LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

TIB

Title

Virtual Knowledge Graph Generation from Heterogeneous Data Sources

Supervisor/Mentor

Oscar Corcho (OEG-UPM, Madrid)

Research Area

Semantic Data Integration

Project

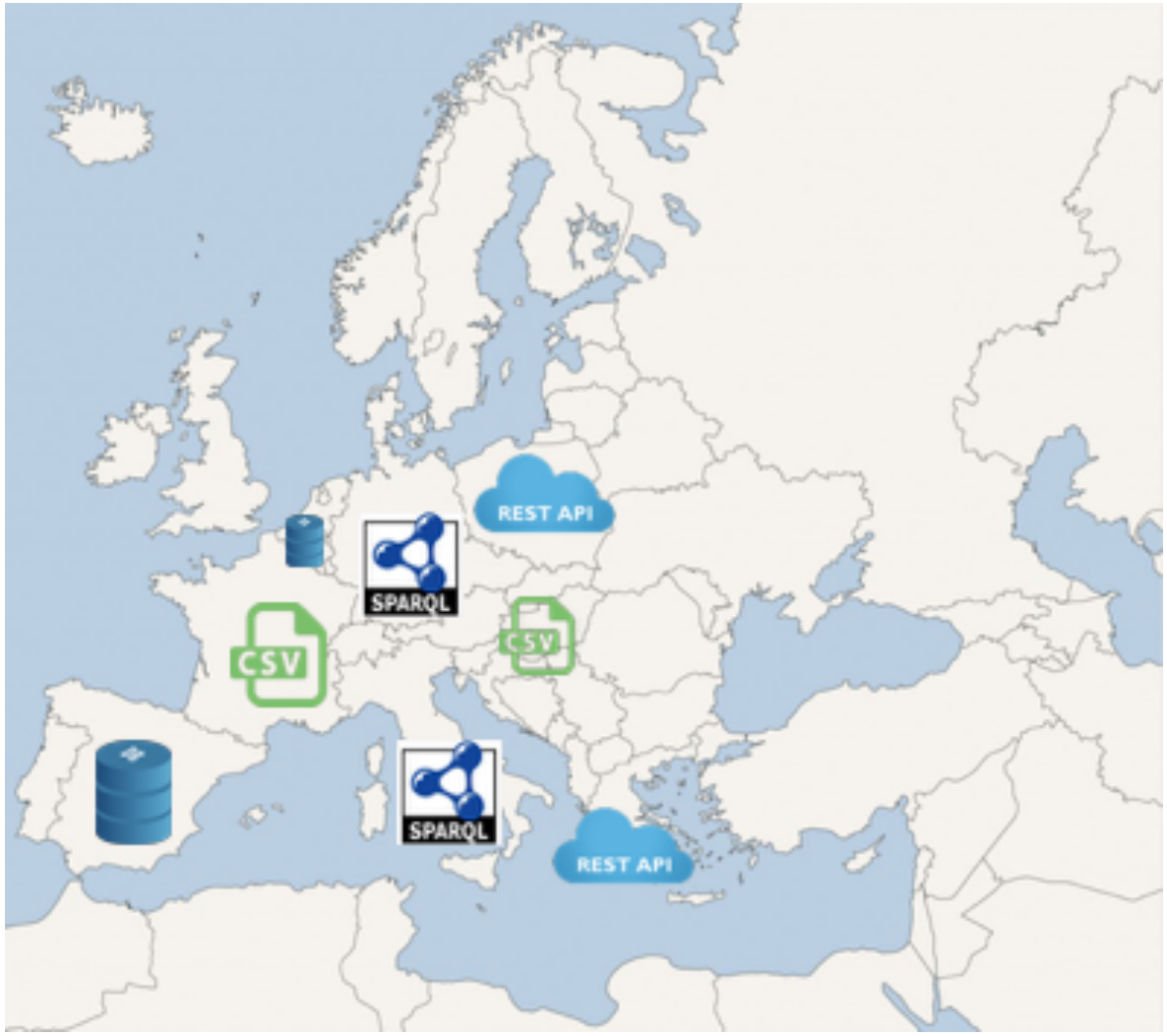
GoF4R, OASIS, SPRINT, SNAP, Datos 4.0

Leibniz Association

Motivating Example

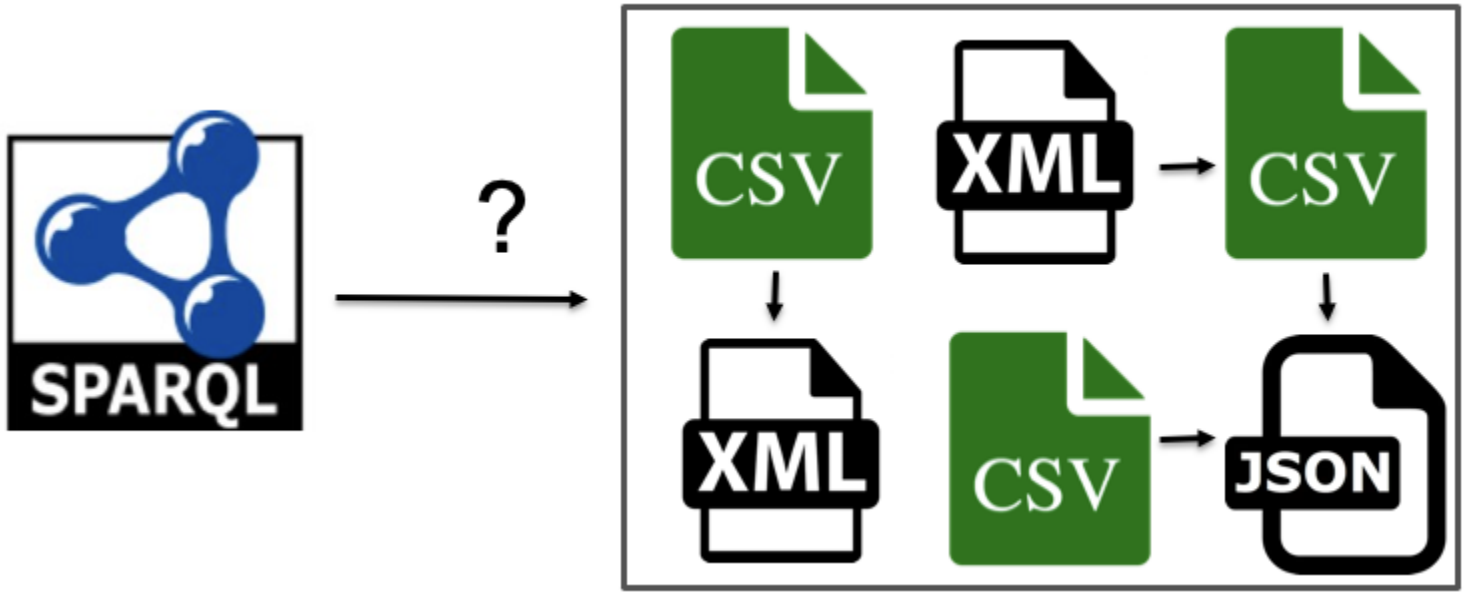
Transport National Access Point and How to Linked with Open City Data:

- JSON
- XML
- CSV
- RDB
- RDF
- REST API



Research Challenges

How to query the data on the Web?



Research Questions

- How can OBDA query-translation over heterogeneous data formats be enabled using or extending existing mapping languages and proposed optimization techniques/engines?

- **How can federated query processing over virtual RDF graphs from heterogeneous data formats be planned adapting the techniques applied over native RDF?**

- How can mappings be translated among different specifications without losing its own properties to be adapted to several scenarios?

Research Goals and Problem Formulation

Goals:

- Define and formalize the Mapping Translation concept.
- Enable OBDA query-translation over CSV using existing proposals of the state-of-the-art and OBDA R2RML-compliant engines
- **Federated query engine over OBDA engines using mapping information.**
- Benchmarking for OBDA query-translation engines that provide access to heterogeneous data formats.

Problem:

How to provide unified access to the data on the web without materializing them.

