



# RMLC: RDF Mapping Language for heterogeneous CSV files

**David Chaves-Fraga, Ontology Engineering Group**  
**Universidad Politécnica de Madrid, Spain**

Oscar Corcho, OEG-UPM

✉ [dchaves@fi.upm.es](mailto:dchaves@fi.upm.es)

🐦 [@dchavesf](https://twitter.com/dchavesf)

📅 20/09/2018

📍 ETSIINF

OBDA



Ontology Based Data Access



SPARQL translation to query language using mapping  
information = Virtualization

Linked Data Generation



Using mapping information for transforming data sources to  
RDF = Materialization

OBDA



Ontology Based Data Access



SPARQL translation to query language using mapping  
information = Virtualization

Linked Data Generation



Using mapping information for transforming data sources to  
RDF = Materialization



- R2RML:
  - Focused on provide access to RDB
  - Allows materialization and virtualization
  - W3C Standard
- RML / YARRRML:
  - Focused on generated linked data from JSON, CSV or XML
  - Allows only materialization
  - De-Facto Standard
- RMLC:
  - Focused on provide access to CSV
  - Allows virtualization (also materialization)
  - Able to deals with the heterogeneity of the format

What is the most used format for representing and exchanging data along the **WEB**?

What is the most used format for representing and exchanging data along the **WEB**?

| Formato                     |
|-----------------------------|
| CSV (8854)                  |
| XLS (6339)                  |
| JSON (4517)                 |
| HTML (3540)                 |
| XML-APP (2593)              |
| PDF (2094)                  |
| ASCII (1909)                |
| PC-Axis (1302)              |
| RDF-Turtle (1192)           |
| XLSX (1186)                 |
| <a href="#">Mostrar más</a> |

What is the most used format for representing and exchanging data along the **WEB**?

| Formato                     |
|-----------------------------|
| CSV (8854)                  |
| XLS (6339)                  |
| JSON (4517)                 |
| HTML (3540)                 |
| XML-APP (2593)              |
| PDF (2094)                  |
| ASCII (1909)                |
| PC-Axis (1302)              |
| RDF-Turtle (1192)           |
| XLSX (1186)                 |
| <a href="#">Mostrar más</a> |

What is the most used format for representing and exchanging data along the **WEB**?

| Formato                     |
|-----------------------------|
| CSV (8854)                  |
| XLS (6339)                  |
| JSON (4517)                 |
| HTML (3540)                 |
| XML-APP (2593)              |
| PDF (2094)                  |
| ASCII (1909)                |
| PC-Axis (1302)              |
| RDF-Turtle (1192)           |
| XLSX (1186)                 |
| <a href="#">Mostrar más</a> |

| Formats                           |
|-----------------------------------|
| CSV (86834)                       |
| TXT (65699)                       |
| ZIP (51600)                       |
| JSON (45900)                      |
| GMZ (44712)                       |
| HTML (42959)                      |
| PDF (35421)                       |
| XLS (24418)                       |
| WMS (21800)                       |
| SHP (19507)                       |
| <a href="#">Show More Formats</a> |



What is the most used format for representing and exchanging data along the **WEB**?

| Formato                     |
|-----------------------------|
| CSV (8854)                  |
| XLS (6339)                  |
| JSON (4517)                 |
| HTML (3540)                 |
| XML-APP (2593)              |
| PDF (2094)                  |
| ASCII (1909)                |
| PC-Axis (1302)              |
| RDF-Turtle (1192)           |
| XLSX (1186)                 |
| <a href="#">Mostrar más</a> |

| Formats                           |
|-----------------------------------|
| CSV (86834)                       |
| TXT (65699)                       |
| ZIP (51600)                       |
| JSON (45900)                      |
| GMZ (44712)                       |
| HTML (42959)                      |
| PDF (35421)                       |
| XLS (24418)                       |
| WMS (21800)                       |
| SHP (19507)                       |
| <a href="#">Show More Formats</a> |

What is the most used format for representing and exchanging data along the **WEB**?

| Formato                     |
|-----------------------------|
| CSV (8854)                  |
| XLS (6339)                  |
| JSON (4517)                 |
| HTML (3540)                 |
| XML-APP (2593)              |
| PDF (2094)                  |
| ASCII (1909)                |
| PC-Axis (1302)              |
| RDF-Turtle (1192)           |
| XLSX (1186)                 |
| <a href="#">Mostrar más</a> |

| Formats                           |
|-----------------------------------|
| CSV (86834)                       |
| TXT (65699)                       |
| ZIP (51600)                       |
| JSON (45900)                      |
| GMZ (44712)                       |
| HTML (42959)                      |
| PDF (35421)                       |
| XLS (24418)                       |
| WMS (21800)                       |
| SHP (19507)                       |
| <a href="#">Show More Formats</a> |

| Formatos   |
|------------|
| CSV (371)  |
| XLS (169)  |
| XML (129)  |
| XLSX (122) |
| WMS (29)   |
| RDF (21)   |
| prj (7)    |
| SHP (7)    |
| SHX (7)    |
| ZIP (7)    |
| dBase (6)  |
| WMTS (5)   |
| JSON (4)   |
| KML (3)    |
| BDF (2)    |
| DAT (2)    |

What is the most used format for representing and exchanging data along the **WEB**?

| Formato                     |
|-----------------------------|
| CSV (8854)                  |
| XLS (6339)                  |
| JSON (4517)                 |
| HTML (3540)                 |
| XML-APP (2593)              |
| PDF (2094)                  |
| ASCII (1909)                |
| PC-Axis (1302)              |
| RDF-Turtle (1192)           |
| XLSX (1186)                 |
| <a href="#">Mostrar más</a> |

| Formats                           |
|-----------------------------------|
| CSV (86834)                       |
| TXT (65699)                       |
| ZIP (51600)                       |
| JSON (45900)                      |
| GMZ (44712)                       |
| HTML (42959)                      |
| PDF (35421)                       |
| XLS (24418)                       |
| WMS (21800)                       |
| SHP (19507)                       |
| <a href="#">Show More Formats</a> |

| Formatos   |
|------------|
| CSV (371)  |
| XLS (169)  |
| XML (129)  |
| XLSX (122) |
| WMS (29)   |
| RDF (21)   |
| prj (7)    |
| SHP (7)    |
| SHX (7)    |
| ZIP (7)    |
| dBase (6)  |
| WMTS (5)   |
| JSON (4)   |
| KML (3)    |
| BDF (2)    |
| DAT (2)    |

# CSV: Comma-Separated Values File

```
stop_id,stop_name,stop_desc,stop_lat,stop_lon,stop_url
20,Ciutadella,Vila Olímpica,Correspondència amb Metro L4,41.387614,2.191922,http://www.tram.cat/linies-i-horaris/linea-t4/
21,Wellington,,41.390167,2.188483,http://www.tram.cat/linies-i-horaris/linea-t4/
22,Marina,Correspondència amb Metro L1,41.394061,2.186917,http://www.tram.cat/linies-i-horaris/linea-t4/
2,Auditori | Teatre Nacional,,41.397536,2.186892,http://www.tram.cat/linies-i-horaris/linea-t4/
3,Glòries,Correspondència amb TRAM T4-T5-T6 i Metro L1,41.401681,2.186897,http://www.tram.cat/linies-i-horaris/linea-t4/
4,Ca l'Aranyó,,41.404450,2.191500,http://www.tram.cat/linies-i-horaris/linea-t4/
5,Pere IV,,41.406144,2.198247,http://www.tram.cat/linies-i-horaris/linea-t4/
6,Fluvià,,41.407822,2.204908,http://www.tram.cat/linies-i-horaris/linea-t4/
7,Selva de Mar,Correspondència amb Metro L4,41.409194,2.210447,http://www.tram.cat/linies-i-horaris/linea-t4/
8,El Maresme,Correspondència amb Metro L4,41.410489,2.215631,http://www.tram.cat/linies-i-horaris/linea-t4/
9,Fòrum,,41.411828,2.219850,http://www.tram.cat/linies-i-horaris/linea-t4/
10,Campus Diagonal-Besòs,,41.415247,2.222928,http://www.tram.cat/linies-i-horaris/linea-t4/
29,Port Fòrum,Correspondència amb TRAM T4-T6,41.418872,2.226161,http://www.tram.cat/linies-i-horaris/linea-t4/
30,Estació de Sant Adrià,Correspondència amb TRAM T4-T6 i Rodalies,41.423953,2.230550,http://www.tram.cat/linies-i-horaris/linea-t4/
3bis,La Farinera,Correspondència amb TRAM T5-T6,41.404200,2.188581,http://www.tram.cat/linies-i-horaris/linea-t5/
23,Can Jaumandreu,Correspondència amb TRAM T5-T6,41.408108,2.193597,http://www.tram.cat/linies-i-horaris/linea-t5/
24,Espronceda,Correspondència amb TRAM T5-T6,41.411361,2.197942,http://www.tram.cat/linies-i-horaris/linea-t5/
25,Sant Martí de Provençals,Correspondència amb TRAM T5-T6,41.416464,2.204744,http://www.tram.cat/linies-i-horaris/linea-t5/
26,Besòs,Correspondència amb TRAM T5-T6 i Metro L4,41.420067,2.209556,http://www.tram.cat/linies-i-horaris/linea-t5/
27,Alfons el Magnànim,Correspondència amb TRAM T5-T6,41.417194,2.213569,http://www.tram.cat/linies-i-horaris/linea-t5/
28,Parc del Besòs,Correspondència amb TRAM T5-T6,41.419431,2.217633,http://www.tram.cat/linies-i-horaris/linea-t5/
13,La Catalana,,41.422650,2.221983,http://www.tram.cat/linies-i-horaris/linea-t5/
14,Sant Joan Baptista,,41.427386,2.225403,http://www.tram.cat/linies-i-horaris/linea-t5/
16,Encants de Sant Adrià,Correspondència amb Metro L2,41.430244,2.222606,http://www.tram.cat/linies-i-horaris/linea-t5/
17,Sant Roc,Correspondència amb Metro L2,41.434944,2.227597,http://www.tram.cat/linies-i-horaris/linea-t5/
19,Gorg,Correspondència amb Metro L2 i 10,41.439967,2.233219,http://www.tram.cat/linies-i-horaris/linea-t5/
11,La Mina,,41.418386,2.221000,http://www.tram.cat/linies-i-horaris/linea-t6/
```

