# Virtual Knowledge Graph Generation from Heterogeneous Data Sources

**David Chaves-Fraga, Ontology Engineering Group**
**Universidad Politécnica de Madrid, Spain**
Freddy Priyatna, OEG-UPM
Oscar Corcho, OEG-UPM

schema:email = lower(substr({name},1,1) || {surname} || '@fi.upm.es')

dchaves@fi.upm.es
@dchavesf

24-25/10/2018
@imec/Ghent University

# Who am I?

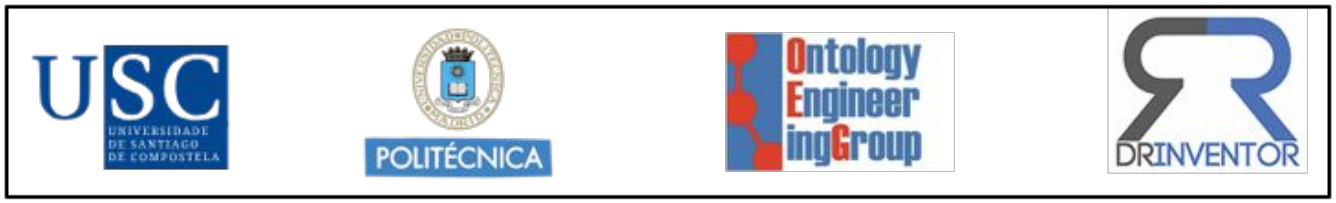
PhD Student and Researcher at OEG-UPM since 2016 (Data Integration team):

- MSc Thesis (2016): Methods and Techniques for the Evaluation of Ontology Learning
- PhD Thesis (2016-2020): Virtual Knowledge Graph Generation from heterogeneous resources

Interests:

- **OBDA**
- **Heterogeneous data**
- SPARQL
- Federated queries
- **Data Integration**
- Public Transport
- Linked Connections
- **R2RML- RML**
- **Virtualization - Access**

2011-2016




2017


2018

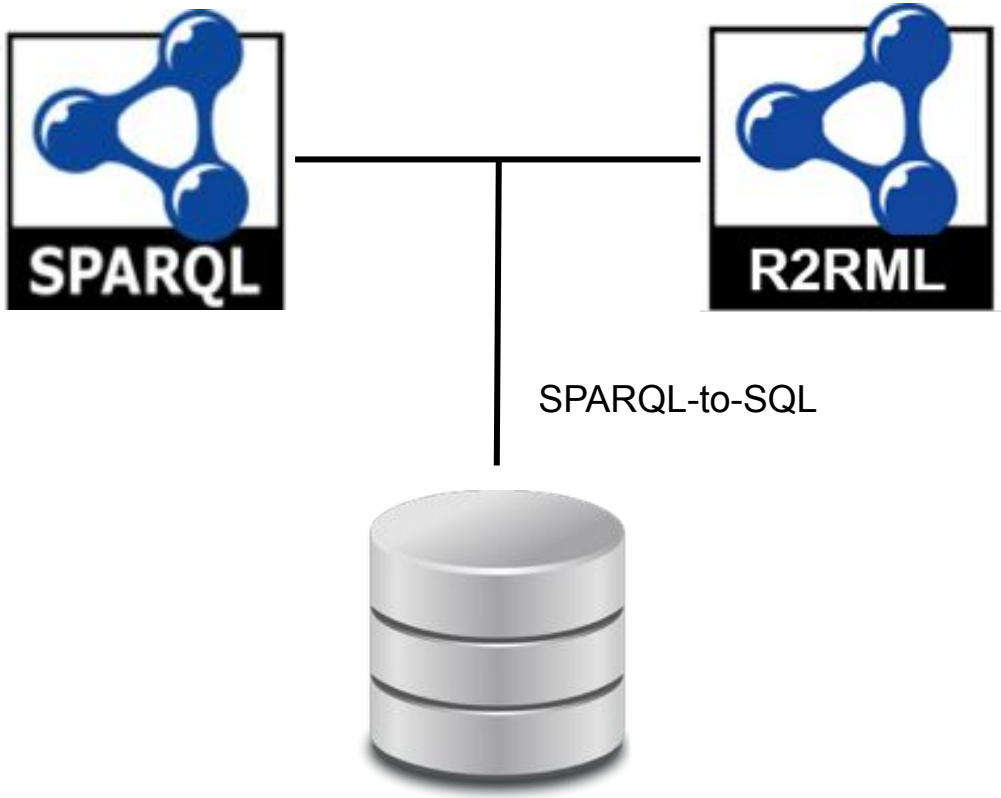

@dchavesf

dchaves@fi.upm.es

dchaves.oeg-upm.net

dachafra

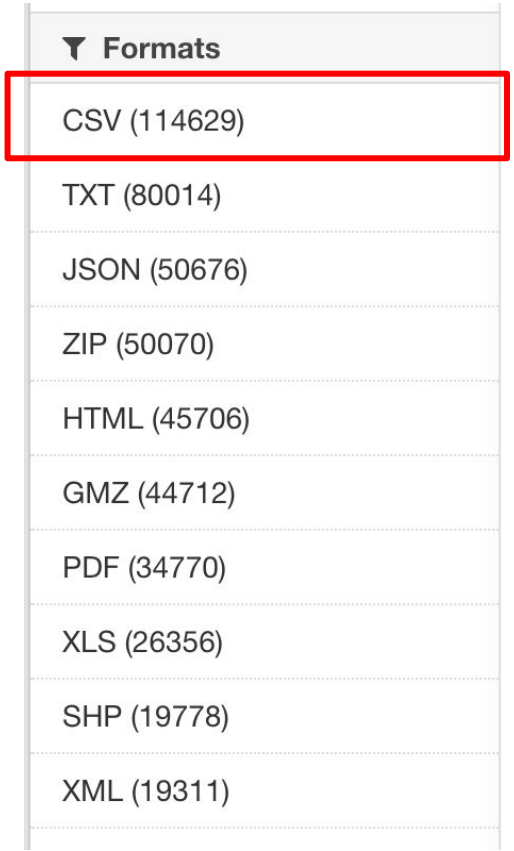
# OBDA… **O**ntology **B**ased **D**ata **A**ccess