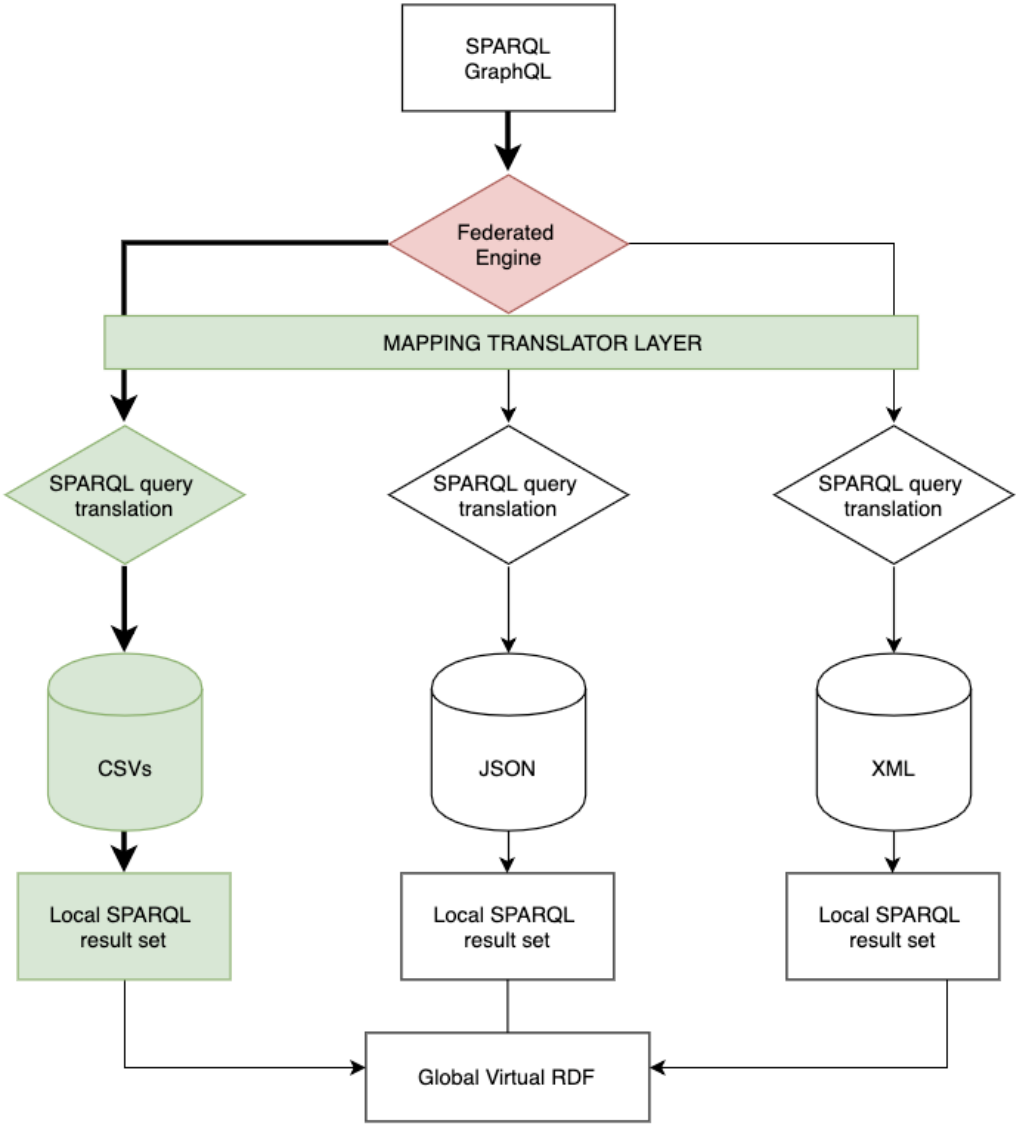
Potential Applications

- Public transport route planners with live information from the city.
- SPARQL/GraphQL-LD access to live open data
- Avoid the necessity of specific engines for each mapping language

What Existing Approaches Have Done?