|   | A       | B         | C                 | D                | E                      | F   | G  | H   | I        |
|---|---------|-----------|-------------------|------------------|------------------------|---|--|---|----------|
| 1 | First   | Last      | Email             | Title            | Company                | Comment   | Phone                                      | Website   | Fax      |
| 2 | Relenta | Helpdesk  | helpdesk@relenta  | Customer service | Relenta                | Nice guys.  |  | <a href="http://www.relenta.com">www.relenta.com</a>  |          |
| 3 | Agent   | Basil     | SpecialAgentBasil | Special Agent    | Ministry of Defense    | DO NOT OFFEND HIS MOTHER!   | 555-1212 (Office)                          | <a href="http://Match.com/Agent">Match.com/Agent</a>  | 555-1235 |
| 4 | James   | Bond      | SirJamesBondDou   | Agent 007        | British Secret Service | 007 on an island populated exclusively by women? We won't see him till dawn!      | 555-1239 (Mobile);+44 43593434534 (Cell)   | <a href="http://www.mi6.uk.gov">www.mi6.uk.gov</a> (Work); <a href="https://relentademo.fres">https://relentademo.fres</a> (FreshBooks) |          |
| 5 | Queen   | Elizabeth | queen.elizabeth@  | Queen            | England                | +44 (0) 23445645937   |  |   |          |
| 6 | Rowan   | Atkinson  | RowYourBoat@Ac    | Actor            |                        | Johnny English was not funny. Really proud about defeating the Nazi's in W W two. | 555-1235                                   | <a href="http://www.TheBestActo">www.TheBestActo</a>  |          |
| 7 | Douglas | Gilmour   | BombsAway@US:     | Commander        | U.S. Strategic Command |   | 555-1237 (Car Phone);555-1238 (Home Phone) |   |          |

This is a screenshot of a CSV file in Microsoft Excel 2016 for Fileinfo.com

# CSV: Comma-Separated Values File

```
stop_id,stop_name,stop_desc,stop_lat,stop_lon,stop_url
20,Ciutadella,Vila Olímpica,Correspondència amb Metro L4,41.387614,2.191922,http://www.tram.cat/linies-i-horaris/linea-t4/
21,Wellington,,41.390167,2.188483,http://www.tram.cat/linies-i-horaris/linea-t4/
22,Marina,Correspondència amb Metro L1,41.394061,2.186917,http://www.tram.cat/linies-i-horaris/linea-t4/
2,Auditori | Teatre Nacional,,41.397536,2.186892,http://www.tram.cat/linies-i-horaris/linea-t4/
3,Glòries,Correspondència amb TRAM T4-T5-T6 i Metro L1,41.401681,2.186897,http://www.tram.cat/linies-i-horaris/linea-t4/
4,Ca l'Aranyó,,41.404450,2.191500,http://www.tram.cat/linies-i-horaris/linea-t4/
5,Pere IV,,41.406144,2.198247,http://www.tram.cat/linies-i-horaris/linea-t4/
6,Fluvià,,41.407822,2.204908,http://www.tram.cat/linies-i-horaris/linea-t4/
7,Selva de Mar,Correspondència amb Metro L4,41.409194,2.210447,http://www.tram.cat/linies-i-horaris/linea-t4/
8,El Maresme,Correspondència amb Metro L4,41.410489,2.215631,http://www.tram.cat/linies-i-horaris/linea-t4/
9,Fòrum,,41.411828,2.219850,http://www.tram.cat/linies-i-horaris/linea-t4/
10,Campus Diagonal-Besòs,,41.415247,2.222928,http://www.tram.cat/linies-i-horaris/linea-t4/
29,Port Fòrum,Correspondència amb TRAM T4-T6,41.418872,2.226161,http://www.tram.cat/linies-i-horaris/linea-t4/
30,Estació de Sant Adrià,Correspondència amb TRAM T4-T6 i Rodalies,41.423953,2.230550,http://www.tram.cat/linies-i-horaris/linea-t4/
3bis,La Farinera,Correspondència amb TRAM T5-T6,41.404200,2.188581,http://www.tram.cat/linies-i-horaris/linea-t5/
23,Can Jaumandreu,Correspondència amb TRAM T5-T6,41.408108,2.193597,http://www.tram.cat/linies-i-horaris/linea-t5/
24,Espronceda,Correspondència amb TRAM T5-T6,41.411361,2.197942,http://www.tram.cat/linies-i-horaris/linea-t5/
25,Sant Martí de Provençals,Correspondència amb TRAM T5-T6,41.416464,2.204744,http://www.tram.cat/linies-i-horaris/linea-t5/
26,Besòs,Correspondència amb TRAM T5-T6 i Metro L4,41.420067,2.209556,http://www.tram.cat/linies-i-horaris/linea-t5/
27,Alfons el Magnànim,Correspondència amb TRAM T5-T6,41.417194,2.213569,http://www.tram.cat/linies-i-horaris/linea-t5/
28,Parc del Besòs,Correspondència amb TRAM T5-T6,41.419431,2.217633,http://www.tram.cat/linies-i-horaris/linea-t5/
13,La Catalana,,41.422650,2.221983,http://www.tram.cat/linies-i-horaris/linea-t5/
14,Sant Joan Baptista,,41.427386,2.225403,http://www.tram.cat/linies-i-horaris/linea-t5/
16,Encants de Sant Adrià,Correspondència amb Metro L2,41.430244,2.222606,http://www.tram.cat/linies-i-horaris/linea-t5/
17,Sant Roc,Correspondència amb Metro L2,41.434944,2.227597,http://www.tram.cat/linies-i-horaris/linea-t5/
19,Gorg,Correspondència amb Metro L2 i 10,41.439967,2.233219,http://www.tram.cat/linies-i-horaris/linea-t5/
11,La Mina,,41.418386,2.221000,http://www.tram.cat/linies-i-horaris/linea-t6/
```

|   | A       | B         | C                 | D                | E                      | F   | G  | H   | I        |
|---|---------|-----------|-------------------|------------------|------------------------|---|--|---|----------|
| 1 | First   | Last      | Email             | Title            | Company                | Comment   | Phone                                      | Website   | Fax      |
| 2 | Relenta | Helpdesk  | helpdesk@relenta  | Customer service | Relenta                | Nice guys.  |  | <a href="http://www.relenta.com">www.relenta.com</a>  |          |
| 3 | Agent   | Basil     | SpecialAgentBasil | Special Agent    | Ministry of Defense    | DO NOT OFFEND HIS MOTHER!   | 555-1212 (Office)                          | <a href="http://Match.com/Agent">Match.com/Agent</a>  | 555-1235 |
| 4 | James   | Bond      | SirJamesBondDou   | Agent 007        | British Secret Service | 007 on an island populated exclusively by women? We won't see him till dawn!      | 555-1239 (Mobile);+44 43593434534 (Cell)   | <a href="http://www.mi6.uk.gov">www.mi6.uk.gov</a> (Work); <a href="https://relentademo.fres">https://relentademo.fres</a> (FreshBooks) |          |
| 5 | Queen   | Elizabeth | queen.elizabeth@  | Queen            | England                | +44 (0) 23445645937   |  |   |          |
| 6 | Rowan   | Atkinson  | RowYourBoat@Ac    | Actor            |                        | Johnny English was not funny. Really proud about defeating the Nazi's in W W two. | 555-1235                                   | <a href="http://www.TheBestActo">www.TheBestActo</a>  |          |
| 7 | Douglas | Gilmour   | BombsAway@US      | Commander        | U.S. Strategic Command |   | 555-1237 (Car Phone);555-1238 (Home Phone) |   |          |

This is a screenshot of a CSV file in Microsoft Excel 2016 for Fileinfo.com.