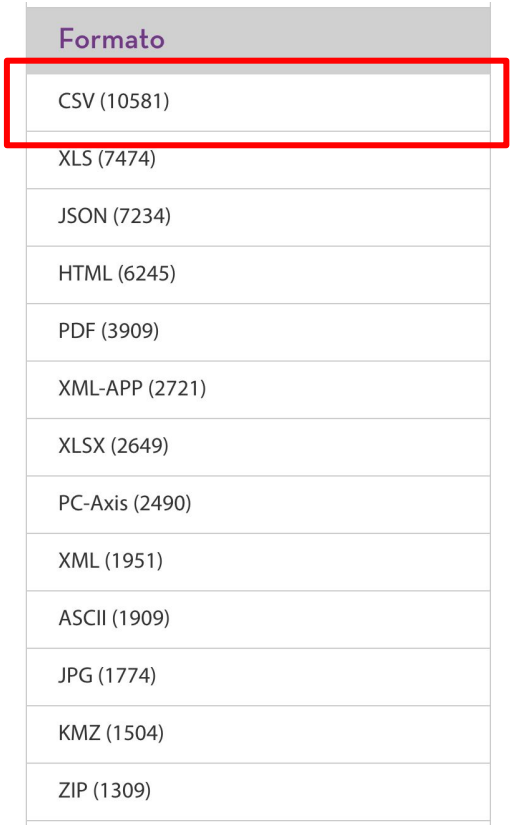



Focused on optimizing the generated SQL query to improve the performance

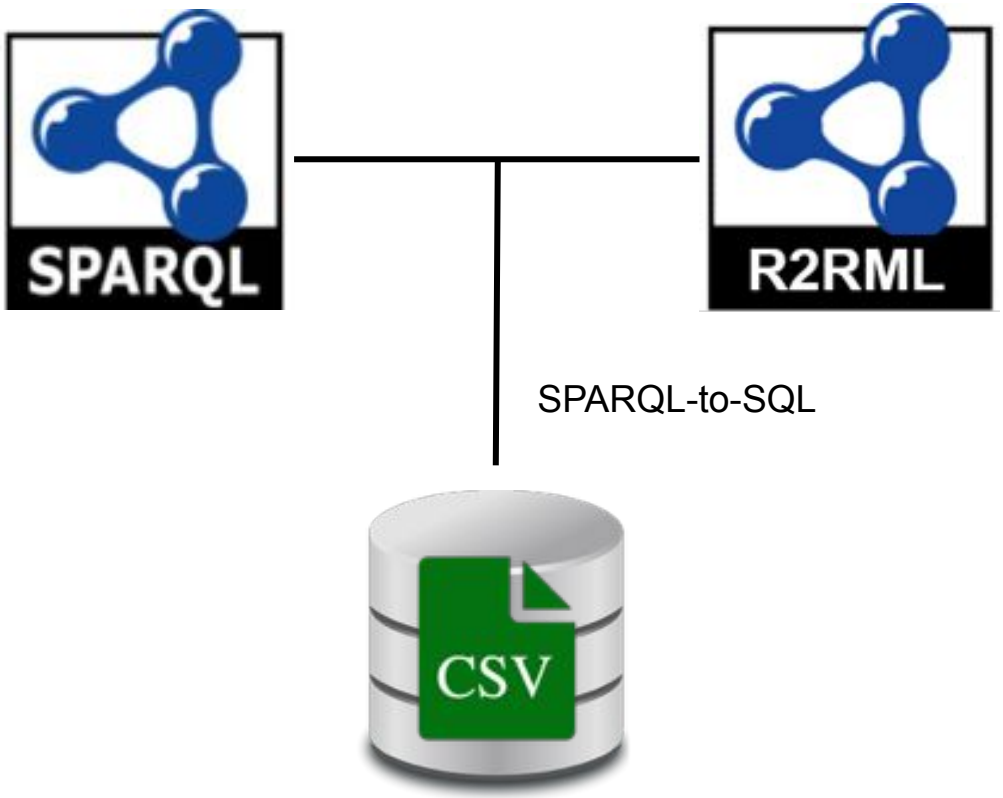
But we are working on… Semantic **WEB**

## How is the data exposed on the Web?

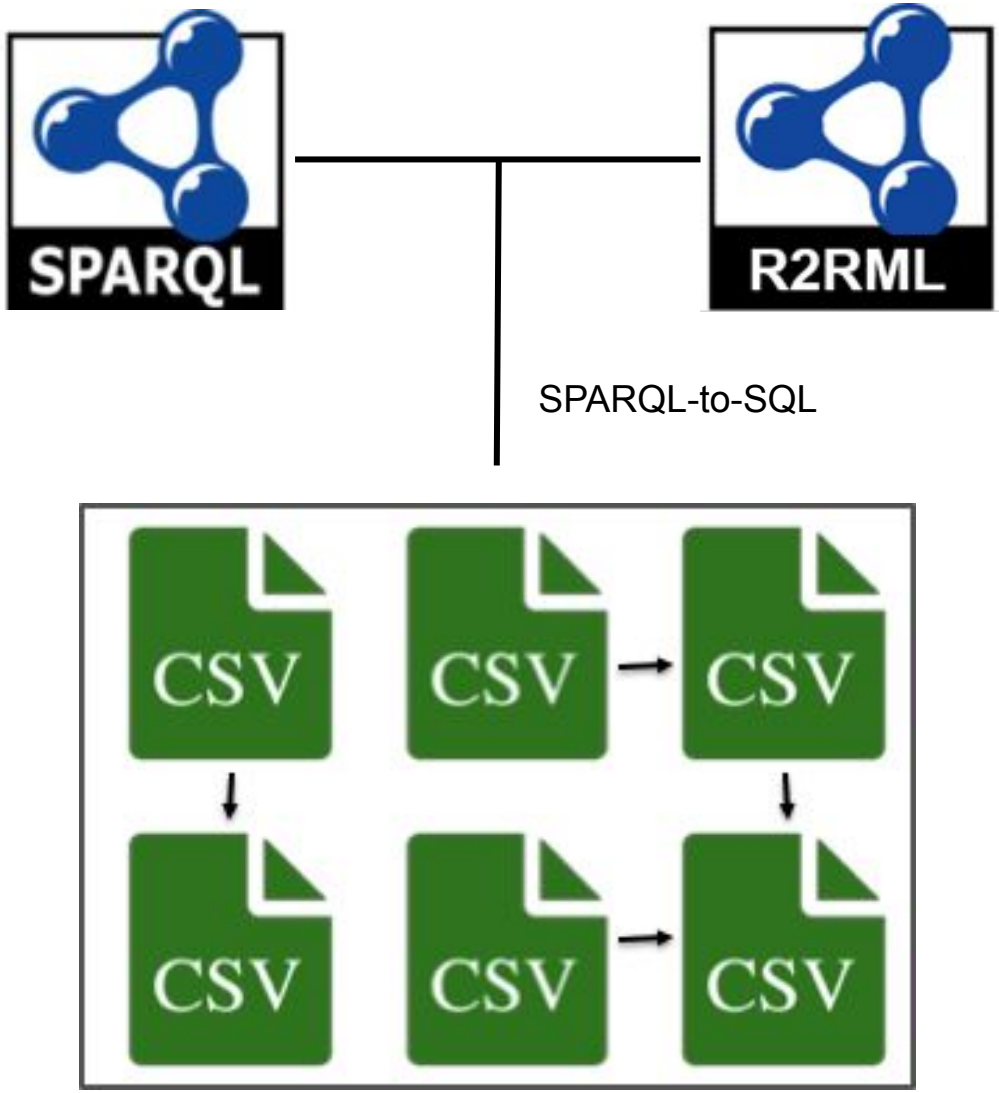



# OBDA… **O**ntology **B**ased **D**ata **A**ccess



# OBDA… **O**ntology **B**ased **D**ata **A**ccess

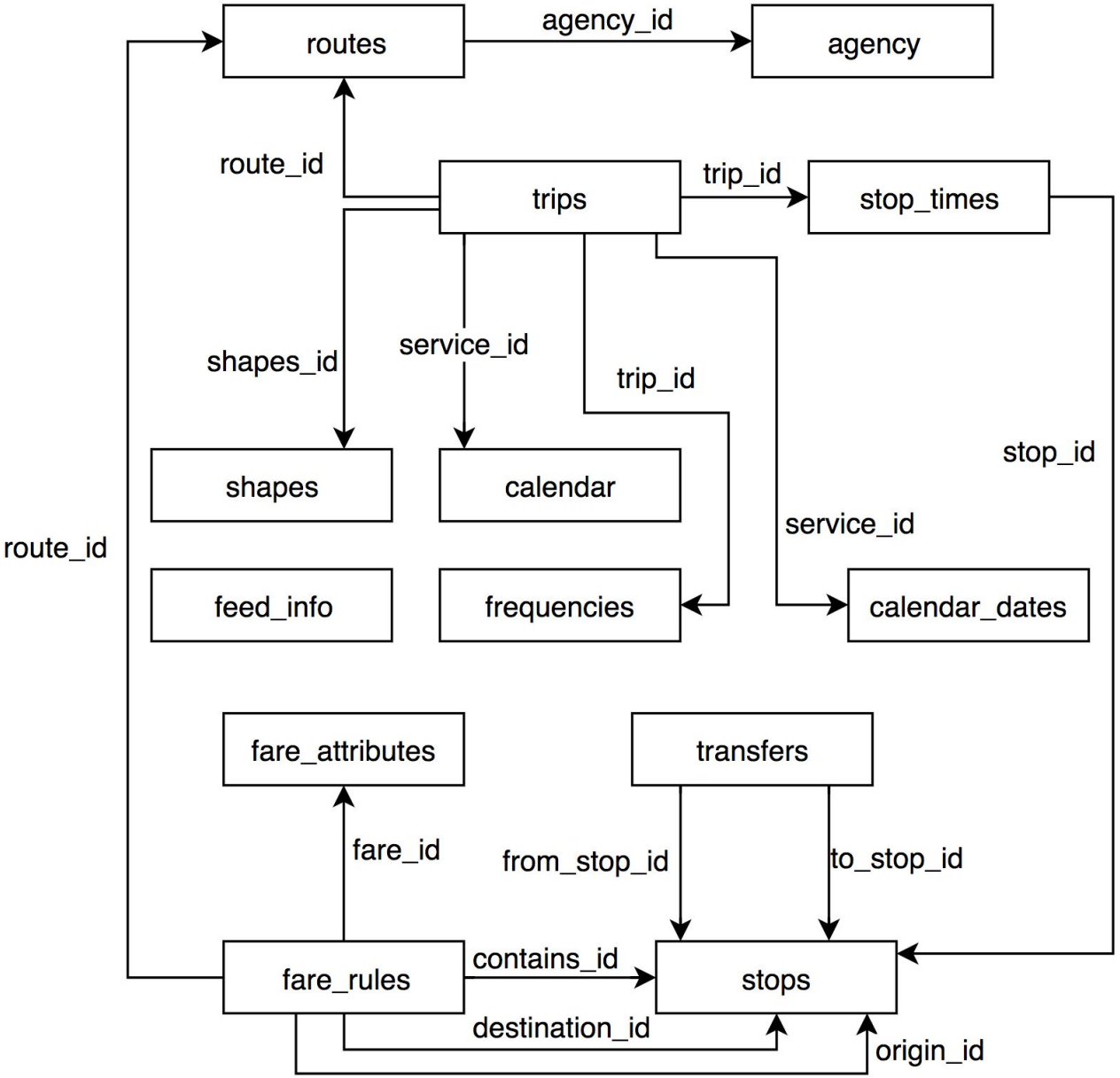


Multiple CSV files with relations among them:

1. Joins are not explicit
2. Constraints are not defined explicitly in the CSVs (PK, FKs)
3. The data may not be in the desirable format (e.g. dates)
4. CSVs are not in 3NF:
    a. PK may be repeated
    b. FKs may not be explicited
    c. FKs could not have a 1:1 cardinality
    d. Lists in column

R2RML is not enough for dealing with CSV(s) in an OBDA approach