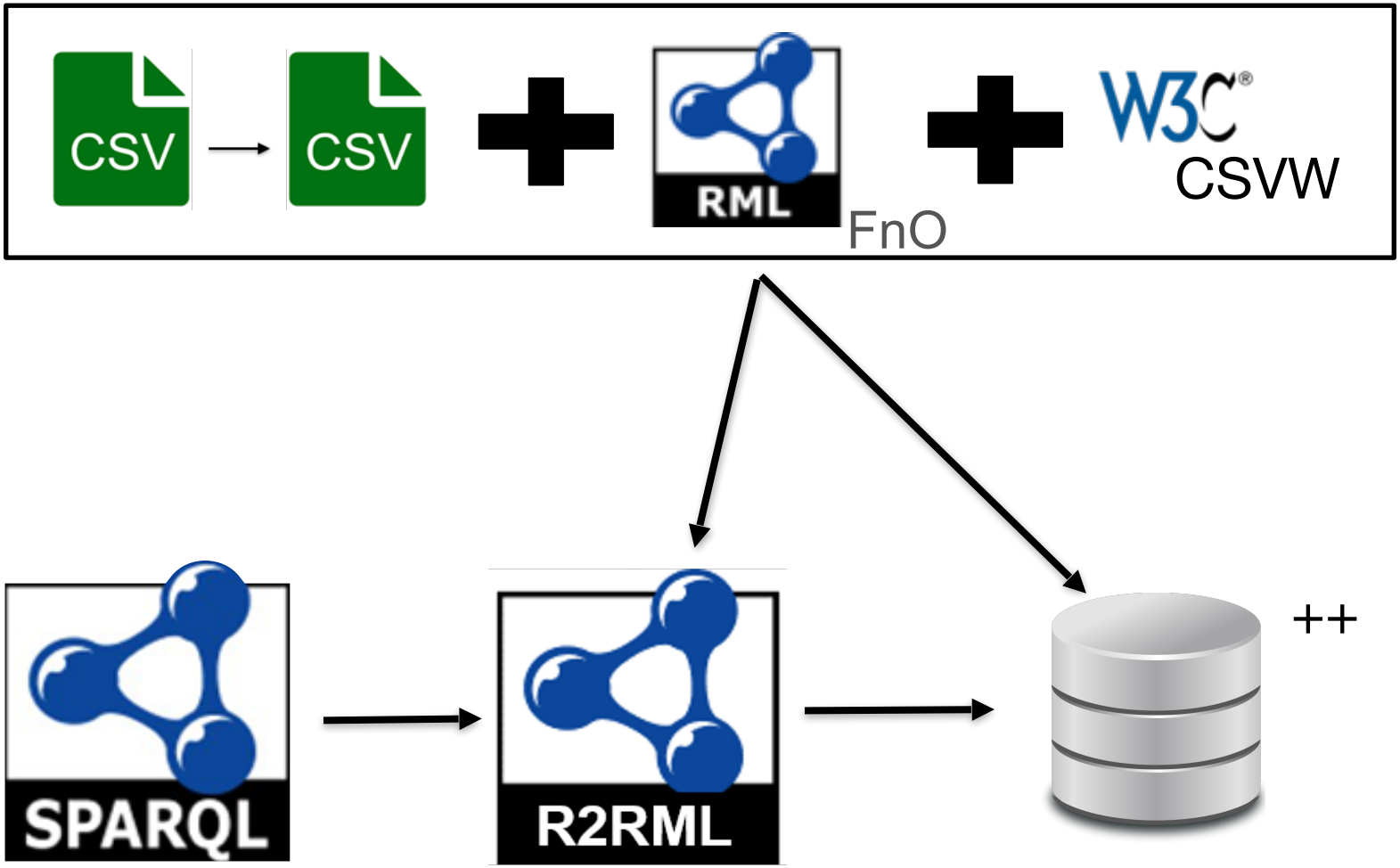
- **RDB2RDF (R2RML)**
  - Advantages
    - Good Optimizations on SPARQL-to-SQL
    - W3C Recommendation
    - Federation delegated to SQL
  - Disadvantages
    - Focused only on RDB
    - CSV != RDB
- **RML**
  - Advantages
    - CSV, RDB, JSON, XML support
    - Emergence of its use (Ontario, CARML, PolyWeb)
  - Disadvantages
    - Focused on RDF materialization
- **CSVW**
  - Advantages
    - W3C Recommendation
    - Specific properties of the field
  - Disadvantages
    - Focused on RDF materialization

