UNIVERSIDADE FEDERAL DO CEARÁ

Towards Automatic Generation of RDB2RDF Mappings

Vânia Maria P. Vidal, José Maria Monteiro, Luís Eufrasio T. Neto

Departamento de Computação

Contents

1)Introduction

2)Related Work

3)RDB2DO Approach

4)SQL Views By Assertions

5)R2RML Direct Mapping

Introduction

Introduction

1)Motivation

2)Use Case

3)R2RML Mapping Language

4)Challenges

5)Proposal

Motivation

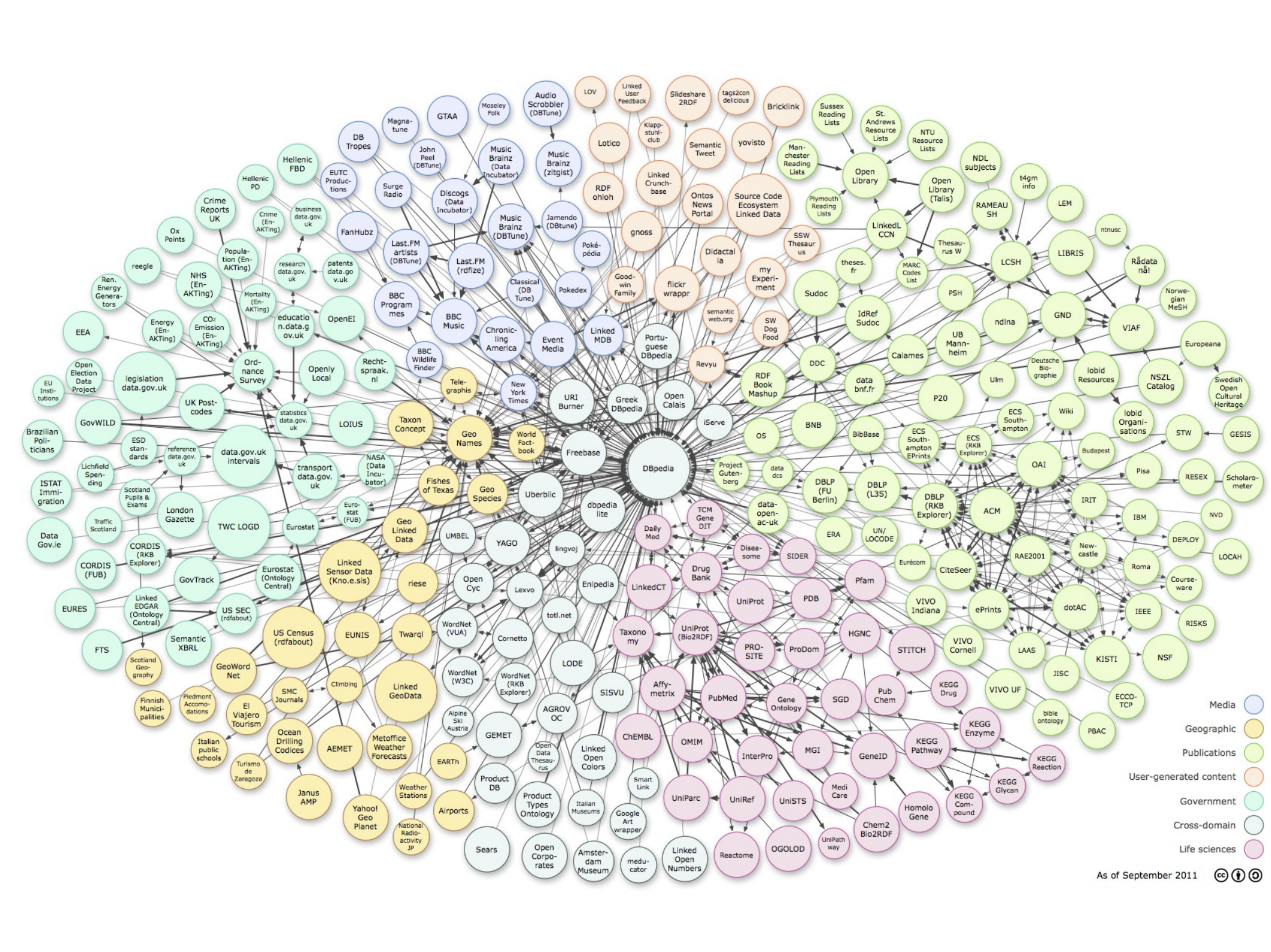
- Linked Data: a term used to describe a recommended best practice for exposing, sharing, and **connecting** pieces of data, information, and knowledge on the Semantic Web using URIs and RDF.
- Relational Databases (RDBs): used in most current enterprise environments to store and manage data.
- The relational data must be published as Linked Data. Why?