Syntactic Interoperability

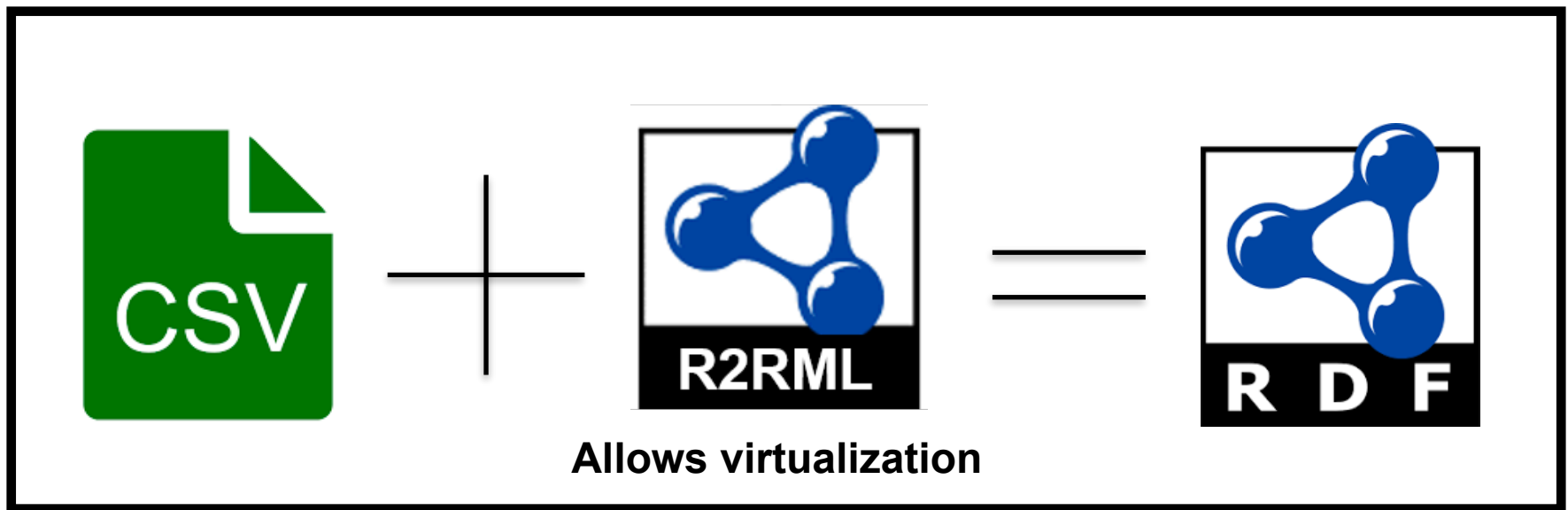


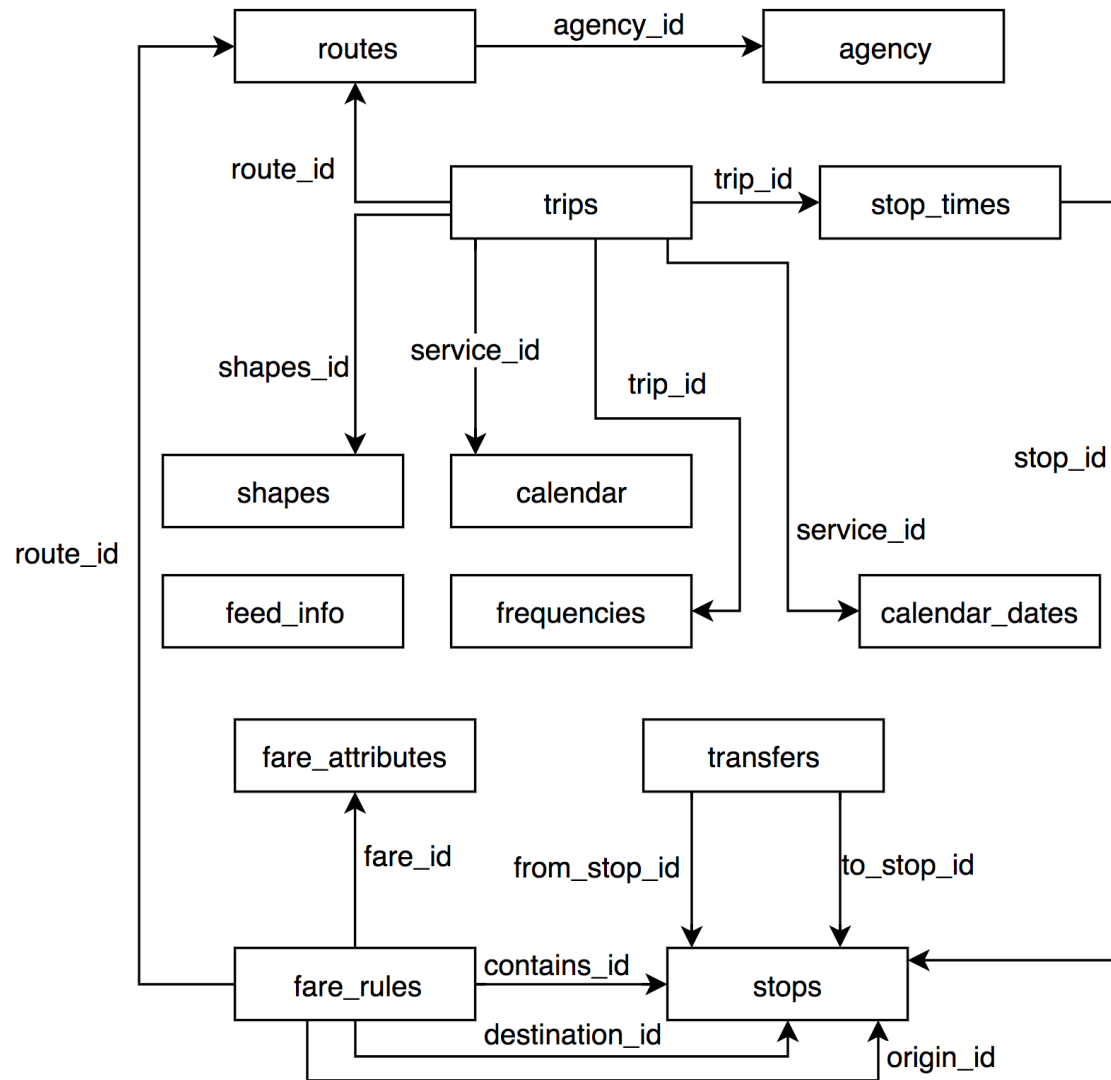












| LinkedGTFS        | Morph - R2RML | RML-Mapper     |
|-------------------|---------------|----------------|
| <b>caceres</b>    | 0,004         | <b>3,739</b>   |
| <b>crtm-metro</b> | 0,026         | <b>2,587</b>   |
| <b>barna-tbs</b>  | 0,068         | 0,786          |
| <b>barna-tbx</b>  | 0,118         | 0,778          |
| <b>crtm-tram</b>  | 0,155         | <b>7,028</b>   |
| <b>crtm-train</b> | 0,217         | <b>12,218</b>  |
| <b>barna-amb</b>  | <b>1,153</b>  | <b>151,541</b> |
| <b>crtm-emt</b>   | <b>12,496</b> | <b>5,118</b>   |

| Data format | Query Engine | Mapping Language   | Virtualization       |
|-------------|--------------|--------------------|----------------------|
| JSON        | MongoDB      | RML                | No                   |
| XML         | XPATH        | RML                | No                   |
| RDB         | SQL          | R2RML              | Yes                  |
| CSV         | ???????      | RML/R2RML/<br>RMLC | Yes (R2RML/<br>RMLC) |



- RMLC: RDF Mapping languages for heterogeneous CSV files
- Two contributions:
  - “Semantic exploitation of implicit joins among heterogeneous CSV files with RMLC” D. Chaves-Fraga and O. Corcho (Expected to ESWC2019)
  - “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)

- RMLC: RDF Mapping languages for heterogeneous CSV files
- Two contributions:
  - “Semantic exploitation of implicit joins among heterogeneous CSV files with RMLC” D. Chaves-Fraga and O. Corcho (Expected to ESWC2019)
  - “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)



Part of the book “Emerging Topics in Semantic Technologies. ISWC 2018 Satellite Events. E. Demidova, A.J. Zaveri, E. Simperl (Eds.), ISBN: 978-3-89838-736-1, 2018, AKA Verlag Berlin”.



# Virtual Statistics Knowledge Graph Generation from CSV files

**Idafen Santana-Perez, Ontology Engineering Group  
Universidad Politécnica de Madrid, Spain**

David Chaves-Fraga, OEG (UPM)

Freddy Priyatna, OEG (UPM)

Oscar Corcho, OEG (UPM)

 [isantana@fi.upm.es](mailto:isantana@fi.upm.es)

 @isantana

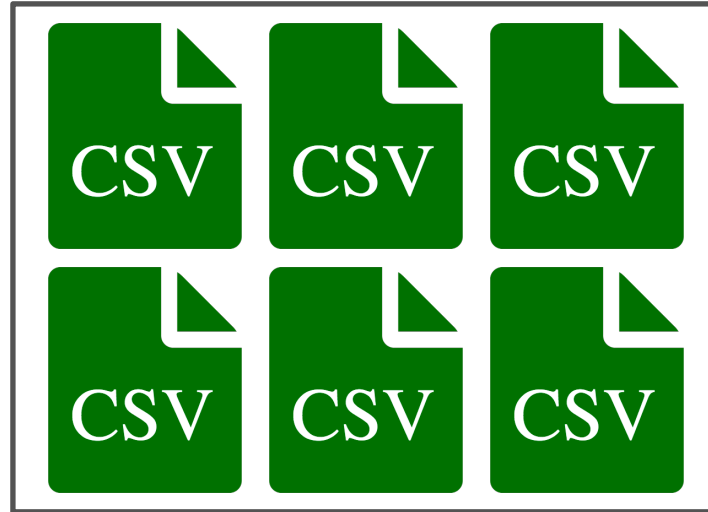
 08/10/2018

 SemStats-ISWC18





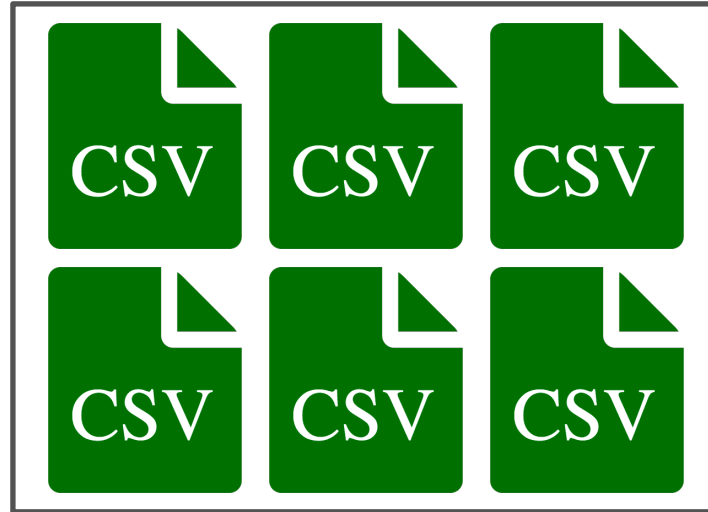
?



Statistics data  
available on  
the Web



?



Statistics data  
available on  
the Web

**Transformation of the CSV files to RDF (e.g. OpenRefine) and load in a Triple Store**

Problems:

1. It is an ad-hoc process
2. How is the data maintained?
3. Who maintains the data updated?
4. What happens if we want to involve other dataset in the process?
5. What happens if the data is volatile?