Proposed Solution and Contributions



Initial Results

- **morph-CSV:** OBDA query translation for CSV files (to submit)



Experimental Results

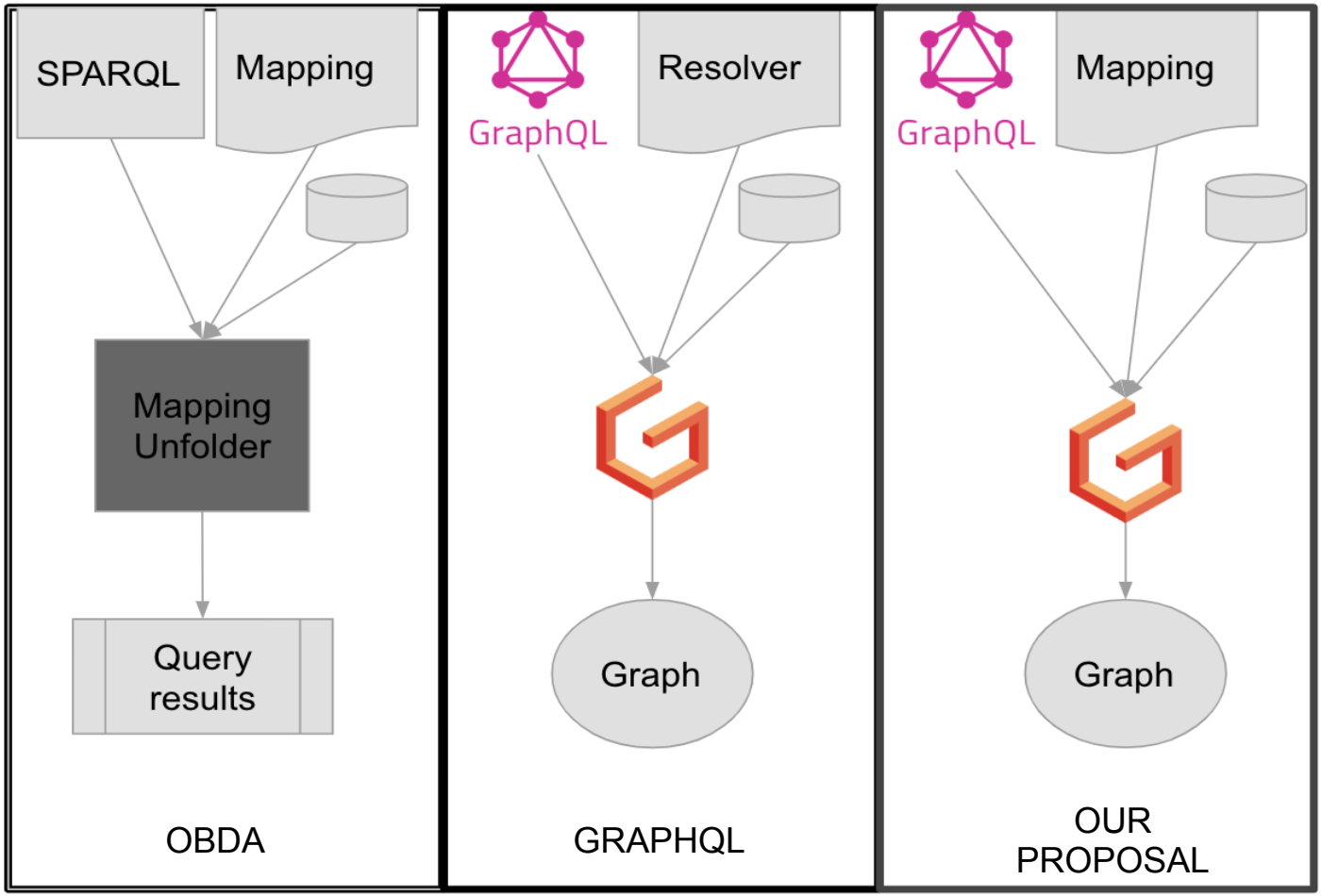
OBDA for CSV (Transport Data):