Motivation

- RDBs are well suited to handle large amounts of data, but they were not designed to preserve the data semantics.
- The meaning of the data is implicit on the application level and not explicitly encoded in the relational model.
- Ontologies and Semantic Web technologies provide explicit semantics in a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.

Motivation

The Linking Open Data cloud diagram
Last updated: 2011-09-19


Media
Geographic
Publications
User-generated content
Government
Cross-domain
Life sciences
As of September 2011

Motivation

- How to Publish Relational Databases as Linked Data?
  1. Create a Mapping from the relational schema to a custom/existing ontology/vocabulary schema
     - Identifying “things” (or Classes)
       - E.g., “Paper”, “Person”, “Conference”
       - Each relational table is mapped to a class
     - Identifying predicates
       - E.g., “a:author”, “a:firstName”, “a:homepage”
     - Reuse terms of known vocabularies
       - E.g., “dc:creator”, “foaf:name”, “foaf:homepage”
  2. Creating RDF using the mapping
     - Creating instances (or objects), assigning unique URIs

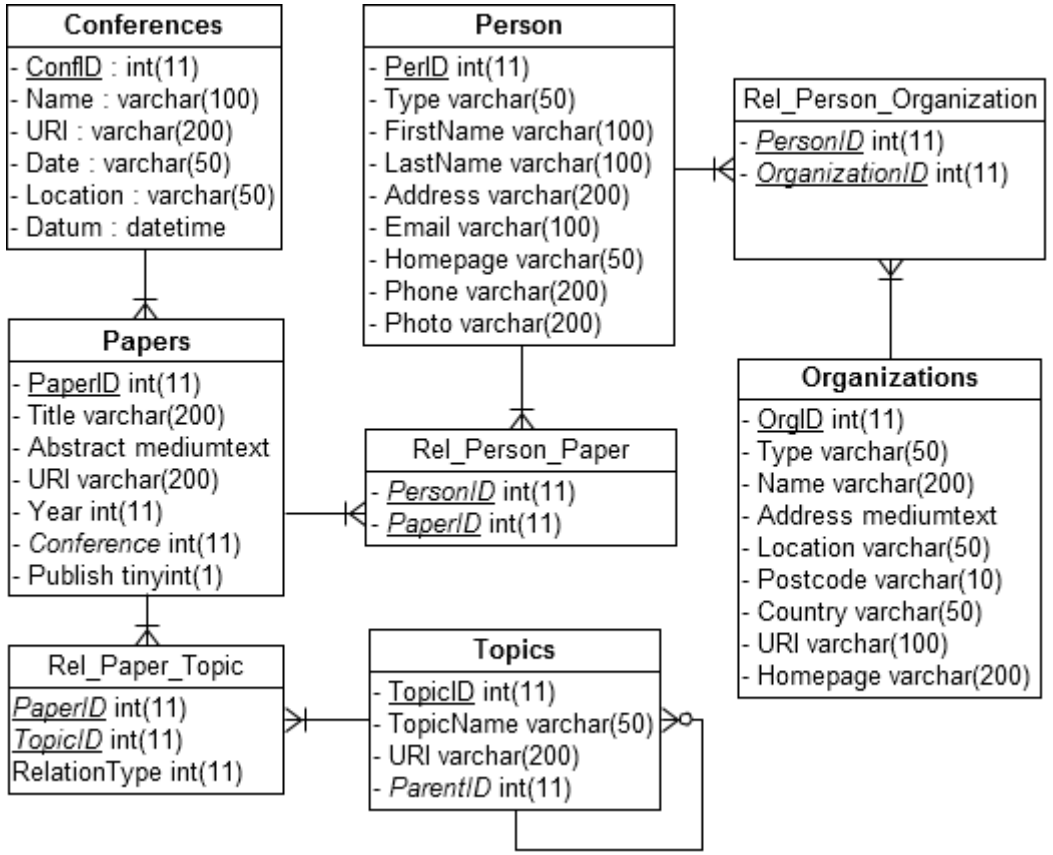
Motivation

- How to Publish Relational Databases as Linked Data?