# Let's give an example...



# LD Generation from GTFS to LinkedGTFS (in hours)

| Dataset (size mg) | Morph-R2RML | RML-Mapper |
|---|---|---|
| D1 (2.3) | 0.004 | 3.739 |
| D2 (2.6) | 0.026 | 2.587 |
| D3 (2.9) | 0.068 | 0.778 |
| D4 (3.4) | 0.118 | 7.026 |
| D5 (4.2) | 0.115 | 7.026 |
| D6 (4.7) | 0.217 | 12.218 |
| D7 (31) | 1.153 | 151.541 |
| D8 (96) | 12.496 | >160 |

# Focused on CSVs files...

Our mission as researchers it to provide solutions for:

- Generate Linked Data when:
  - Quality is important
  - The data is static

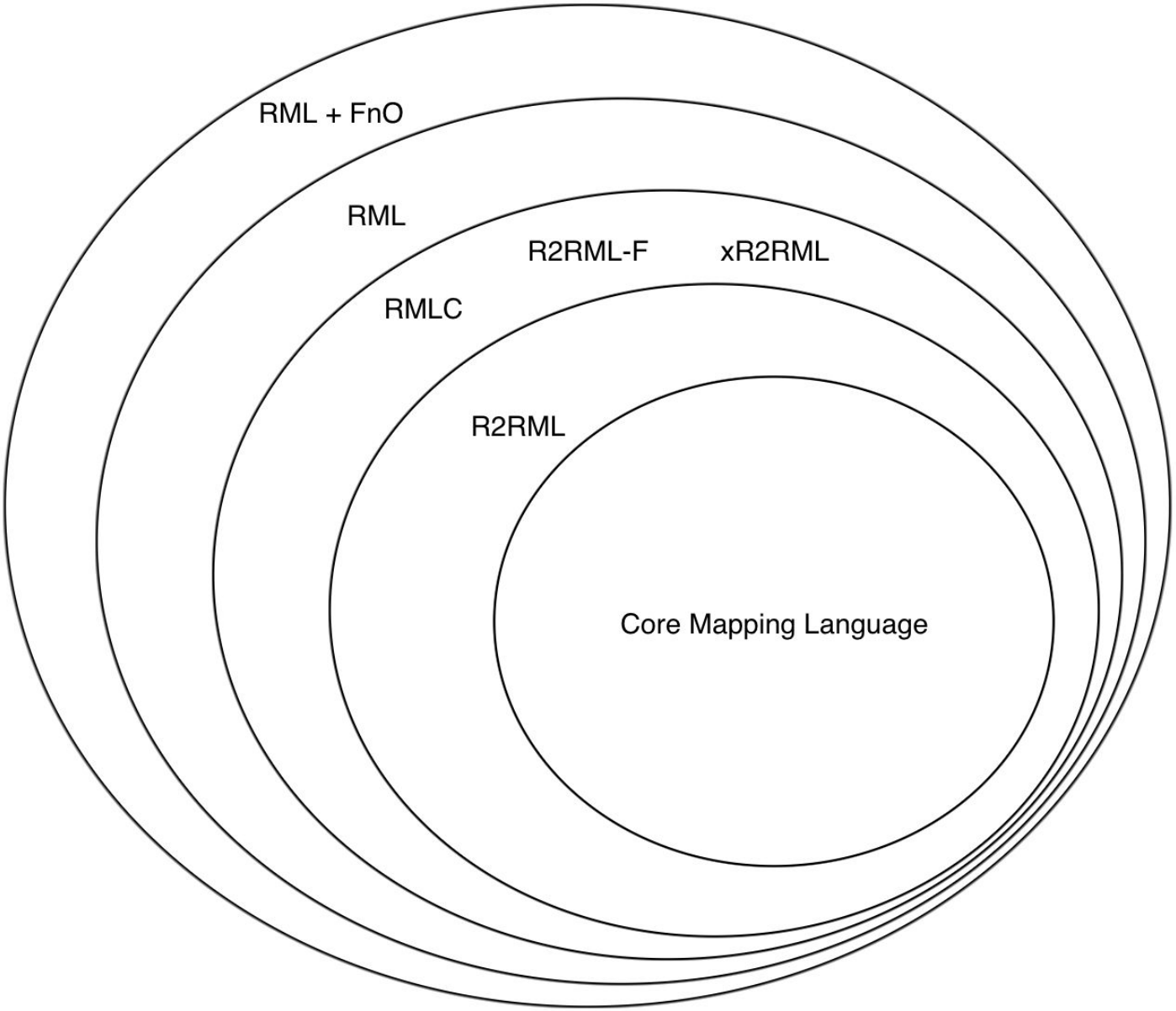
- Access to data using a graph query language when:
  - Data is volatile
  - Performance is relevant
  - Underlying query engine for translation exists

**Common point: The mapping language!**

How to provide **access**/generation to heterogeneous data exposed on the web with relations among them using semantic technologies?

