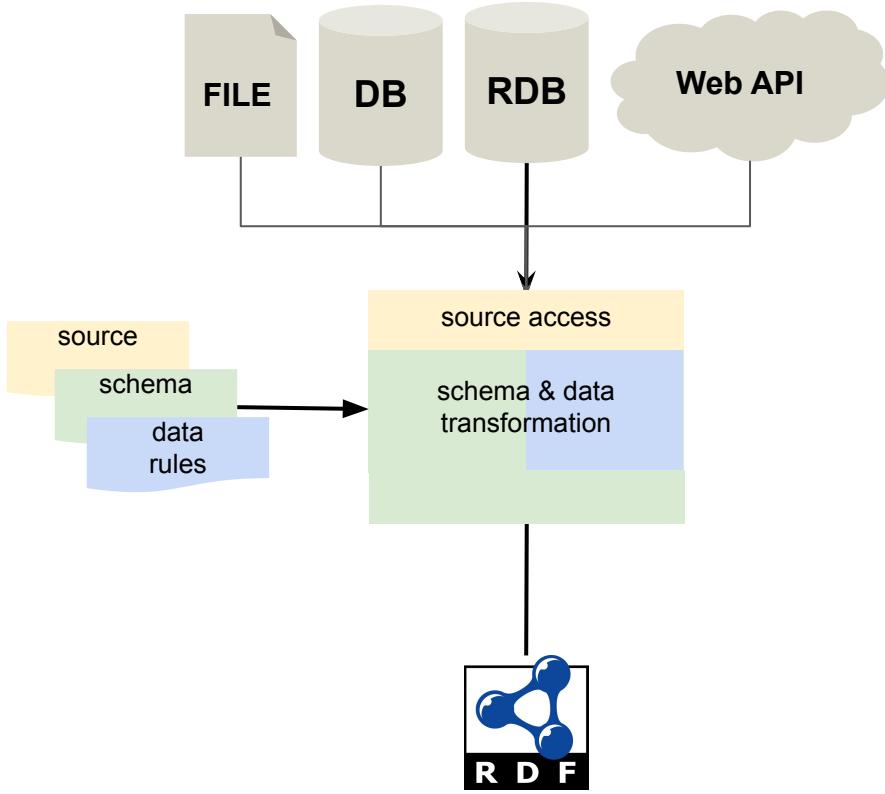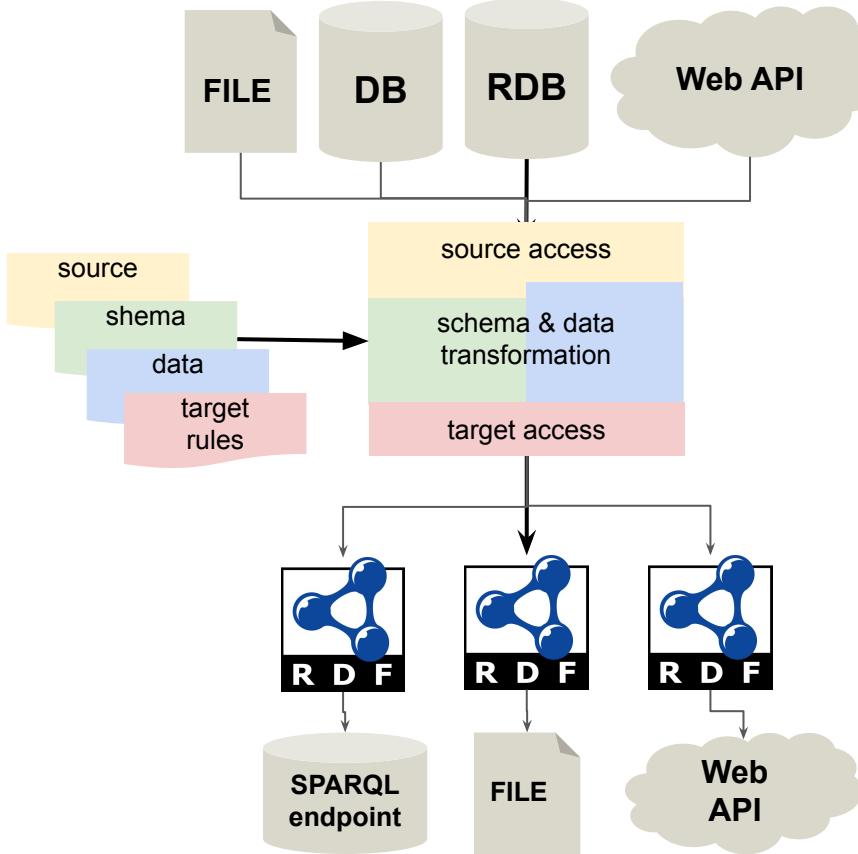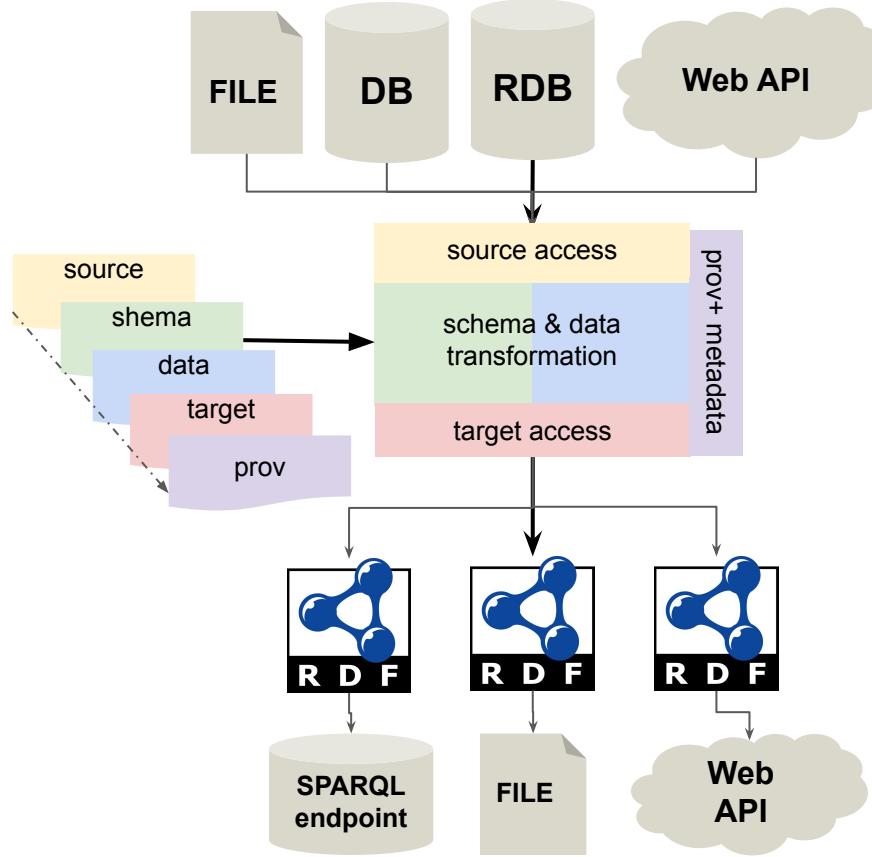
Machine-interpretable dataset & service descriptions for heterogeneous data access and retrieval. A. Dimou et al. SEMANTICS 2015