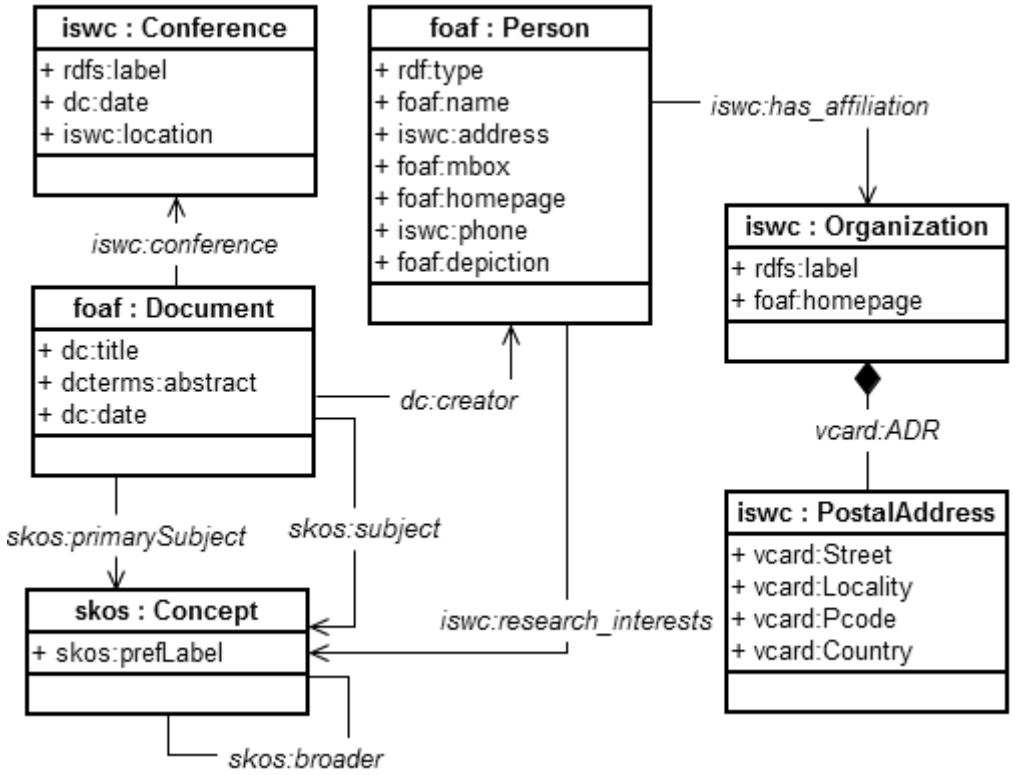
  2. Creating RDF using the mapping
     - Creating instances (or objects), assigning unique URIs
     - Can be performed once in an offline process
       - Triplification
       - Materialized into RDF Triple Stores
       - Must be updated
     - Or on-the-fly in an online fashion
       - Virtual
       - End Point SPARQL
       - Translations from SPARQL to SQL and from ResultSet to RDF GRAPH

Use Case

ISWC_DB



ISWC Ontology



Use Case

- Person Tuples

PerID	FirstName	LastName	Email	Homepage
1	Yolanda	Gil	gil@isi.edu	http://www.isi.edu/~gil
2	Varun	Ratnakar	varunr@isi.edu	http://www.isi.edu/~varunr
3	Jim	Blythe	blythe@isi.edu	http://www.isi.edu/~varunr

- Triples of PerID = 1:

<http://example.org/person/1> rdf:type foaf:Person .

<http://example.org/person/1> foaf:name “Yolanda Gil” .

<http://example.org/person/1> foaf:mbox <mailto:gil@isi.edu> .

<http://example.org/person/1> foaf:homepage <http://www.isi.edu/~gil> .

Use Case

- Paper Tuples

PaperID	Title	Year
1	Trusting Information Sources One Citizen at a Time	2002
2	Automatic Generation of Java/SQL based Inference Engines from RDF Schema and RuleML	2002

- Rel_Person_Paper Tuples

PersonID	PaperID
1	1
2	1

Use Case

- Triples of PaperID = 1:

<http://example.org/paper/1> rdf:type foaf:Document .

<http://example.org/paper/1> dc:title

“Trusting Information Sources One Citizen at a Time” .

<http://example.org/paper/1> dc:date “2002”^^xsd:gYear .