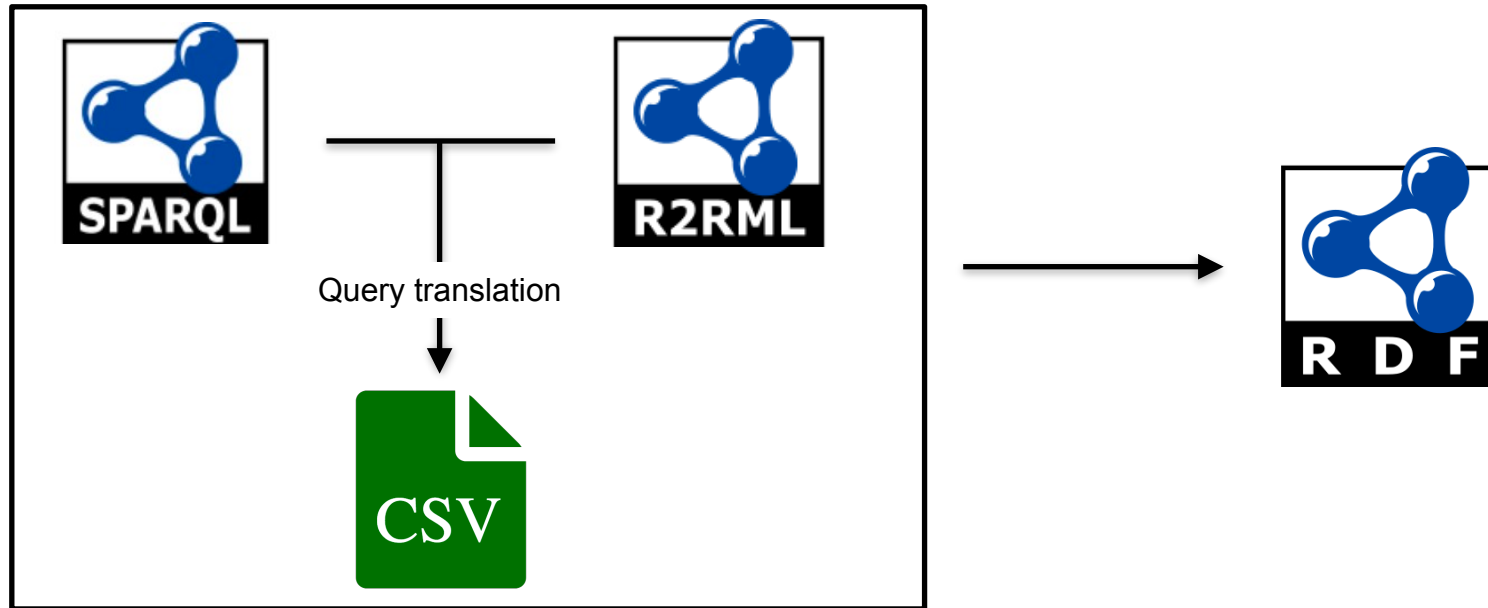
Virtualization: SPARQL translation to source query engine using mapping info

Why is virtualization a good solution?

- General purpose based on mapping standard (R2RML)
- Removes the need of synchronization between CSV and SKG
- Only mapping maintenance is needed
- Provides a way to materialize the data (CSV2RDF)

| Features        | Ad-Hoc            | R2RML           |
|-----------------|-------------------|-----------------|
| Processor Type  | Solution Specific | General Purpose |
| # Processors    | 1                 | Many            |
| Materialization | Yes               | Yes             |
| Virtualization  | No                | Yes             |

## OBDA Engine



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



Difficulty of maintenance and creation

- Approach 1 (Base Line): Use R2RML to answer SPARQL over Virtual SKG
- Approach 2: Reduce the size of the R2RML mapping using RMLC-Iterator and answer SPARQL queries over Virtual SKG.

- R2RML standard approach for querying SKG
- One TriplesMap for each column corresponding to a slice of a dimension.
- Each TriplesMap has:
  - A rr:logicalTable property specifies the source
  - A rr:subjectMap specifies qb:Observation as the generated triples' RDF type
  - A set of rr:predicateObjectMap mappings that specify a slice of a dimension and its values
  - A rr:predicateObjectMap that specifies which dataset the generated triples belongs to

```
<TriplesMapJanuary>
```

```
rr:logicalTable [
  rr:tableName "Statistics2016"
];

rr:subjectMap [
  a rr:Subject; rr:template "www.ex.com/January";
  rr:class qb:Observation;
];

rr:predicateObjectMap[
  rr:predicate ex:month;
  rr:objectMap [ rr:constant "interval:January"; ];
];

rr:predicateObjectMap[
  rr:predicate ex:numberOfArrivals;
  rr:objectMap [ rr:column "Jan"; ];
];

rr:predicateObjectMap[
  rr:predicate qb:dataSet;
  rr:objectMap [ rr:constant "ex:Arrivals"; ];
];
```

<TriplesMapJanuary>

```
rr:logicalTable [
  rr:tableName "Statistics2016"
];
```



Source specification

```
rr:subjectMap [
  a rr:Subject; rr:template "www.ex.com/January";
  rr:class qb:Observation;
];
```

```
rr:predicateObjectMap[
  rr:predicate ex:month;
  rr:objectMap [ rr:constant "interval:January"; ];
];
```

```
rr:predicateObjectMap[
  rr:predicate ex:numberOfArrivals;
  rr:objectMap [ rr:column "Jan"; ];
];
```

```
rr:predicateObjectMap[
  rr:predicate qb:dataSet;
  rr:objectMap [ rr:constant "ex:Arrivals"; ];
];
```



```
<TriplesMapJanuary>
```

```
rr:logicalTable [
  rr:tableName "Statistics2016"
];
```

```
rr:subjectMap [
  a rr:Subject; rr:template "www.ex.com/January";
  rr:class qb:Observation;
];
```

→ SubjectMap for observations

```
rr:predicateObjectMap[
  rr:predicate ex:month;
  rr:objectMap [ rr:constant "interval:January"; ];
];
```

```
rr:predicateObjectMap[
  rr:predicate ex:numberOfArrivals;
  rr:objectMap [ rr:column "Jan"; ];
];
```

```
rr:predicateObjectMap[
  rr:predicate qb:dataSet;
  rr:objectMap [ rr:constant "ex:Arrivals"; ];
];
```

```
<TriplesMapJanuary>
```

```
rr:logicalTable [
  rr:tableName "Statistics2016"
];
```

```
rr:subjectMap [
  a rr:Subject; rr:template "www.ex.com/January";
  rr:class qb:Observation;
];
```

```
rr:predicateObjectMap[
  rr:predicate ex:month;
  rr:objectMap [ rr:constant "interval:January"; ];
];
```

```
rr:predicateObjectMap[
  rr:predicate ex:numberOfArrivals;
  rr:objectMap [ rr:column "Jan"; ];
];
```

```
rr:predicateObjectMap[
  rr:predicate qb:dataSet;
  rr:objectMap [ rr:constant "ex:Arrivals"; ];
];
```

→ Set of predicateObjectMaps

<TriplesMapJanuary>

```
rr:logicalTable [
  rr:tableName "Statistics2016"
];

rr:subjectMap [
  a rr:Subject; rr:template "www.ex.com/January";
  rr:class qb:Observation;
];

rr:predicateObjectMap[
  rr:predicate ex:month;
  rr:objectMap [ rr:constant "interval:January"; ];
];

rr:predicateObjectMap[
  rr:predicate ex:numberOfArrivals;
  rr:objectMap [ rr:column "Jan"; ];
];
```

```
rr:predicateObjectMap[
  rr:predicate qb:dataSet;
  rr:objectMap [ rr:constant "ex:Arrivals"; ];
];
```

→ Dataset predicateObjectMap

```
<TriplesMapJanuary>
```

```
rr:logicalTable [
  rr:tableName "Statistics2016"
];

rr:subjectMap [
  a rr:Subject; rr:template "www.ex.com/January";
  rr:class qb:Observation;
];

rr:predicateObjectMap[
  rr:predicate ex:month;
  rr:objectMap [ rr:constant "interval:January"; ];
];

rr:predicateObjectMap[
  rr:predicate ex:numberOfArrivals;
  rr:objectMap [ rr:column "Jan"; ];
];

rr:predicateObjectMap[
  rr:predicate qb:dataSet;
  rr:objectMap [ rr:constant "ex:Arrivals"; ];
];
```

1 TriplesMap for each Month



12 Similar TriplesMap

- Extension of RMLC, an RDF Mapping Language for heterogeneous CSV files.
- Four new properties in the Logical Table:
  - `rmlc:columns`
  - `rmlc:columnRange`
  - `rmlc:dictionaryFile`
  - `rmlc:dictionary`
- Two variables for identifying independently each TriplesMap and provide access to the CSV data
  - `{column}`
  - `{alias}`
- RMLC-Iterator can easily be converted to R2RML

```

<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;
  ];
];

```

```

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

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

```

Accessed columns

```

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

```

→ Dictionary with alias

```

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;
  ];
];

```



```
<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;
```

```
  ];
```

```
];
```

Reference to columns

```
<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;
  ];
];
```

Reference to alias

```

<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;
  ];
];

```

## 1 TriplesMap for 12 Months

## Tools:

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

## Datasets and queries:

- D1: Statistics from the Sri Lanka Tourism Development Authority
- D2: EuroStat - Immigration Statistics
- Q1/Q2: 6 SPARQL queries

All data, mappings, queries and tools are available at:

<https://github.com/oeg-upm/rmlc-statistic>

## Statistics from the Sri Lanka Tourism Development Authority

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

## EuroStat - Immigration Statistics

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

- We identify the size problem of R2RML mappings in statistics data.
- **RMLC-Iterator drastically reduces the size of the mapping** improving its maintenance and creation time.
- **RMLC-Iterator is aligned to R2RML**
- We provide a tool for transforming RMLC-Iterator to R2RML
- We follow a **virtualization approach avoiding synchronization** between CSV and RDF data.
- The performance of the virtualization process is not affected