FILE DB RDB Web API

source access

source
schema
data
rules

schema & data
transformation

RDF
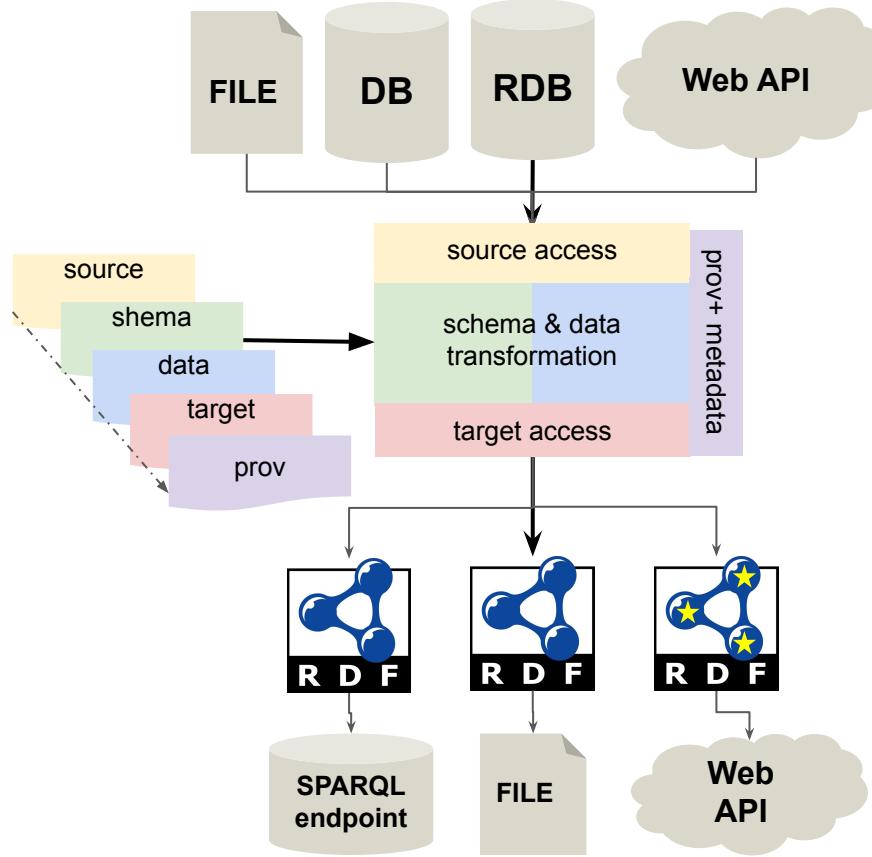
An Ontology to Semantically Declare & Describe Functions
B. De Meester, A. Dimou, R. Verborgh, E. Mannens & R. Van De Walle. ESWC P&D 2016

Leveraging Web of Things W3C recommendations for knowledge graphs generation.
D.Van Assche et al. ICWE2021

Automated Metadata Generation for Linked Data Generation and Publishing Workflows. A. Dimou et al. LDOW 2016
Detailed Provenance Capture of Data Processing. B. De Meester, A. Dimou, R. Verborgh, E. Mannens. SemSci2017

Automated Metadata Generation for Linked Data Generation and Publishing Workflows. A. Dimou et al. LDOW 2016
Detailed Provenance Capture of Data Processing. B. De Meester, A. Dimou, R. Verborgh, E. Mannens. SemSci2017

Mapping Challenges*

limitations of R2RML
limitation of generalizing R2RML

* Extracted from: García-González, H. (2021). A ShExML Perspective on Mapping Challenges: Already Solved Ones, Language Modifications and Future Required Actions. *In Proceedings of the 2nd International Workshop on Knowledge Graph Construction co-located with 18th Extended Semantic Web Conference*

Mapping Challenges I: Dynamic Datatype

Generate datatype from input data

Dynamic vs static

```
{
  "persons": [
    {
      "firstname": "John",
      "lastname": "Doe",
      "lang": "en",
      "num": 3,
      "dt": "http://www.w3.org/2001/XMLSchema#integer"
    },
    {
      "firstname": "Jane",
      "lastname": "Smith",
      "lang": "fr",
      "num": "3.14",
      "dt": "http://www.w3.org/2001/XMLSchema#decimal"
    }
  ]
}
```

Mapping Challenges II: Dynamic Language tag

Generate language tag from input data

Dynamic vs static

Different possible format inputs

- en
- English

```
{
    "persons": [
        {
            "firstname": "John",
            "lastname": "Doe",
            "lang": "en",
            "num": 3,
            "dt": "http://www.w3.org/2001/XMLSchema#integer"
        },
        {
            "firstname": "Jane",
            "lastname": "Smith",
            "lang": "fr",
            "num": "3.14",
            "dt": "http://www.w3.org/2001/XMLSchema#decimal"
        }
    ]
}
```