<http://example.org/paper/1> dc:creator <http://example.org/person/1> .

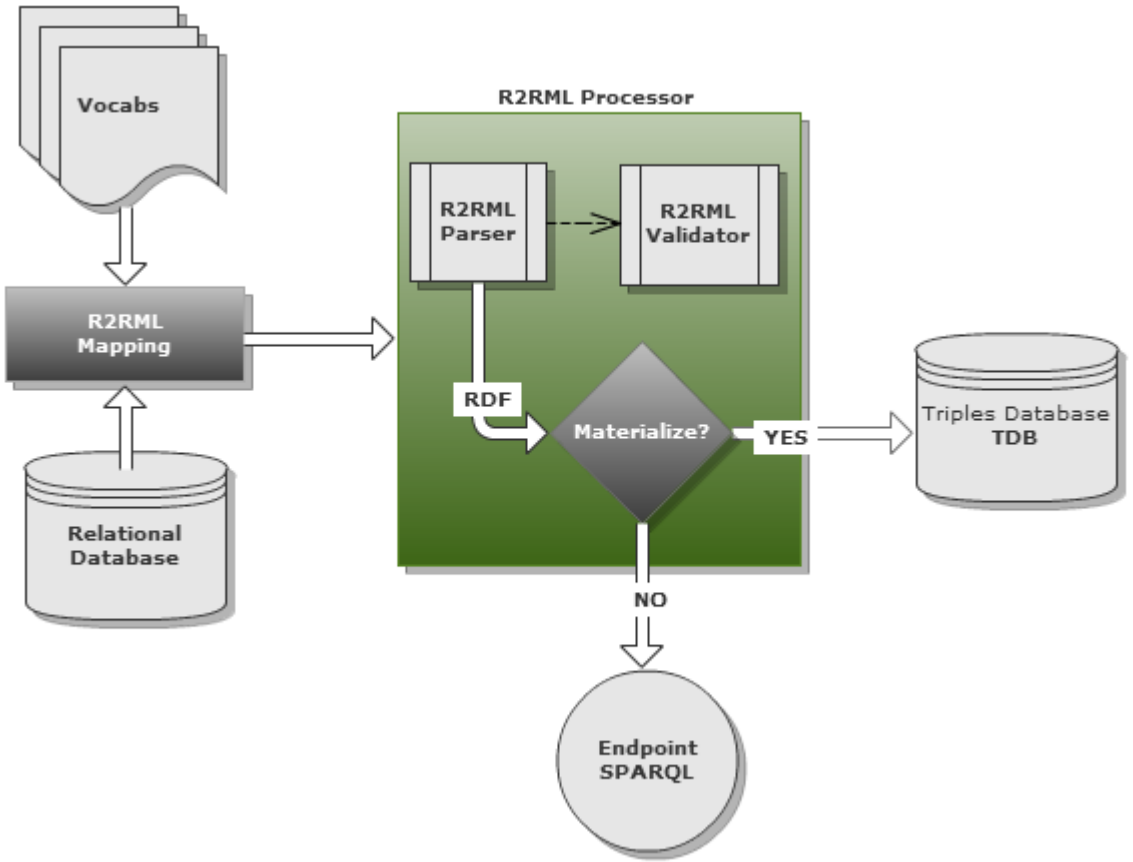
<http://example.org/paper/1> dc:creator <http://example.org/person/2> .

R2RML

- A language for expressing customized mappings from relational databases to RDF datasets.
- Every R2RML mapping is tailored to a specific database schema and target vocabulary.
- The input to an R2RML mapping is a relational database that conforms to that schema.
- The output is an RDF dataset that uses predicates and types from the target vocabulary.

R2RML

R2RML processors are free to materialize the output data, or to offer virtual access through an interface that queries the underlying database:



R2RML

Overview and Example

R2RML

An R2RML mapping refers to logical tables to retrieve data from the input database. A logical table can be one of the following:

1. A base table,
2. a view, or
3. a valid SQL query (called an “R2RML view” because it emulates a SQL view without modifying the database).