How can we extend standard Mapping Languages maintaining their semantics for using **OBDA engines** or LD generators?

# Core Mapping Language + extensions



| Feature | R2RML | RML | RML-C |
|---|---|---|---|
| Data format | RBD | JSON,CSV, XML | CSV |
| Materialization | Yes | Yes | Yes |
| Virtualization | Yes | No | Yes |
| Functions | No | Yes (FnO) | Yes (SQL Functions) |
| Specification | Yes | Partially? (FnO+RML?) | Partially |

"Virtual Statistics Knowledge Graph Generation from CSV files" D. Chaves-Fraga, F. Priyatna, I. Santana-Perez and O.Corcho at *SemStats Workshop co-located with ISWC18* (Best Paper Award)

"SATET: Providing access to multiple CSV on the Web using OBDA" D. Chaves-Fraga and O.Corcho (on-going work)

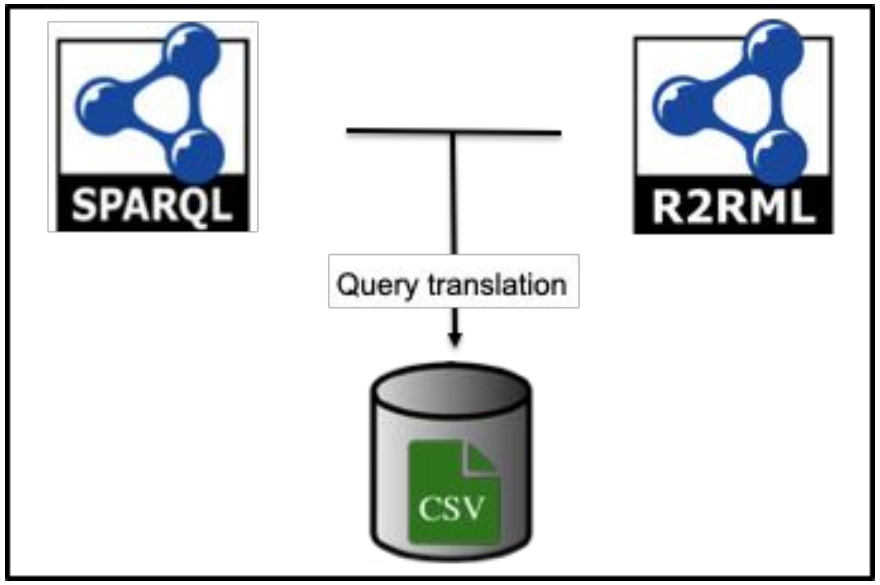

# Virtual Statistics Knowledge Graph Generation from CSV files

**David Chaves-Fraga, Ontology Engineering Group**
**Universidad Politécnica de Madrid, Spain**
Freddy Priyatna, OEG-UPM
Idafen Perez-Santana, OEG-UPM
Oscar Corcho, OEG-UPM

dchaves@fi.upm.es
@dchavesf

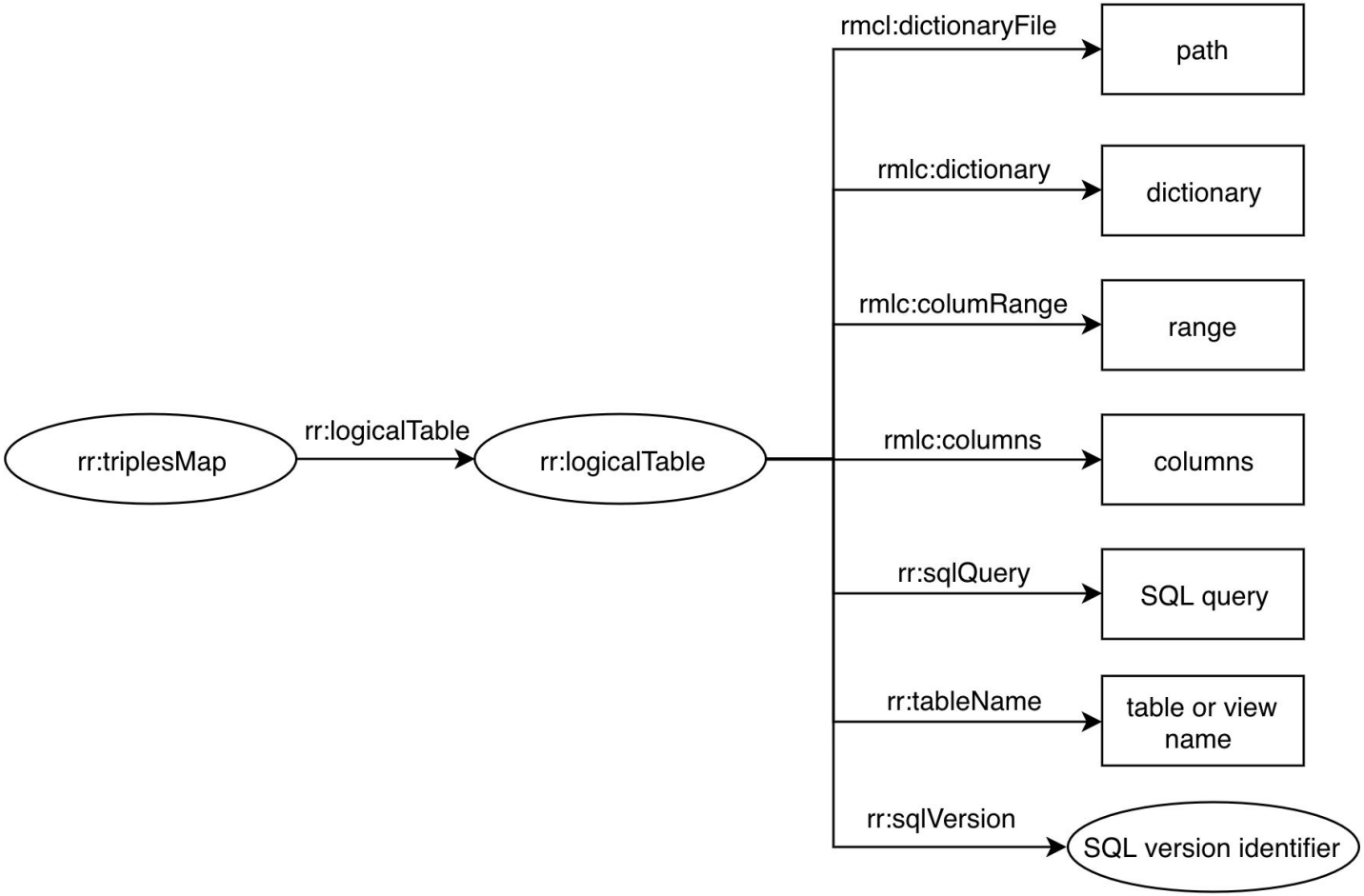
08/10/2018
ISWC18-SemStats

# Virtual Statistics Knowledge Graph Generation




The **size** of the R2RML mapping depends on the **number of columns** in the CSV


Difficulty of maintenance and creation




Two variables for identifying independently each TriplesMap and provide access to the CSV data: {$column},{$alias}