R: Running Time
I: Index Time
T: Total Time

D	RML-Mapper			Morph-RDB			RMLC-CSV		
	R	I	T	R	I	T	R	I	T
d1	182.00	0	182.00	1.07	0	1.07	0.98	0.02	**1.00**
d2	42.64	0	42.64	2.27	0	2.27	0.85	0.01	**0.86**
d3	50.01	0	50.01	2.89	0	2.89	1.09	0.01	**1.10**
d4	343.79	0	343.79	2.02	0	**2.02**	2.52	0.13	2.64
d5	594.47	0	594.47	1.34	0	1.34	0.67	0.01	**0.68**
d6	132.02	0	132.02	0.42	0	0.42	0.51	0.01	**0.53**
d7	>10h	0	>10h	54.97	0	54.97	6.00	0.42	**6.42**
d8	>10h	0	>10h	>10h	0	>10h	23.32	0.01	**23.34**

Initial Results

- **morph-GraphQL:** GraphQL resolver generation from R2RML (submitted)



Initial Results

morph-CSV: RMLC-Iterator for statistics CSV files (published)

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

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;
      ];
];
```

```
<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”; ];
];
```

Experimental Results

RMLC-Iterator:

Features	Statistics from Sri Lanka Touism		EautoStat-Inmigration	
	R2RML	**RMLC**	**R2RML**	**RMLC**
Total lines	~700	74	~2800	<70
TriplesMap	12	1	>40	1
PredicateObject Maps	60	5	>170	4

Lessons Learned

- **OBDA isn’t only SPARQL-to-SQL**
- **CSV are one of the most formats for exposing data on the web**
  - This isn’t going to change soon (the administration won’t publish RDF)
  - We wouldn’t be data providers (RDF materialization)
- **Decentralization is needed**
  - Google Maps spend days updating its data
- **Mapping languages contain useful information**
- **Mapping languages don’t have to resolve ad-hoc problems**
- **Mapping languages should be aligned among others**
  - Try to reuse optimized frameworks

Limitations and Future Work

Limitations:

- Generation of mappings/rules
- Optimization techniques over the query translation process
- CSV supported by RDBMS
- Lack of Benchmarking for an objective evaluation
- Focused only in CSV format

Future work:

- Benchmarking
- Mapping translator concept
- Federation

Timeline for Future Work

Organization:

- VKG2019 - Tutorial at ESWC2019 (OEG team)
- Workshop on KG Building at ESWC2019 (OEG, RML and J. Sequeda)
- Open Summer of Code 2019 (Madrid - July 2019)

OEG 2019:

- morph-CSV: Enabling OBDA query-translation over CSV files (written)
- Mapping Translation Concept (written)
- OBDA Benchmarking for Heterogeneous Data Sources (on-going)

TIB (March-June):

- Federated query processing over OBDA engines and Heterogeneous Data Sources
- Intensive collaboration with Scientific Data Management Group

Publications

Published:

- David Chaves-Fraga et al. **Virtual Statistics Knowledge Graph Generation from CSV files**. In:Emerging Topics in Semantic Technologies: ISWC2018 Satellite Events. Best paper award at SemStats2018
- Rojas Melendez, J. A., Chaves, D., Colpaert, P., Verborgh, R., & Mannens, E. (2017). **Providing reliable access to real-time and historic public transport data using linked connections**. In ISWC2017 (Demo)
- 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 DeSemWeb2017
- Chaves-Fraga, D., Gutierrez, C., & Corcho, O. (2017). **On the Role of the GRAPH Clause in the Performance of Federated SPARQL Queries.** In PROFILES@ ISWC2017.

Under Review:

- GraphQL Schema and Resolver Generation fromR2RML Mappings (SEKE2019)
- Conformance Test Cases for the RDF Mapping Language (RML) (KGSWC2019)