R2RML

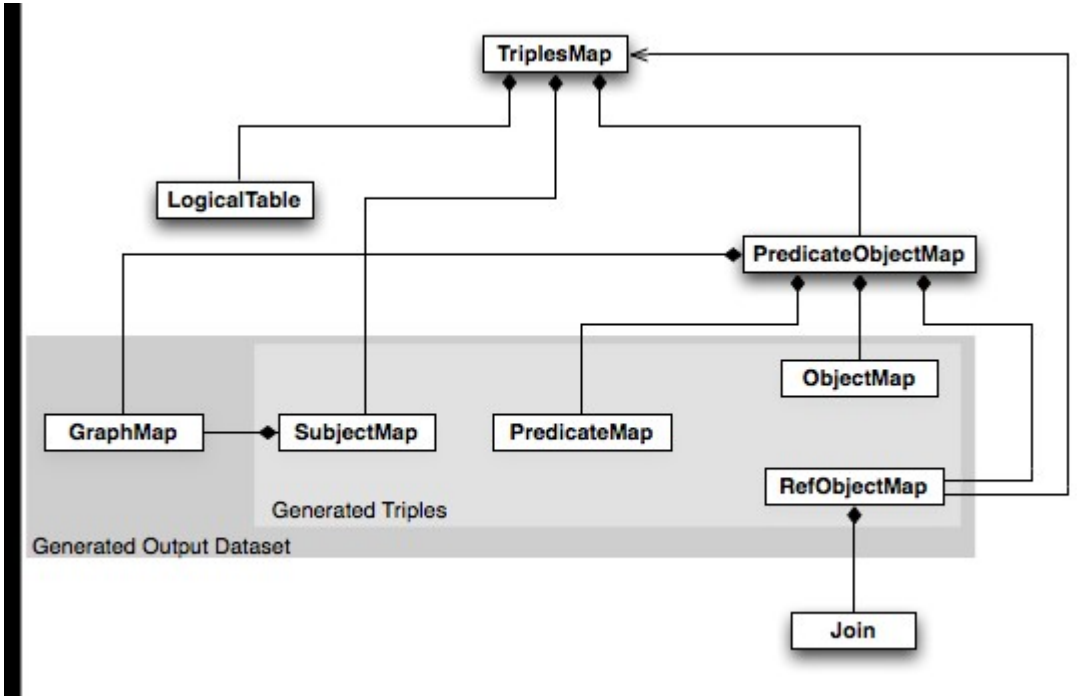
- Each logical table is mapped to RDF using a triples map.
- The *triples map* is a rule that maps each row in the logical table to a number of RDF triples.
- The rule has two main parts:
  1. A *subject map* that generates the subject of all RDF triples that will be generated from a logical table row. The subjects often are IRIs that are generated from the primary key column(s) of the table.
  2. Multiple *predicate-object maps* that in turn consist of *predicate maps* and *object maps* (or referencing object maps).

R2RML

- Triples are produced by combining the subject map with a predicate map and object map, and applying these three to each logical table row.
- For example:
  - Subjects: http://example.org/person/{PerID} is the template used to generate subject IRIs from the *PerID* column
  - Predicates: The constant vocabulary IRI *foaf:mbox* is used
  - Objects: The value of the *Email* column is used to produce an RDF literal

R2RML

- Overview



R2RML

- Simple Table: Conferences Tuples

ConfID	Name	Date	Location
23541	International Semantic Web Conference 2002	June 9-12, 2002	Sardinia
23542	15th International World Wide Web Conference	May 23-26, 2006	Edinburgh

- Triples of ConfID = 1:

<http://example.org/conf/23541> rdf:type iswc:Conference .

<http://example.org/conf/23541> rdfs:label

“International Semantic Web Conference 2002” .

<http://example.org/conf/23541> dc:date “June 9-12, 2002” .

<http://example.org/conf/23541> iswc:location “Sardinia” .

R2RML

- Simple Table Mapping:

```
<#ConfTriplesMap>
  rr:logicalTable [ rr:tableName "Conferences" ];
  rr:subjectMap [
    rr:template "http://example.org/conf/{ConfID}";
    rr:class iswc:Conference;
  ];
  ...
```

R2RML

- Simple Table Mapping:

```
<#ConfTriplesMap>
  ...
  rr:predicateObjectMap [
    rr:predicate rdfs:label;
    rr:objectMap [ rr:column "Name" ];
  ];
  rr:predicateObjectMap [
    rr:predicate dc:date;
    rr:objectMap [ rr:column "Date" ];
  ];
  ...
```

R2RML

- Simple Table Mapping:

```
<#ConfTriplesMap>
  ...
  rr:predicateObjectMap [
    rr:predicate iswc:location;
    rr:objectMap [ rr:column "Location" ];
  ].
```

R2RML

- Linking Two Tables: Papers and Conference

	PaperID	Conference
▶	1	23541
	2	23541

select PaperID, Conference *from* Papers

- Triples of PaperID IN (1, 2):

<http://example.org/paper/1> iswc:conference <http://example.org/conf/23541> .