# RMLC-Iterator Example

```
<TriplesMap2016{$column}>
rr:logicalTable [
    rr:tableName "\"2016-P21\"";
    rmlc:columns ("Jan","Oct","Dec");
    rmlc:dictionary {"Jan":"January","Oct":"October","Dec":"December"};
];

rr:subjectMap [
    a rr:Subject;
    rr:template "http://ex.org/2016{$column}";
    rr:termType rr:IRI;
    rr:class qb:Observation;
];

rr:predicateObjectMap[
    rr:predicate sltsv:month;
    rr:objectMap [
        rr:termType rr:IRI;
        rr:constant "http://reference.data.gov.uk/def/intervals/{$alias}";
    ];
];

rr:predicateObjectMap[
    rr:predicate sltsv:numberOfArrivals;
    rr:objectMap [
        rr:termType rr:Literal;
        rr:column {$alias};
        rr:datatype xsd:integer;
    ];
];
```

Accessed columns

Dictionary with alias

Reference to columns

1 TriplesMap for 12 Months

Reference to alias

Outputs:

- RMLC-Iterator for transforming the mappings to R2RML
- Morhp-RDB as OBDA engine for the query translation

Results:

D1

| Features | R2RML | RMLC |
|---|---|---|
| Total Lines | ~700 | 74 |
| #TriplesMaps / #SubjectMaps | 12 | 1 |
| #PredicateObjectMaps | 60 | 5 |

D2

| Features | R2RML | RMLC |
|---|---|---|
| Total Lines | >2800 | <70 |
| #TriplesMaps / #SubjectMaps | >40 | 1 |
| #PredicateObjectMaps | >170 | 4 |

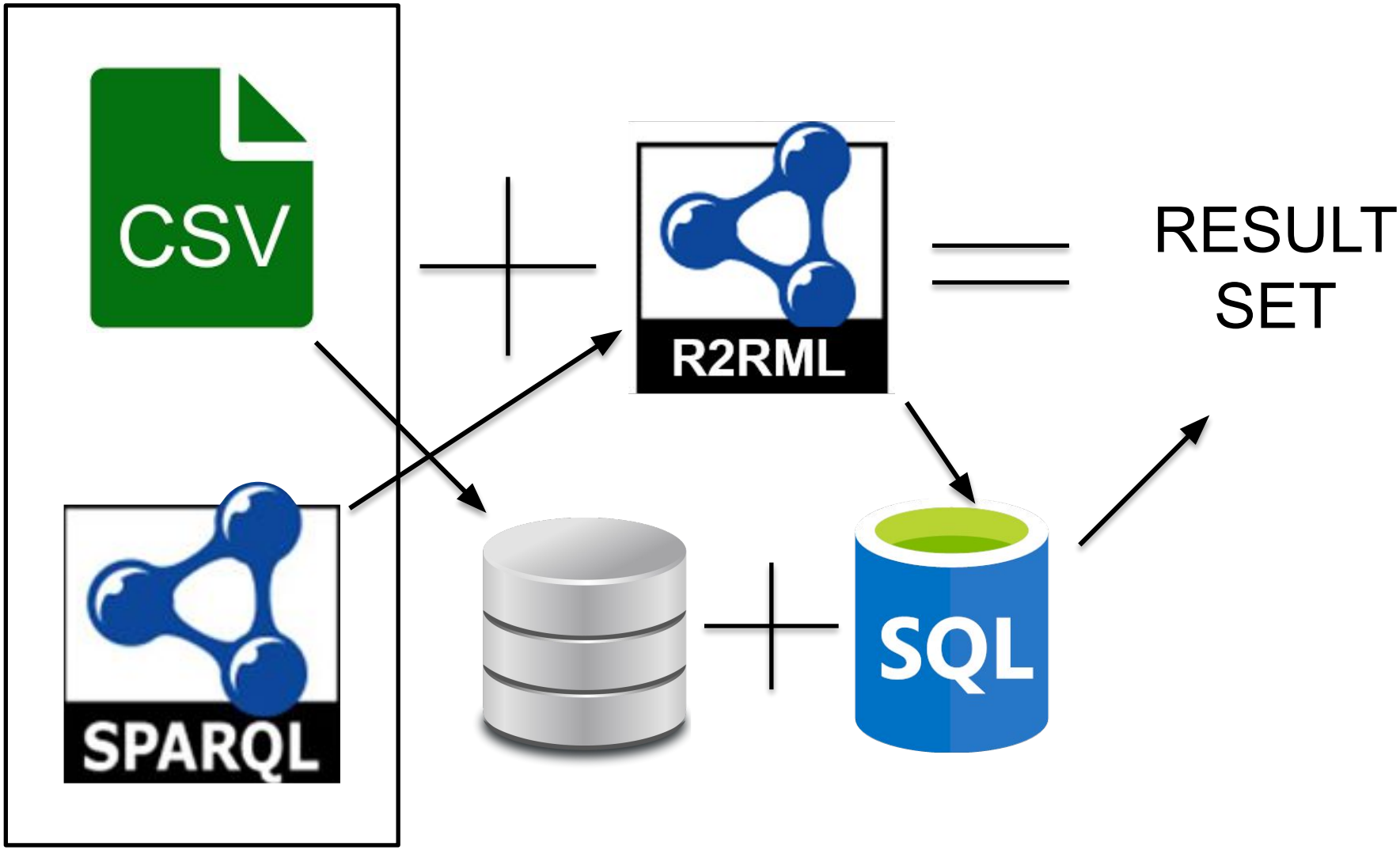

# SATET: Providing access to multiple CSV on the Web using OBDA

**David Chaves-Fraga, Ontology Engineering Group**
**Universidad Politécnica de Madrid, Spain**
Freddy Priyatna, OEG-UPM
Oscar Corcho, OEG-UPM