# Semantic exploitation of implicit joins among heterogeneous CSV files with RMLC

**David Chaves-Fraga, Ontology Engineering Group**  
**Universidad Politécnica de Madrid, Spain**

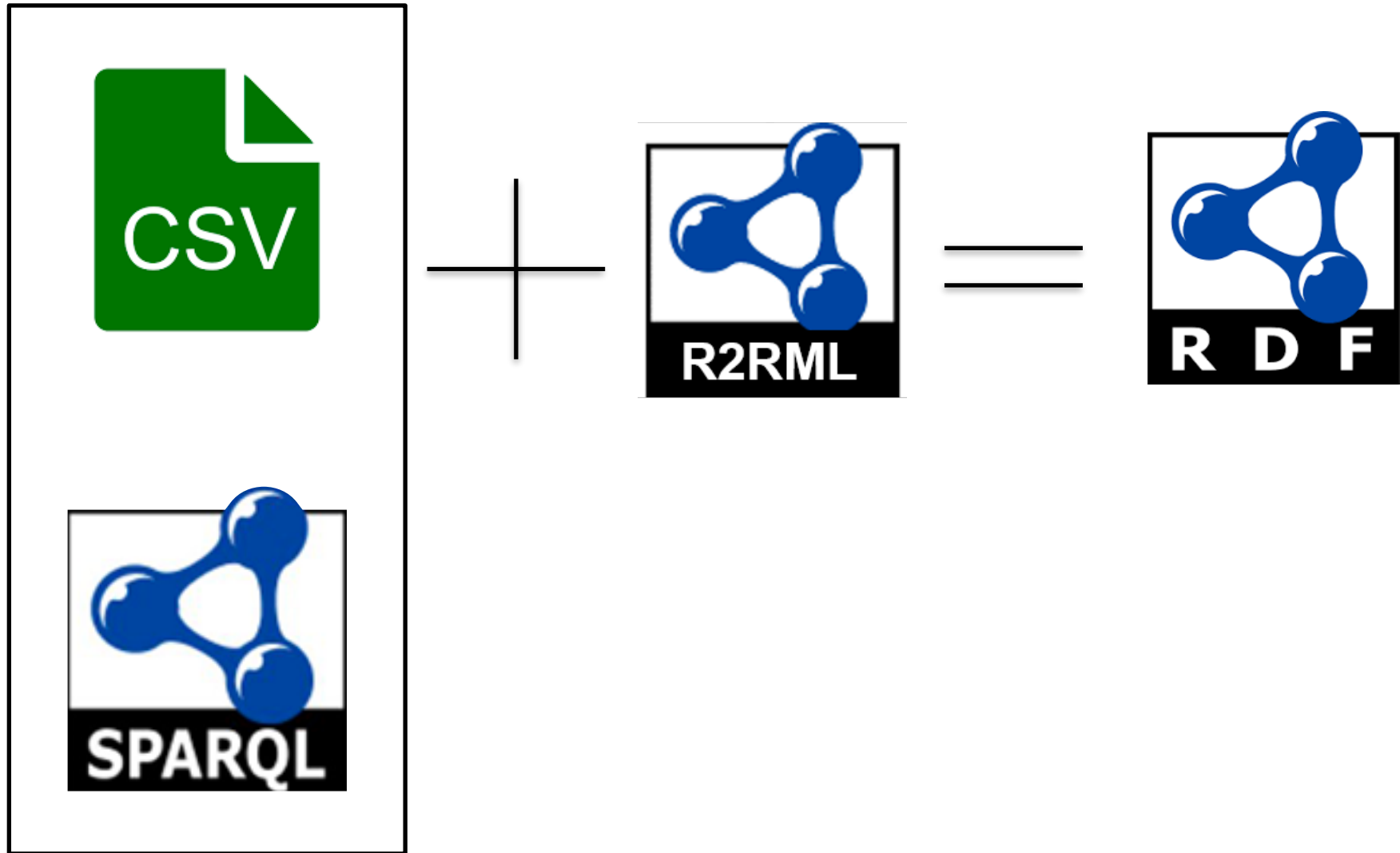
Oscar Corcho, OEG-UPM

✉ [dchaves@fi.upm.es](mailto:dchaves@fi.upm.es)

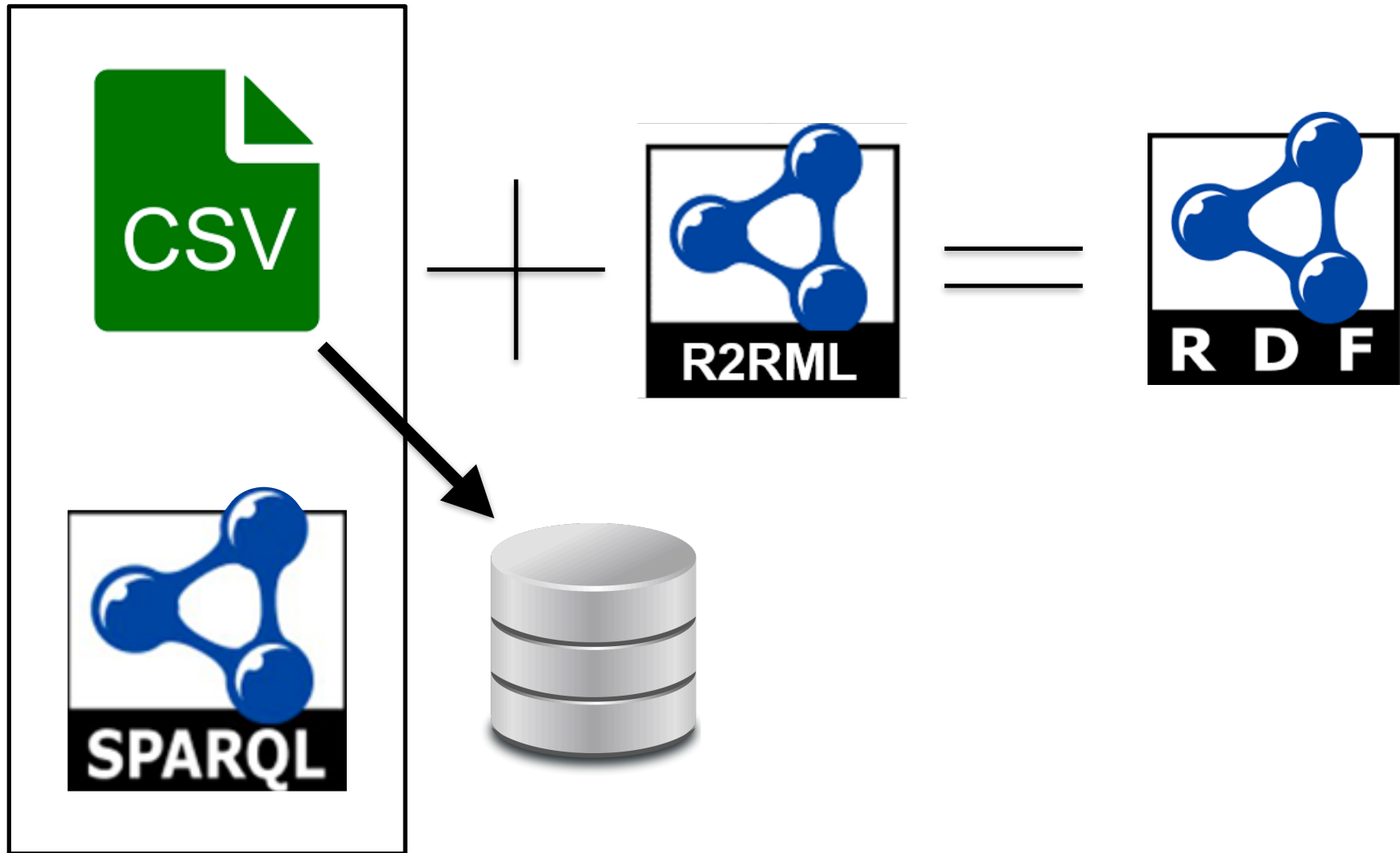
🐦 [@dchavesf](https://twitter.com/dchavesf)