Mapping Challenges III: Generate multiple values

Multi-language or multi-datatype values for the same subject

Additionally, default languages or datatypes

```
exPerson:John    ex:name    "John"@fr , "John"@en .
```

```
{
    "lastname": "Doe",
    "firstname": [
        {
            "label": "John",
            "lang": "en"
        },
        {
            "label": "John",
            "lang": "fr"
        }
    ]
}
```

```
{
    "firstname": "John",
    "lastname": "Doe",
    "lang": "fr"
}
```

Mapping Challenges IV: Join on Literals

Joins, by default, generate resources

There is no way to output literals instead

```
experson:2   :affiliation   "Uni2" ;
         :lastName       "Dane" .

experson:1   :affiliation   "Uni1" ;
         :lastName       "Doe" .
```

```
{
    "author": [
        {
            "id": 1,
            "firstname": "John",
            "affiliation": "Uni1"
        },
        {
            "id": 2,
            "firstname": "Jane",
            "affiliation": "Uni2"
        }
    ],
    "people": [
        {
            "firstname": "John",
            "familyName": "Doe"
        },
        {
            "firstname": "Jane",
            "familyName": "Dane"
        }
    ]
}
```

Mapping Challenges V: Multi-value references

How to deal with the expected output in a hierarchical file

Cartesian product or respect the current relation

Join condition poses problems in JSON files as it is not possible to go upwards

```json
{
    "labName": "AmazingLab1",
    "articles": [
        {
            "title": "article1",
            "authors": [
                {
                    "name": "Alice",
                    "affiliation": [
                        {
                            "label": "Uni1"
                        },
                        {
                            "label": "Company2"
                        }
                    ]
                },
                {
                    "name": "Bob",
                    "affiliation": [
                        {
                            "label": "Uni3"
                        },
                        {
                            "label": "Company4"
                        }
                    ]
                }
            ]
        },
        { … } // 2 items
    ]
}
```

Mapping Challenges VI: Access fields outside iterators

How to access upper fields

JSON path doesn't allow going upwards

From cars how to reach owners?

```json
{
    "records": [
        {
            "id": "1",
            "enteredBy": "Alice",
            "cars": [
                {
                    "make": "Mercedes"
                },
                {
                    "make": "Honda"
                }
            ]
        },
        {
            "id": "2",
            "enteredBy": "Bob",
            "cars": [
                {
                    "make": "Mercedes"
                },
                {
                    "make": "Toyota"
                }
            ]
        }
    ]
}
```

Mapping Challenges VII: RDF Collections

Generate collections from multi- value references

Different RDF Collections and Containers

- List
- Bag, Seq, Alt

```
{
    "labName": "AmazingLab1",
    "article": {
        "title": "article1",
        "authors": [
            {
                "name": "Alice"
            },
            {
                "name": "Bob"
            }
        ]
    }
}
```