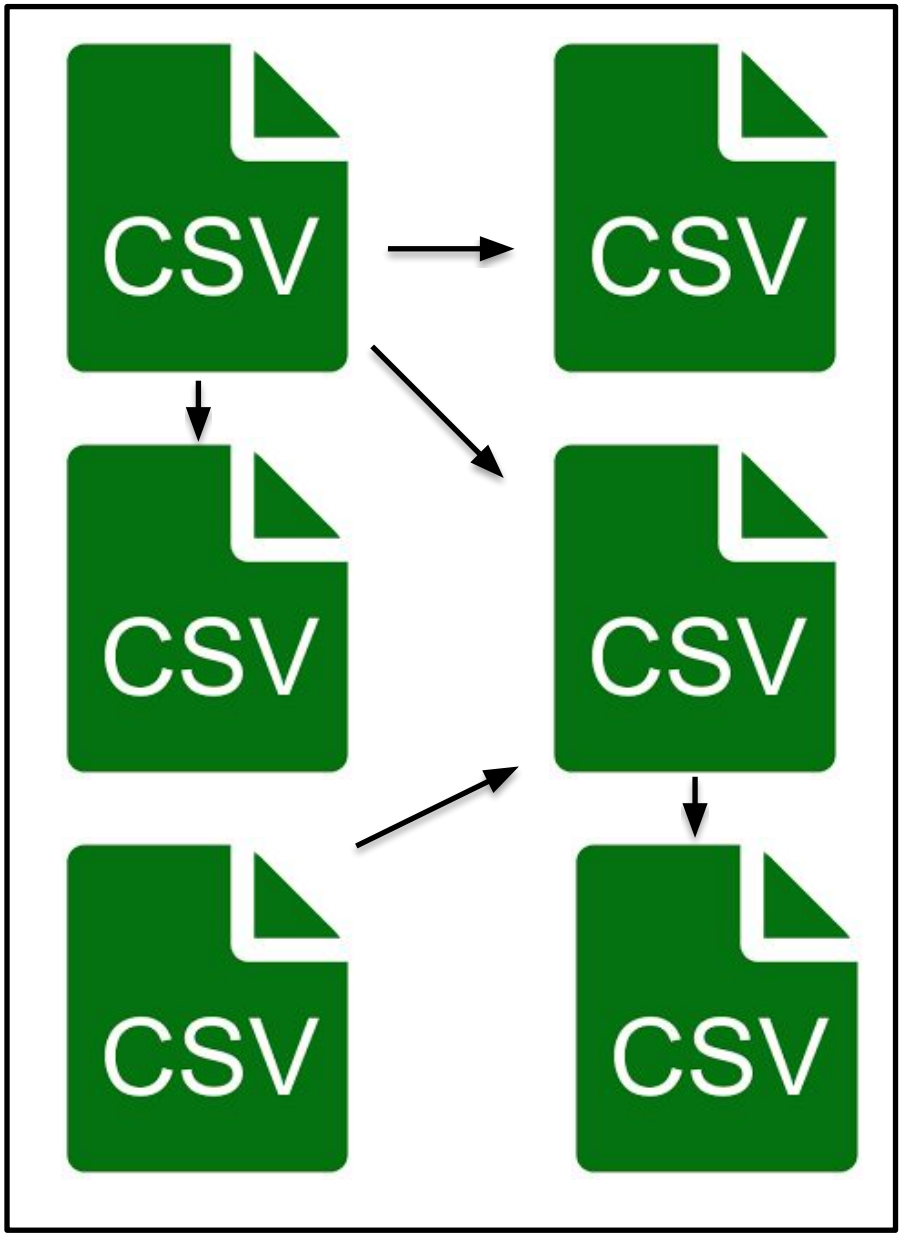
dchaves@fi.upm.es
@dchavesf

???
???

# Semantic Interoperability in CSV files - Virtualizing



# Multiple CSV files with relations

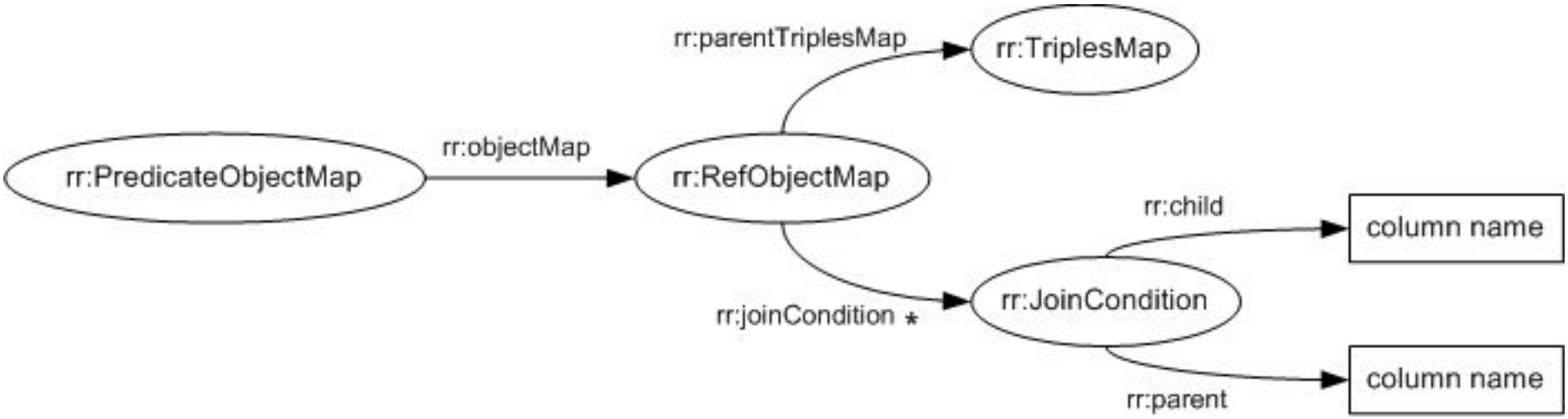



???

# SATET: Semantic Access for heTErogeneous Tabular data

- RMLC: Extension of R2RML for including SQL functions
  - Discover implicit joins among CSV files
  - Transforming CSV columns to RDF objects
- Generation of an enriched database schema using the mapping info (optimization)
- Semantic preservation of R2RML

# Join in R2RML and RML



# Discovering implicit joins between CSV files

Relational Database

```
id,name,surname,birthdate,location
1,david,chaves-fraga,27-11-1993,SDC
```

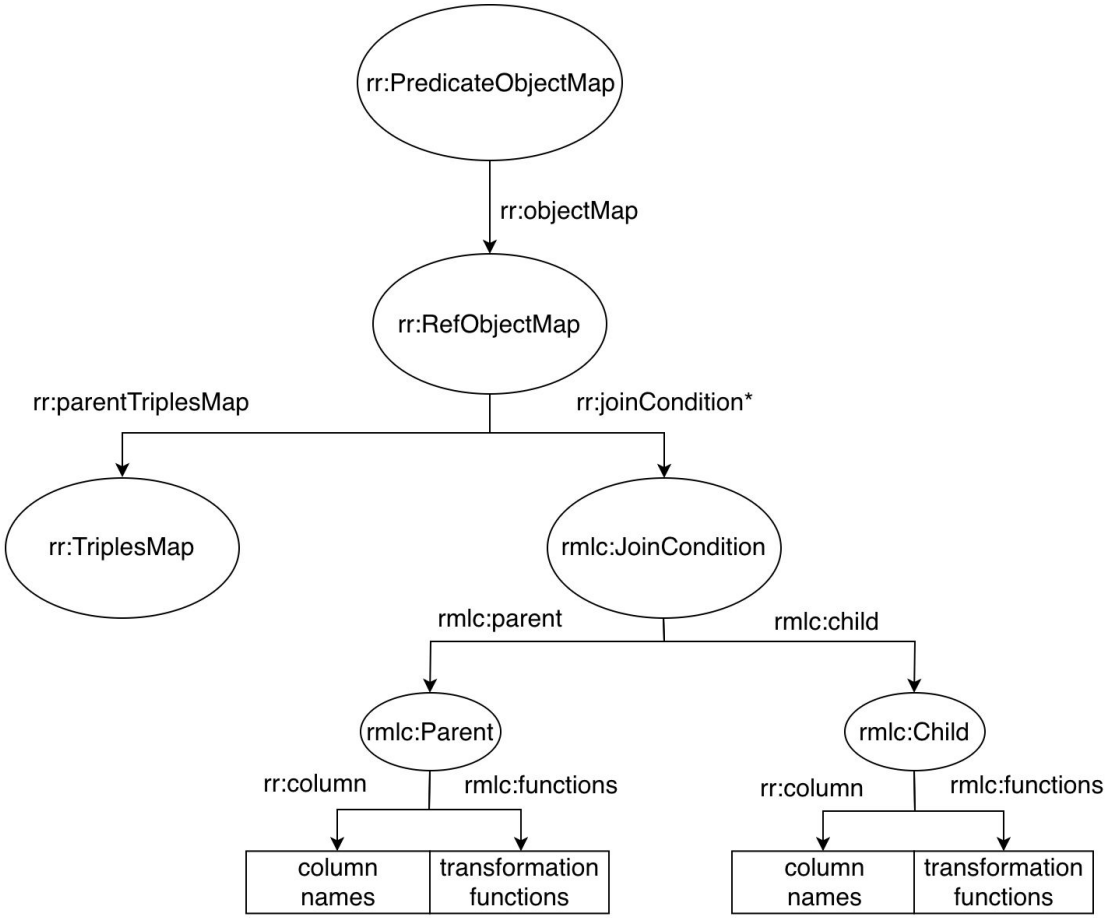
```
id,job
1,phd_student
```

CSV files

```
name,surname,birthdate,location
david,chaves_fraga,27111993,SDC
```

```
full_name,job
"David Chaves Fraga","phd_student"
```