<http://example.org/paper/2> iswc:conference <http://example.org/conf/23541> .

R2RML

- Linking Two Tables Mapping:

```
<#PapersTriplesMap>
  rr:logicalTable [ rr:tableName "Papers" ];
  rr:subjectMap [
    rr:template "http://example.org/paper/{PaperID}";
    rr:class foaf:Document;
  ];
  ...
```

R2RML

- Linking Two Tables Mapping:

```
<#PapersTriplesMap>
  ...
  rr:predicateObjectMap [
    rr:predicate iswc:conference;
    rr:objectMap [
      rr:parentTriplesMap <#ConfTriplesMap>;
        rr:joinCondition [
          rr:child "Conference";
          rr:parent "ConfID";
        ];];].
```

Challenges

- R2RML is the RDB to RDF Mapping Language W3C Proposed Recommendation.
- But their use involves some challenges:
  - The user must have advanced knowledge about the R2RML language.
  - There isn't any GUI Tool for creating R2RML mappings.
  - Problems can arise from heterogeneity between relational schema and the ontology schema.
  - When changes occur in any of the schemas, the mapping must be updated manually.

Proposal

- To overcome these challenges we propose:
  - A graphical tool where the user can create their mappings through correspondence assertions
  - A three-tier architecture that includes the creation of SQL views from the correspondence assertions
  - A publishing process that enables the use of the tool, the architecture implementation and automation of R2RML mapping creation.

Related Work

Related Work

- Virtuoso
  - http://virtuoso.openlinksw.com
- Spyder
  - http://www.revelytix.com/content/spyder
- Triplify
  - http://triplify.org/
- D2RQ
  - http://d2rq.org/

Related Work

- Virtuoso
  - A commercial database system from Open-Link Software
  - Supports a multi-model data server
    - XML, SQL, RDF, plain text
  - Has a declarative meta-schema language used to map terms of an ontology to concepts in the database schema
  - Open source version with limited support
  - Does not have a GUI Component to support mappings creation

Related Work

- Spyder
  - Free tool, but not open source.
  - Implements R2RML
  - Supports only a subset of SPARQL 1.1
  - Does not implement the R2RML current version
  - Does not have a GUI Component

Related Work

- Triplify
  - It is a plugin for Web applications that exposes their information in RDF.
  - Uses a set of application-specific SQL queries to extract data from the underlying RDB to generate RDF data from the results.
  - The SQL queries have to be defined manually for each Web application.
  - The RDF generation is performed automatically according to a fixed process.
  - Reuse of existing ontologies is possible via result column renaming in the SQL queries.

Related Work

- D2RQ
  - It is a robust platform and open source.
  - Widely used by the community.
  - Creates End Points SPARQL
  - Supports multiple databases.
  - Uses D2R Map Language to create mappings
  - Will be updated to support R2RML
  - Doesn't have a GUI Tool to create the mappings

RDB2DO Approach

RDB2DO Approach

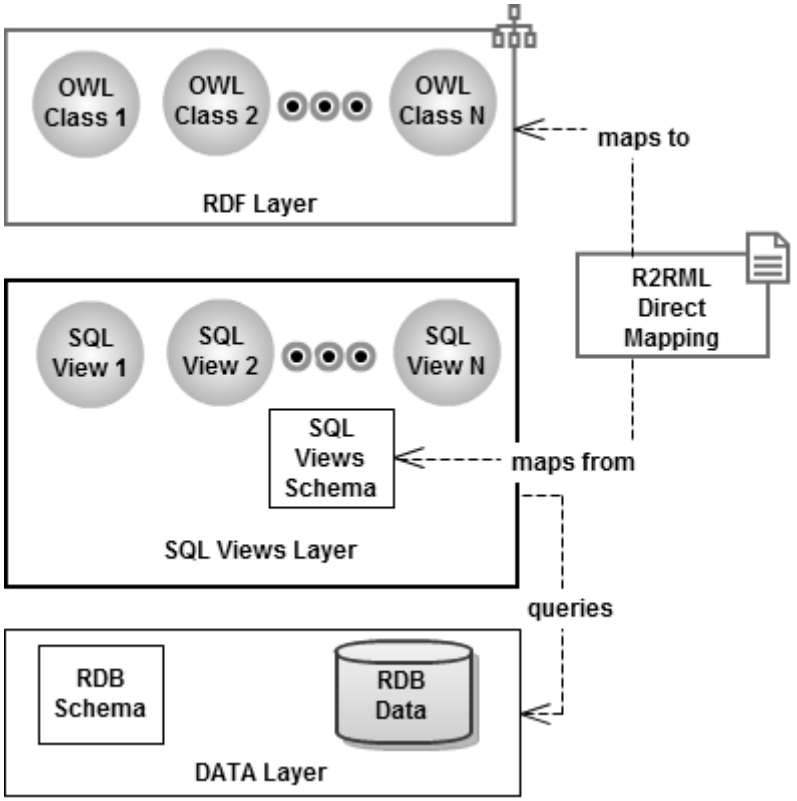
- RDB to Domain Ontology (RDB2DO) is our approach to map a RDB schema in terms of the Domain Ontology (DO) schema.
- Comprises:
  - Three-tier Architecture
  - RDB2DO Publication Process