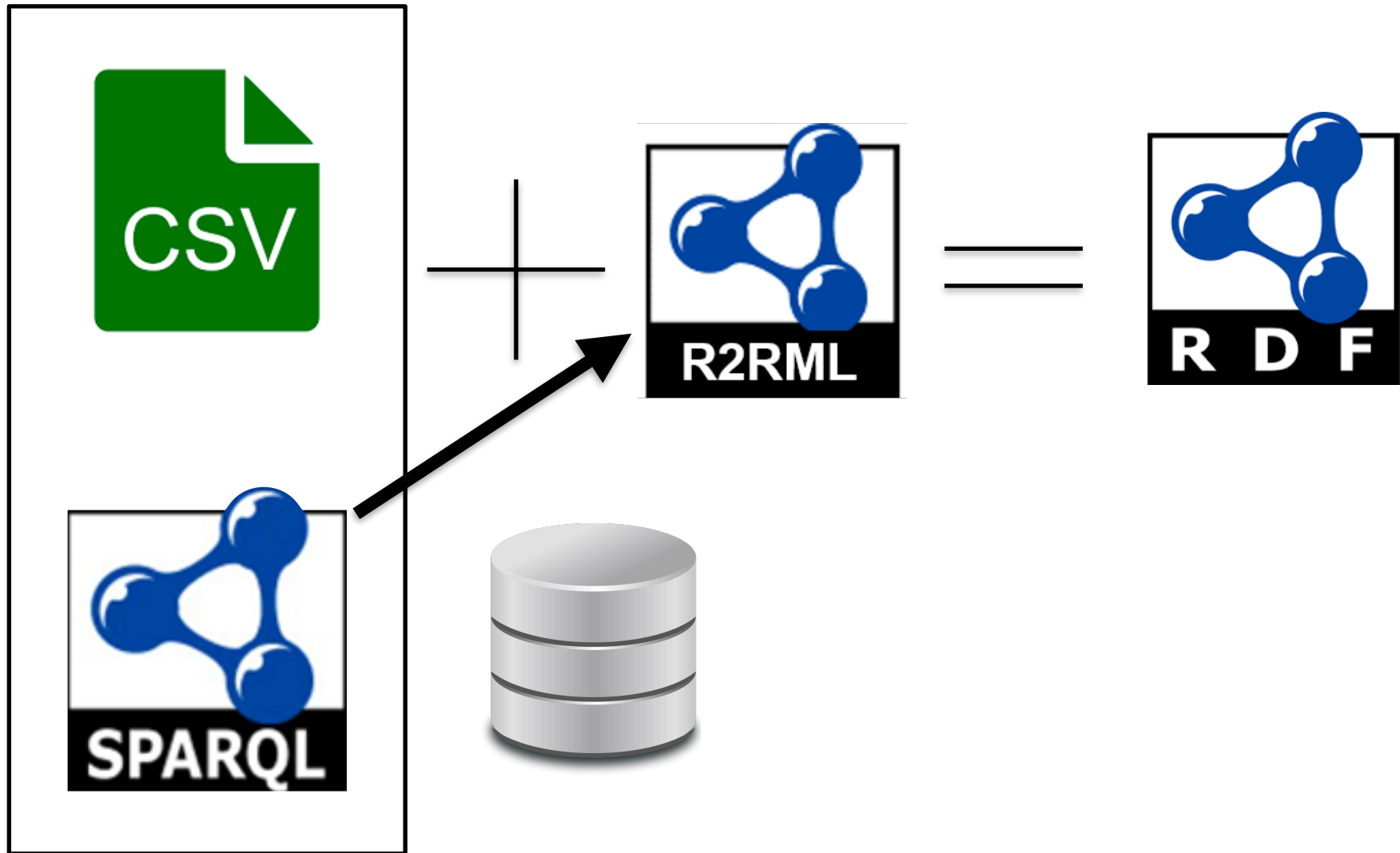
📅 20/09/2018

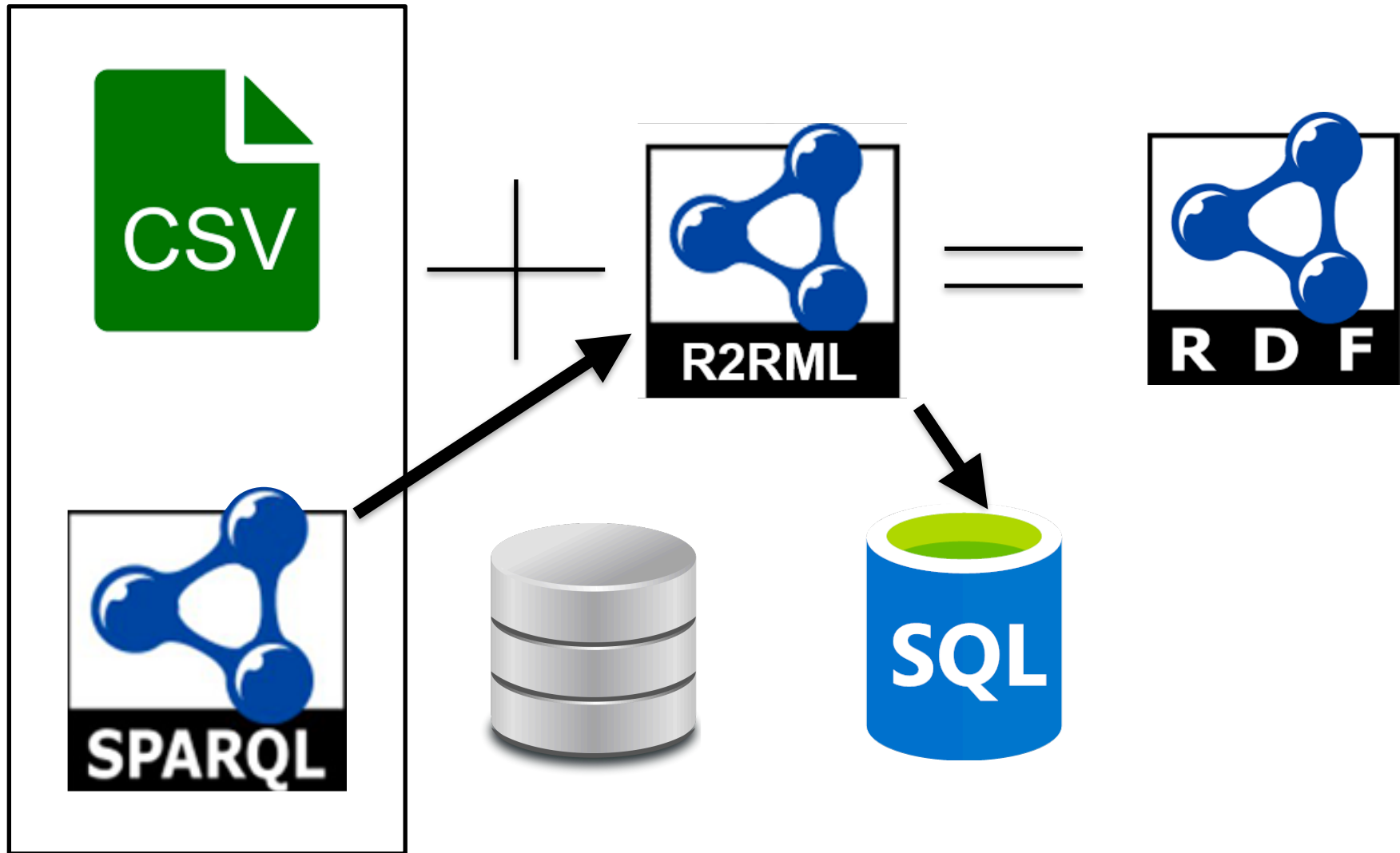
📍 ETSIINF

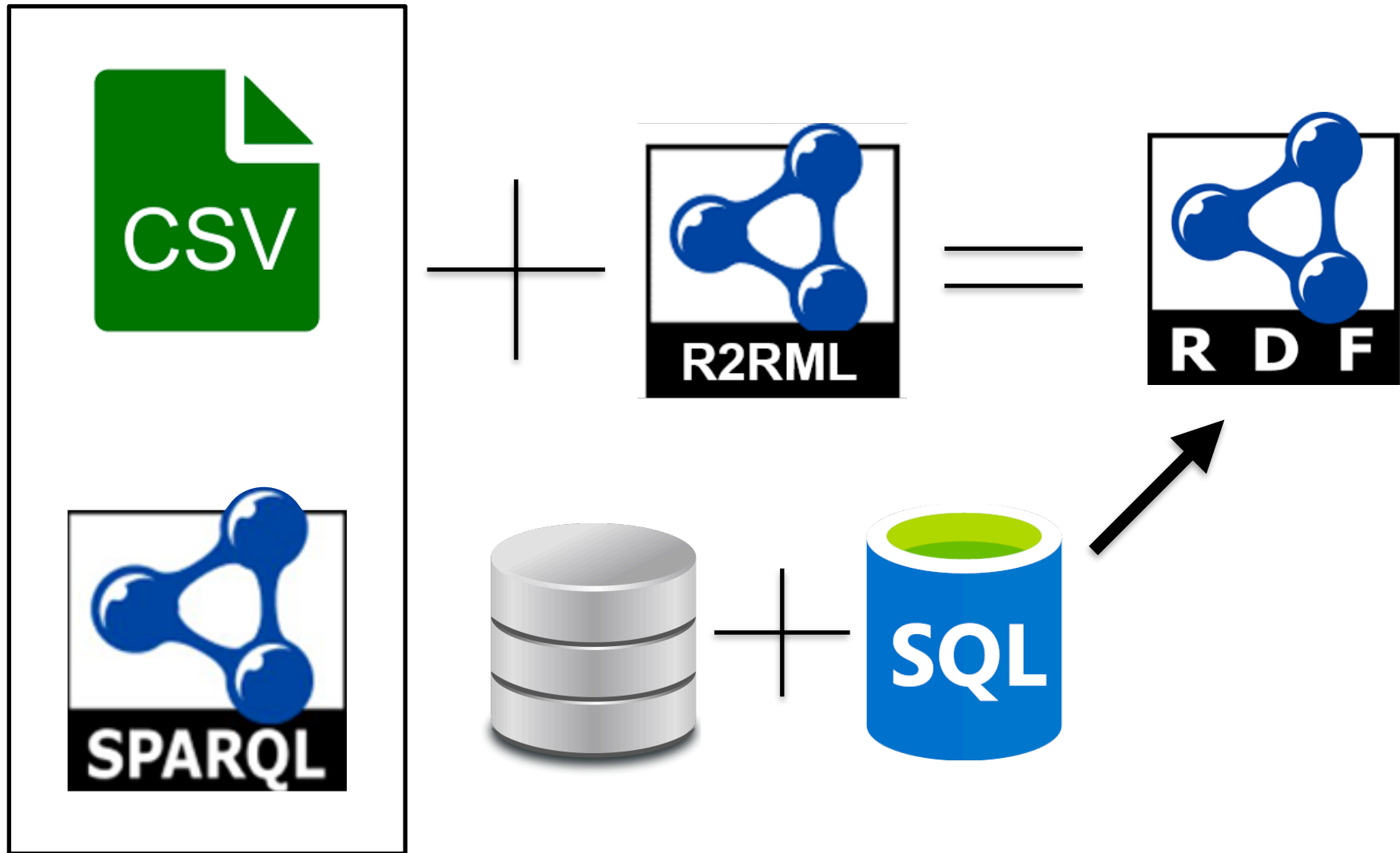


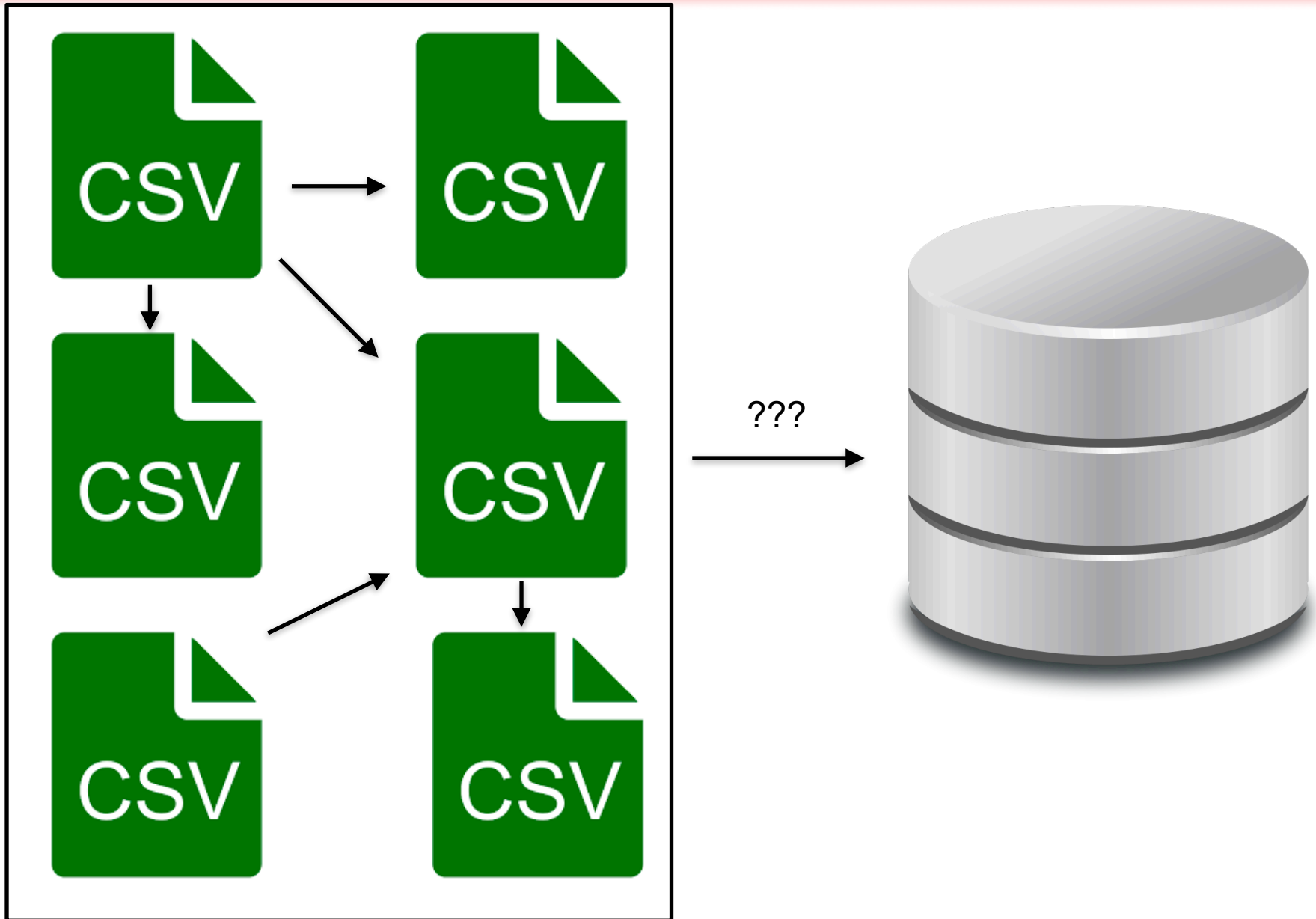












## Relational Database

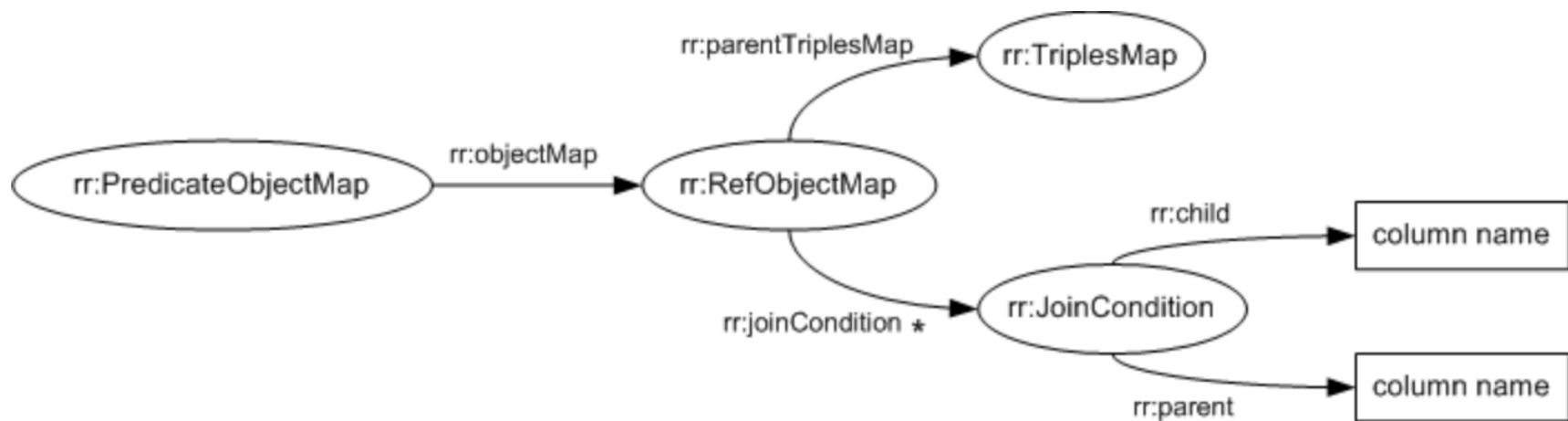
```
id,name,surname,birthdate,location  
1,david,chaves-fraga,27-11-1993,SDC
```

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

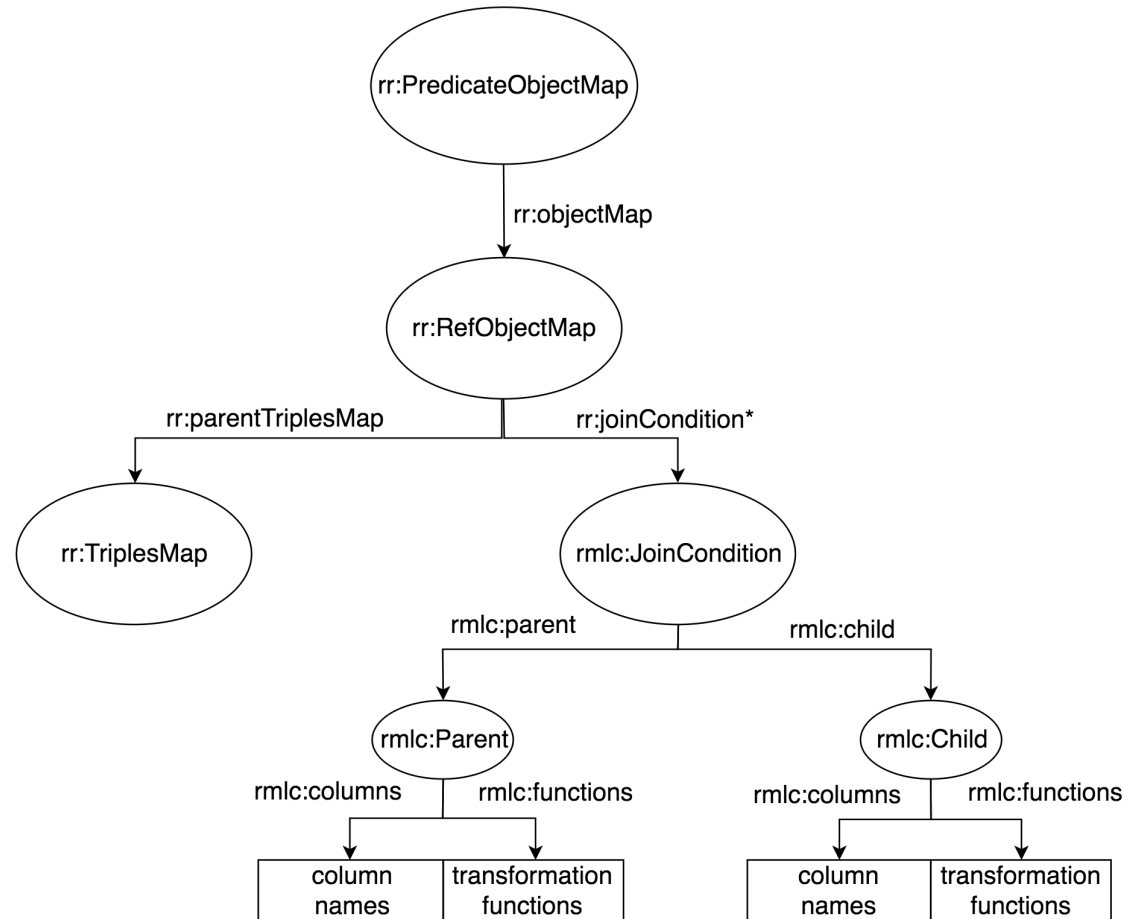
## CSV files

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

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



# RMLC: RDF Mapping Language for heterogeneous CSV files



Are the functions programmed? NO! They are SQL basic 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 .  
}
```



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

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

```
<#TriplesMap1>
....
rr:predicateObjectMap[
  rr:predicate foaf:name;
  rr:objectMap [
    rr:parentTriplesMap <#TriplesMap2>;
    rr:joinCondition [
      rmlc:child [
        rmlc:columns "['FULL_NAME']";
        rmlc:functions "LOWER(column[0])";
      ];
      rmlc:parent [
        rmlc:columns "['NAME','SURNAME']";
        rmlc:functions
          "CONCAT(column[0],' ',REPLACE(column[1],'_',' '))";
      ];
    ];
  ];
];
.
```

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