# RMLC: RDF Mapping Language extension for heterogeneous CSV files




The functions are SQL basic transformation functions

Table 1

```
name,surname,birthdate,location
david,chaves_fraga,27111993,SDC
```

Table 2

```
full_name,job
"David Chaves Fraga","phd_student"
```

```
SELECT ?name ?birthday ?job WHERE {
    ?name ?p1 ?birthday.
    ?name ?p2 ?job .
}
```

# Example

```
SELECT ?name ?birthday ?job
WHERE {
      ?name ex:birthday ?birthday.
      ?name ex:job ?job .
}
```

```
<#TriplesMap1>
....
  rr:predicateObjectMap[
    rr:predicate foaf:name;
    rr:objectMap [
      rr:parentTriplesMap <#TriplesMap2>;
      rr:joinCondition [
        rmlc:child [
          rmlc:functions "LOWER({FULL_NAME})";
        ];
        rmlc:parent [
          rmlc:functions “CONCAT({NAME},' ',REPLACE({SURNMAE},'_',' '))";
        ];
      ];
    ];
  ];
.
```

```
SELECT name, birthday, table2.job FROM table1
INNER JOIN table2 ON
CONCAT(table1.name,' ',REPLACE(table1.surname,'_',' ')) =  LOWER(table2.full_name)
```

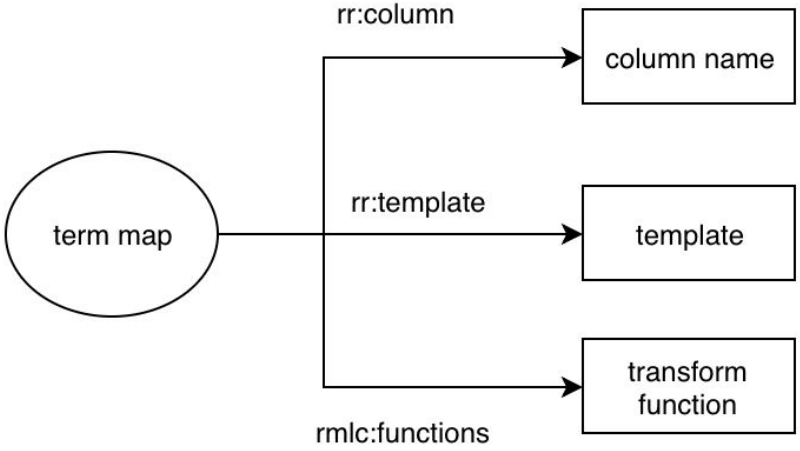
# Example with GraphQL

```
{
  listSocialMediaPosting {
    identifier
    comment
    author {
      identifier
      email
      familyName
      givenName
      name
      telephone
    }
  }
}
```


```
{
  "data": {
    "listSocialMediaPosting": [
      {
        "identifier": "http://ex.org/1",
        "comment": "Hallo Dunia@id",
        "author": {
          "identifier": "http://ex.org/Person/1",
          "email": "fpriyatna@fi.upm.es",
          "familyName": "Priyatna",
          "givenName": "Freddy",
          "name": "Freddy Priyatna",
          "telephone": "8141"
        }
      },
      {
        "identifier": "http://ex.org/2",
        "comment": "Hola Mundo@es",
        "author": {
          "identifier": "http://ex.org/Person/2",
          "email": "dchaves@fi.upm.es",
          "familyName": "Chaves",
          "givenName": "David",
          "name": "David Chaves",
          "telephone": "9063"
        }
      },
      {
        "identifier": "http://ex.org/3",
```


```
SELECT
  "listSocial"."id" AS "id",
  'http://ex.org/' || "listSocial".id || '' AS "identifier",
  "listSocial"."mensaje" AS "comment",
  "author"."id" AS "author__id",
  'http://ex.org/Person/' || "author".id || '' AS "author__identifier",
  lower(substr("author".nombre,1,1) || "author".apellido || '@fi.upm.es') AS "author__email",
  "author"."apellido" AS "author__familyName",
  "author"."nombre" AS "author__givenName",
  '' || "author".nombre || ' ' || "author".apellido || '' AS "author__name",
  "author"."telephone" AS "author__telephone"
FROM comentarios "listSocial"
LEFT JOIN personas "author" ON "listSocial".usuario = lower(substr("author".nombre,1,1) || "author".apellido)
```

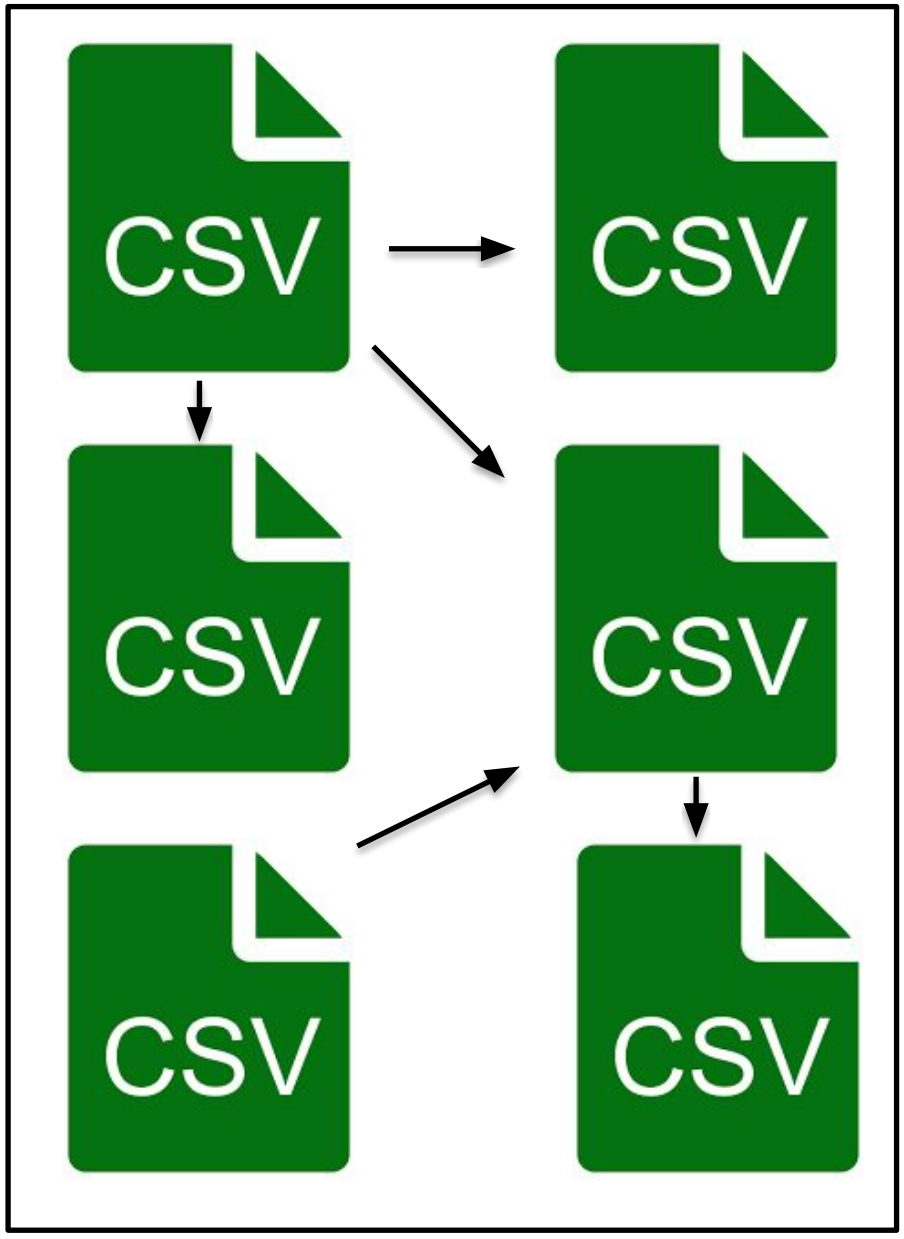
# Transforming CSV columns to RDF objects




```
<TriplesMap1>
...
rr:predicateObjectMap[
    rr:predicate ex:shortName;
    rr:objectMap [
        rr:datatype xsd:string;
        rmlc:functions "REPLACE(SUBSTRING(LOWER{FULL_NAME},1,5),' ','−')";
    ]; ];
rr:predicateObjectMap[
    rr:predicate ex:yearofBirthday;
    rr:objectMap [
        rmlc:functions "YEAR({birthday})";  ]; ];
```