RDB2DO Approach

- Three-tier Architecture
  - RDF Layer: *n* OWL classes. A subset of DO classes.
  - SQL Views Layer: Views schema and *n* SQL views.
  - Data Layer: RDB scheme and data.

RDB2DO Approach

- Three-tier Architecture

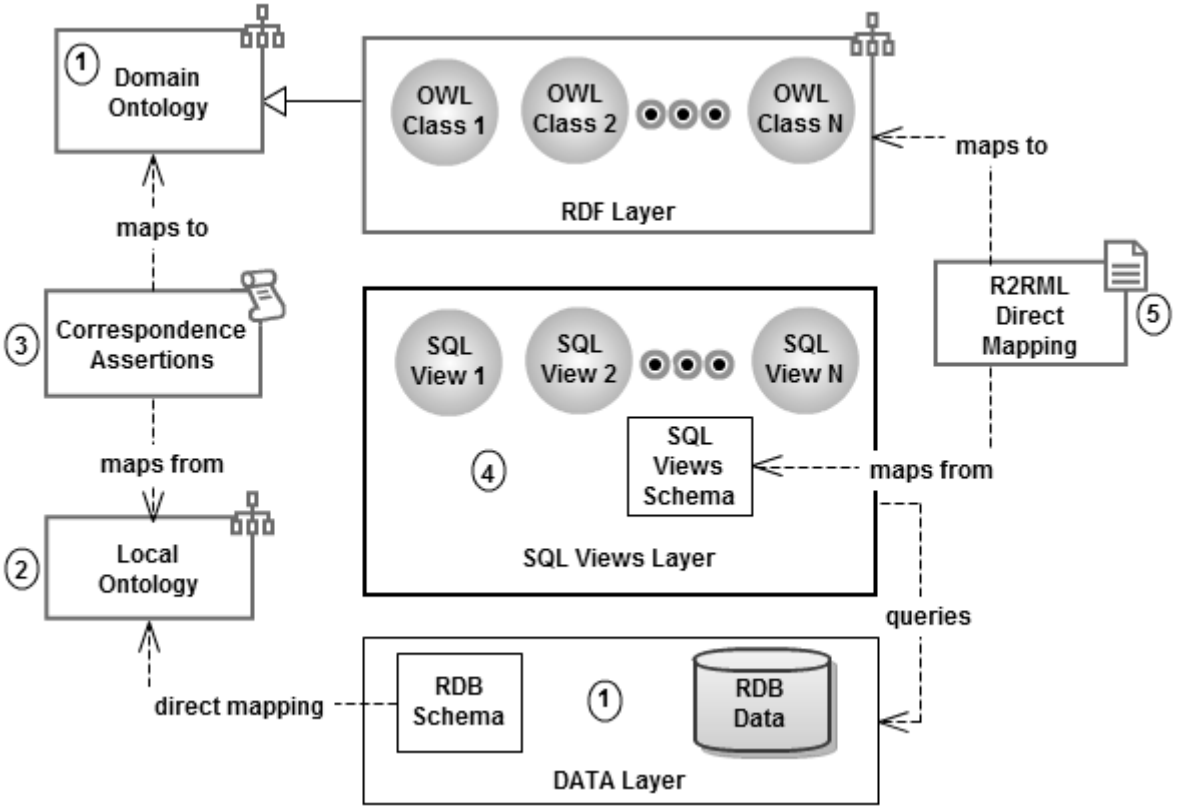


RDB2DO Approach

- RDB2DO Publication Process
  1. RDB and DO Selection
  2. Local Ontology Generation
  3. Mapping by Correspondence Assertions
  4. SQL Views and Schema Generation
  5. R2RML Direct Mapping Generation