```
<#TriplesMap1>
....
rr:predicateObjectMap[
  rr:predicate foaf:name;
  rr:objectMap [
    rr:parentTriplesMap <#TriplesMap2>;
    rr:joinCondition [
      rmlc:child [
        rmlc:columns "[FULL_NAME]";
        rmlc:functions "LOWER(column[0])";
      ];
      rmlc:parent [
        rmlc:columns "[NAME,'SURNAME]";
        rmlc:functions
          "CONCAT(column[0],' ',REPLACE(column[1],'_',' '))";
      ];
    ];
  ];
];
.
```



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

- Other features
  - Transforming CSV columns to RDF objects
  - Maintain the semantics of R2RML
  - Create an enriched database schema using the mapping info (optimizations)

```

<TriplesMapBands>
...
rr:predicateObjectMap[
  rr:predicate ex:name;
  rr:objectMap [
    rr:datatype xsd:string;
    rmlc:columns ["BAND"];
    rmlc:functions "REPLACE(SUBSTRING(columns[0],1,5),' ','-')";
  ] ];
rr:predicateObjectMap[
  rr:predicate ex:leader;
  rr:objectMap [
    rmlc:columns ["NAME","SURNAME"];
    rmlc:template
    "http://ex.org/leader/{CONCAT(columns[0],'-',columns[1])}";
  ] ];

```

## RMLC

```

rr:predicateObjectMap[
  rr:predicate gtfs:service;
  rr:objectMap [
    rr:parentTriplesMap <TriplesMap1>;
    rr:joinCondition [
      rmlc:child [
        rmlc:columns "SERVICE_ID";
      ];
      rmlc:parent [
        rmlc:columns "SERVICE_ID";
      ];
    ]
  ]
];

```

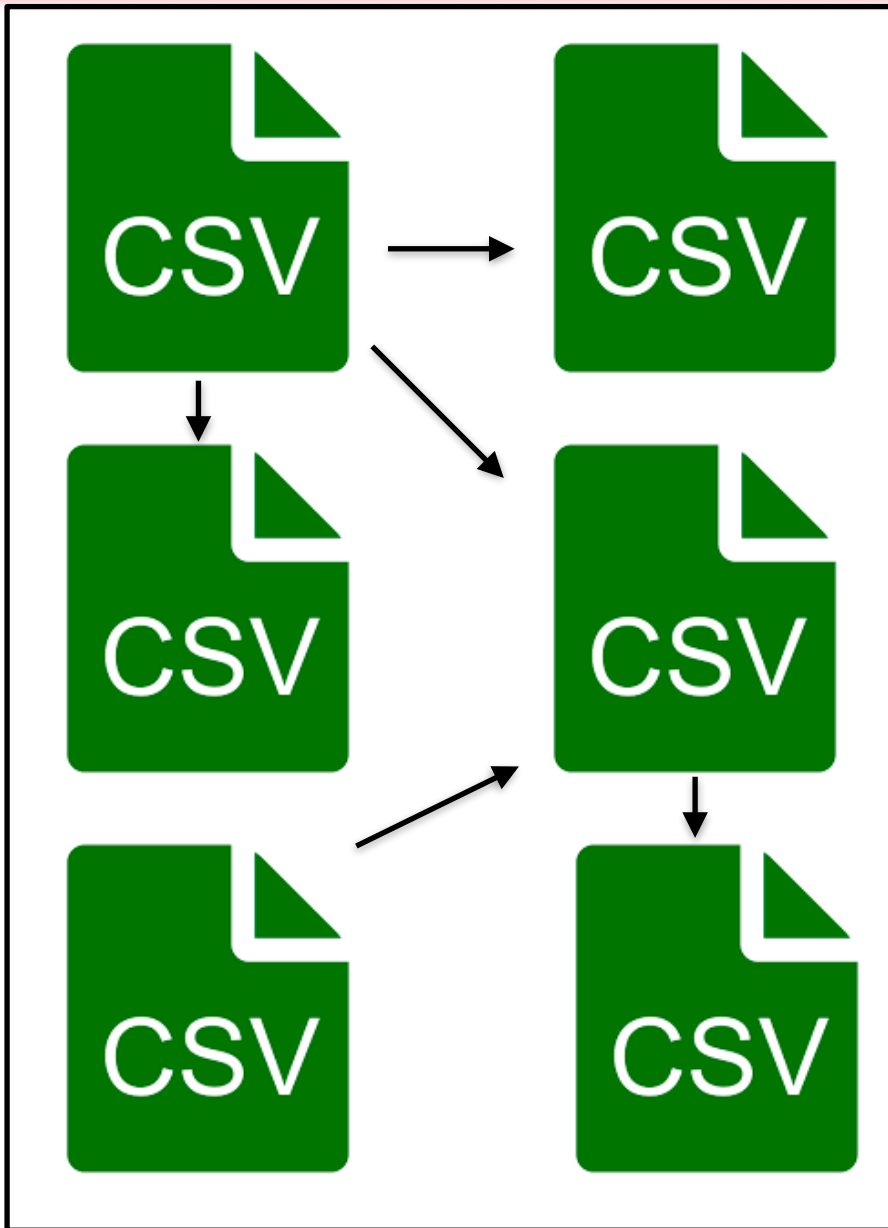
## R2RML

```

rr:predicateObjectMap[
  rr:predicate gtfs:service;
  rr:objectMap [
    rr:parentTriplesMap <TriplesMap1>;
    rr:joinCondition [
      rr:child "SERVICE_ID";
      rr:parent "SERVICE_ID";
    ];
  ];
];

```





+

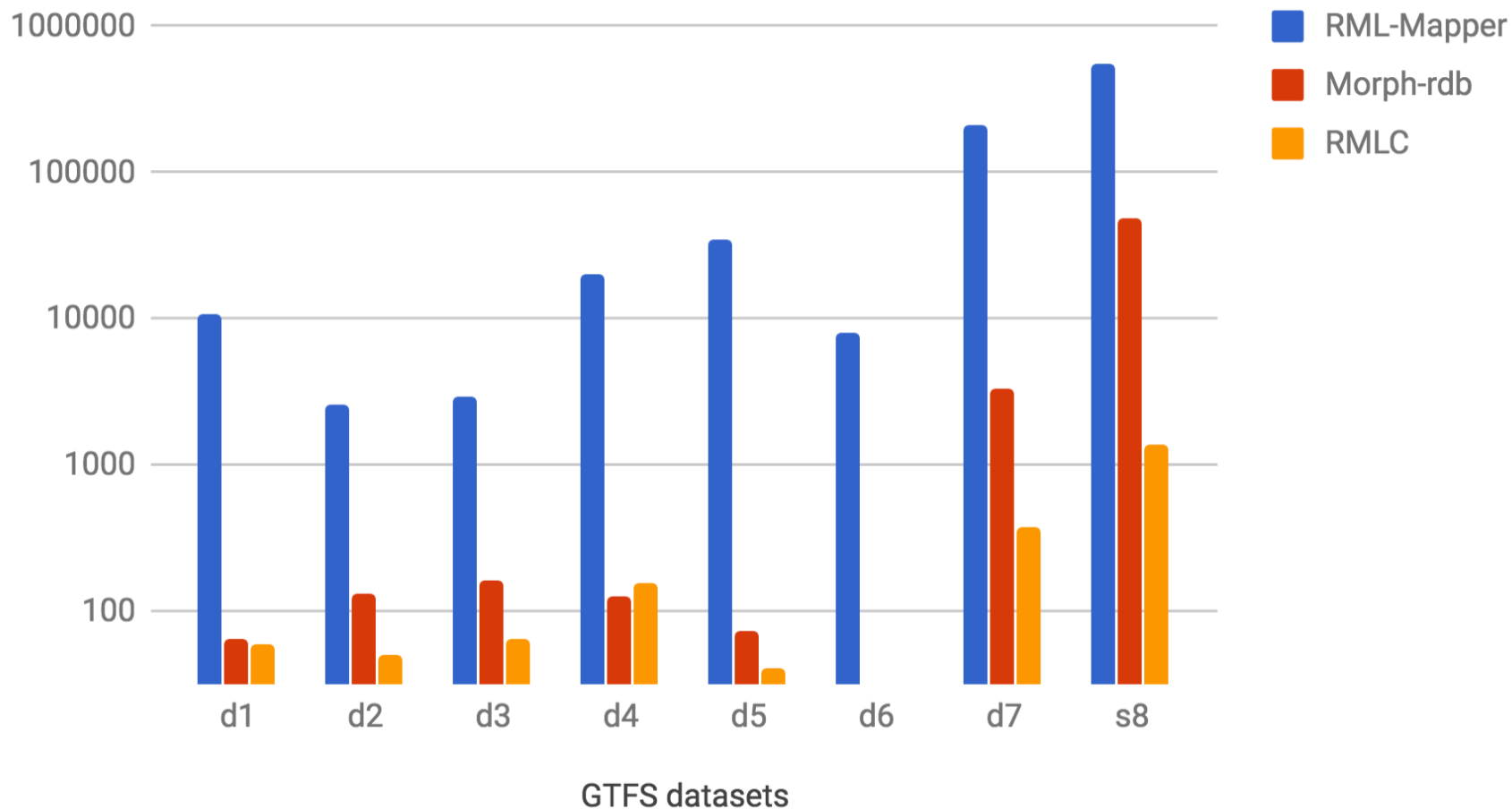
RMLC

=



- Datatypes
- Primary Keys
- ¿Foreign Keys?

## GTFS to RDF materialization



# RMLC: RDF Mapping Language for CSV files

## Main Contributions:

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

## Future Work:

- RMLC processor (SPARQL-to-SQL with functions)
- Comparison between RMLC and R2RML (in terms of performance)
- Optimizations in the resulted SQL queries