Mapping Challenges: Current status
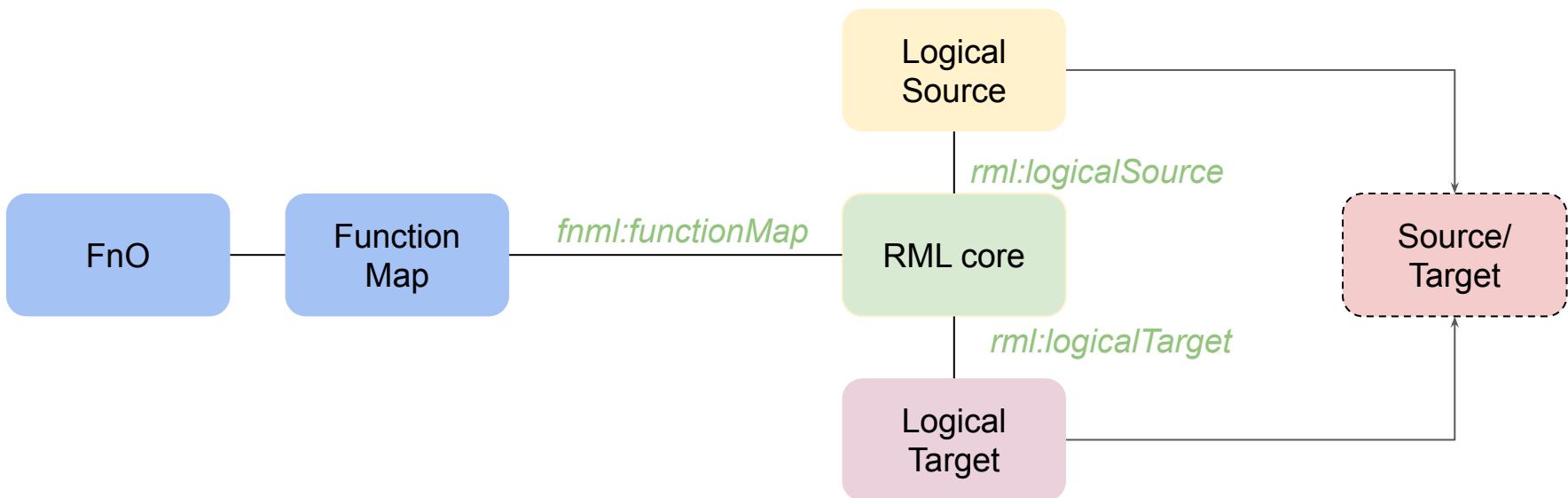
Partial solutions proposed:

- RML fields
- Extension of ShexML
- RML extension
- xR2RML (not actually an extension)

More info and details https://github.com/kg-construct/mapping-challenges

Current spec: https://kg-construct.github.io/rml-fno-spec/

New RML(?) specification

2ⁿᵈ Knowledge Graph Construction Workshop

Keynote by Jesús Barrasa@Neo4j (15:00 - 16:00 CEST)
**Knowledge graphs 2021: The great convergence"**

Machine Learning for KG construct panel (17:30 - 18:30 CEST)

Maria-Esther Vidal
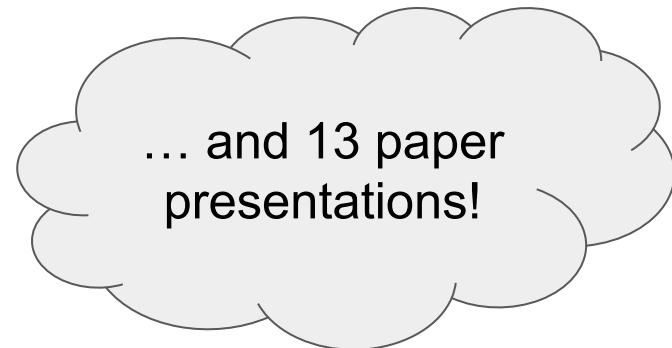
Heiko Paulheim

Francesco Osborne

Ernesto Jiménez-Ruiz

… and 13 paper presentations!

2nd edition of the KGC Workshop

With an special focus on **user's role** and **mapping challenges**

- 16 papers submitted (1 withdraw, 2 rejected, 13 accepted):
  - 5 research papers
  - 4 experience papers
  - 3 position papers
  - 1 demo paper

- Open-Review process:
  - 42 distinguish researchers part of the Program Committee
  - All papers received between 3 and 4 reviews + meta-review
    - Besides some exceptions, each paper was reviewed by
      1 professor, 1 senior researcher, 1 junior researcher, 1 reviewer from industry

- ~40 attendees and 9 hours of discussion

Workshop outcomes I: Bridging the gaps

Any RDF graph can be automatically imported into a PG DB in a lossless manner

Any Property Graph can be automatically serialised as RDF (or RDF*) in a lossless manner.

Workshop outcomes II: Jesús Barrasa's Observations

O1: #KnowledgeGraph construction is largely an engineering task! Is it?

O2: There is no model reuse!