RDB2DO Approach

- RDB2DO Publication Process



SQL Views By Assertions

SQL Views By Assertions

- Mappings formaly defined by Correspondence Assertions (CAs).
- One View for each OWL Class into the Application Ontology.
- SQL clauses of Views derived by the CAs.

SQL Views By Assertions

- Some CAs of Person Class

```
1.foaf:Person(person) ← db:Person(person)

2.foaf:mbox(person, email) ←
    db:email(person, email), db:Person(person)

3.foaf:name(person, fConcat(firstN, lastN) ←
  db:firstName(person, firstN) ^
  db:lastName(person, firstN),
  db:Person(person)

4.iswc:has_affiliation(person, ) ←
  db:Person(person) ^
  db:fk_org_person_inversa(person, rel_p_o) ^
  db:fk_person_org(rel_p_o, org),
  db:Organization(org)
```

SQL Views By Assertions

- Foaf_Person View

```
CREATE VIEW Foaf_Person AS
SELECT email as foaf_mbox, ---------------------(2)
CONCAT(FirstName, LastName) as foaf_name, ----(3)
(select rpo.OrganizationID
 from Rel_Person_Organization as rpo
 where rpo.PersonID=p.PerID)
 as iswc_affiliation ----------------------------(4)
FROM Person p ---------------------------------(1)
```

References

1. Resource Description Framework (RDF) Model and Syntax Specification, W3C Proposed Recommendation 05 January 1999. Disponível em http://www.w3.org/TR/PR-rdf-syntax/. Acessado em 2012.

2. Tim Berners-Lee: Linked Data, Design Issues. Disponível em:

<http://www.w3.org/DesignIssues/LinkedData.html >. Acessado em 2012.

3. A Direct Mapping of Relational Data to RDF, Marcelo Arenas, Eric Prud'hommeaux, Juan Sequeda, Editors. World Wide Web Consortium, 24 de Março de 2011. Disponível em http://www.w3.org/TR/rdb-direct-mapping/.

References

4. Percy E. Salas, Karin K. Breitman, Marco A. Casanova, José Viterb: StdTrip: Promoting the Reuse of Standard Vocabularies in Open Government Data.

5. Sacramento, E., Vidal, V. M., Macêdo, J. A., Lóscio, B. F., Lopes, F. L. R., Casanova, M. A., and Lemos, F.: Towards automatic generation of application ontologies. In: Journal of Information and Data Management, Vol. V, No. N, Month 20YY, pp.1-16, 2010.

6. David Beckett, Tim Berners-Lee: Turtle - Terse RDF Triple Language. Disponível em http://www.w3.org/TeamSubmission/2008/SUBM-turtle-20080114/ .

7. M. Dürst and M. Suignard: Internationalized Resource Identifiers (IRIs), Internet-Draft, June 2003, expires December 2003. This document is http://www.w3.org/International/iri-edit/draft-duerst-iri-04.

References

8. Hull, R. and Yoshikawa, M. Ilog: declarative creation and manipulation of object identifiers. In Proceedings of the 16th International Conference on Very Large Databases (VLDB). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 455#468, 1990.

9. Leme, L. A. P. P., Casanova, M. A., Breitman, K. K., and Furtado, A. L. Instance-based OWL schema matching. In Proceedings of the 11th International Conference on Enterprise Information Systems. Lecture Notes in Business Information Processing, vol. 24. pp. 14-26, 2009.

10. Das Souripriya, Sundara Seema, Cyganiak Richard (2011), R2RML: RDB to RDF Mapping Language. Disponível em < http://www.w3.org/TR/r2rml/>

References

11. Bizer, C., Cyganiak, R.: D2R server – publishing relational databases on the Semantic Web. Disponível em:

http://www4.wiwiss.fu-berlin.de/bizer/pub/Bizer-Cyganiak-D2R-Server-ISWC2

12. Jena - A Semantic Web Framework for Java. Disponível em:

<http://jena.sourceforge.net>