.

# Enriched database schema




\+ RMLC =

SATET ++

- Primary Keys
- Foreign Keys
- Datatypes

# Alignment with R2RML and OBDA engines

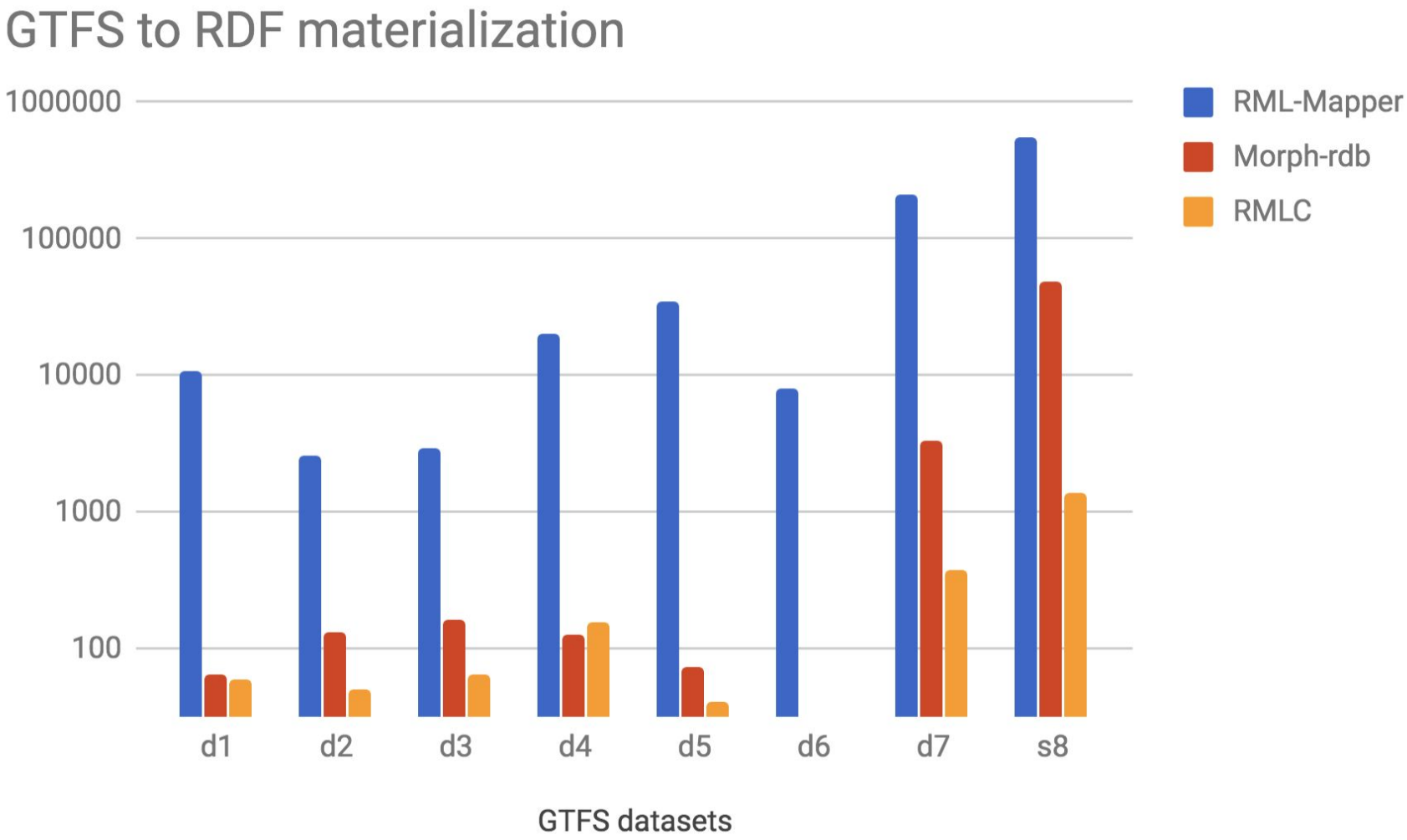
- RMLC maintains the semantics of R2RML
- It's aligned with R2RML:
  - ObjectMaps with Functions → new column in the table with the name of the predicate
  - Joins with Functions → new columns in the tables
  - SATET transforms RMLC to R2RML
- SATET can be introduced on the top of state-of-art OBDA engines (morph/ontop) for using their optimizations to efficiently access to CSV files

# Example

```
SELECT ?name ?birthday ?job
WHERE {
      ?name ex:birthday ?birthday.
      ?name ex:job ?job .
}
```

```
<#TriplesMap1>
....
  rr:predicateObjectMap[
    rr:predicate ex:fullName;
    rr:objectMap [
      rr:parentTriplesMap <#TriplesMap2>;
      rr:joinCondition [
        rmlc:child [
          rmlc:functions "LOWER({FULL_NAME})";
        ];
        rmlc:parent [
          rmlc:functions “CONCAT({NAME},' ',REPLACE({SURNMAE},'_',' '))";
        ];
      ];
    ];
  ];
.
```

```
SELECT name, birthday, table2.job FROM table1
INNER JOIN table2 ON table1.fullName =  table2.fullName
```


GTFS to RDF materialization
1000000
100000
10000
1000
100
d1
d2
d3
d4
d5
d6
d7
s8
GTFS datasets
RML-Mapper
Morph-rdb
RMLC

# SATET: Semantic Access for heTErogeneous Tabular data

Main Contributions:

- Discover implicit joins
- Apply transformation functions to individual columns
- Enriched database schema from mapping information
- Semantic preservation of R2RML

Future Work:

- Alignment with FnO → full specification (Possible collaboration)
- Alignment with RML (without FnO) for LD Generation from RDB/CSV
- Optimizations over generated SQL queries
- Query answering over SATET
- Applying to transport domain for linking potential datasets during a route planning creation
- Define the core for the mapping languages of SW (Possible collaboration)

# Virtual Knowledge Graph Generation from heterogeneous data sources

**David Chaves-Fraga, Ontology Engineering Group**
**Universidad Politécnica de Madrid, Spain**
Freddy Priyatna, OEG-UPM
Oscar Corcho, OEG-UPM

dchaves@fi.upm.es
@dchavesf

24-25/10/2018
@imec/Ghent University