O3: #KnowledgeGraph construction is augmented by automation e.g., #MachineLearning

O4: #KnowledgeGraphs are constructing reusing fragments of other #KnowledgeGraphs
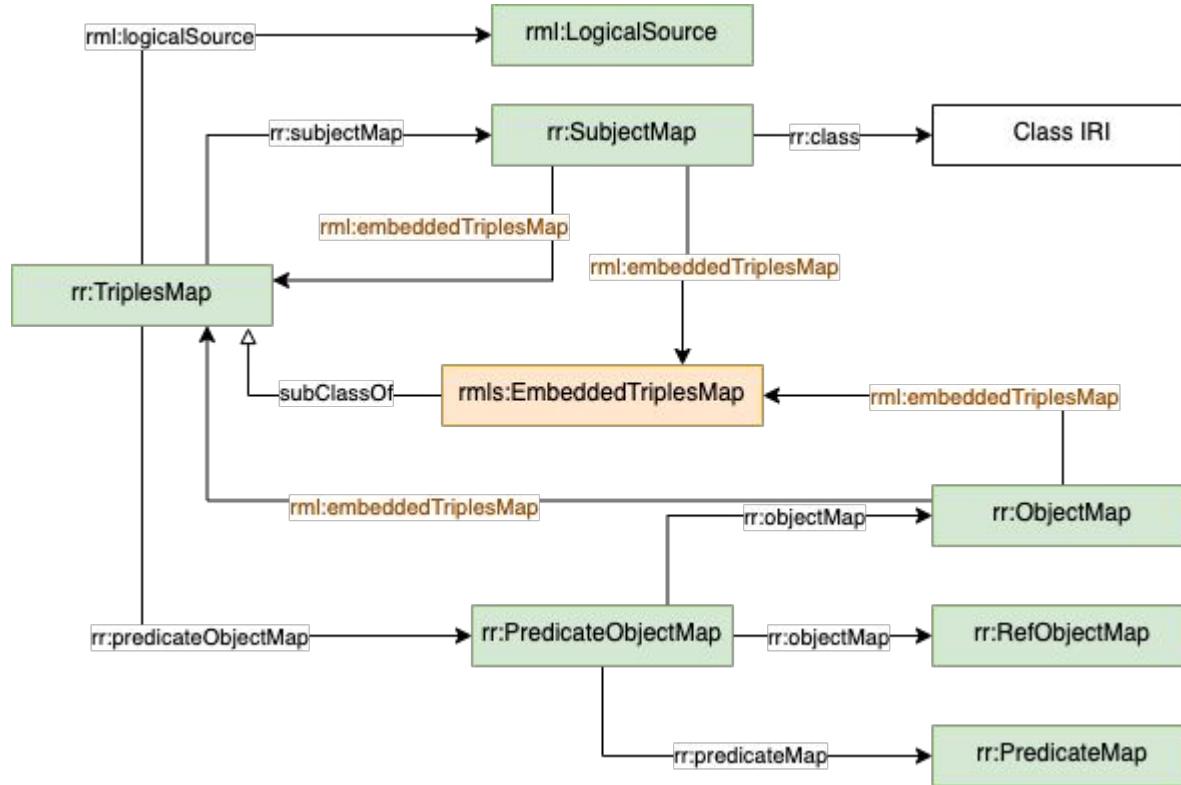
O5: **It is not #propertyGraphs Vs #RDF graphs but #propertyGraphs AND #RDF graphs! & RDF* makes it a lot easier!**

O6: There is many valuable #graphs!

Workshop outcomes III: Panel

- Current mapping language specifications **are not enough to covering real use cases** (by Ernesto)

- We need **hybrid solutions** for enhancing current KGC approaches,
  symbolic WITH subsymbolic (all panelists)

- **(mapping) rules are really relevant**, and they have to be explicitly defined (by Heiko)

- **Users in the loop 100%**: UI for non-experts and DSL or YARRRML for developers (by Juan)

- #SemTab should be redefine with more complex tasks.
  R/P/F-Measure are not enough (by Ernesto and Juan)

- Current best systems are the ones with a **big amount structured knowledge in the backend**,
  although they are quite simple in the front (by Maria-Esther)

- The relevance of **high education courses about Knowledge Science/Graphs** for the new
  generation of data scientists

RML-star for RDF-star (w/ Ana&Julián)

Thanks! Questions?