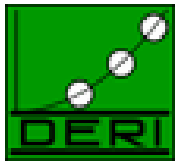
An overview of RDB2RDF techniques and tools

DERI Reading Group Presentation

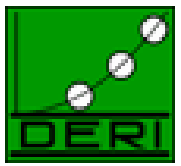
Nuno Lopes

August 26, 2009



Main purpose of RDB2RDF WG

"...standardize a language for mapping Relational Database schemas into RDF and OWL." [2]

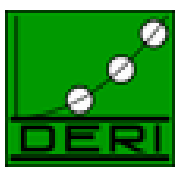


Main purpose of RDB2RDF WG

"...standardize a language for mapping Relational Database schemas into RDF and OWL." [2]

Enables to:

- Publish on the Web the vast amounts of information stored in Relational Databases (RDB)

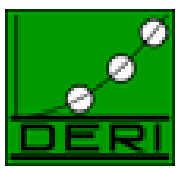


Main purpose of RDB2RDF WG

"...standardize a language for mapping Relational Database schemas into RDF and OWL." [2]

Enables to:

- Publish on the Web the vast amounts of information stored in Relational Databases (RDB)
- Integrate data from different RDBs

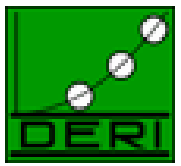


Main purpose of RDB2RDF WG

"...standardize a language for mapping Relational Database schemas into RDF and OWL." [2]

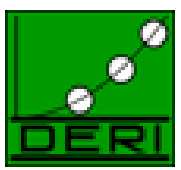
Enables to:

- Publish on the Web the vast amounts of information stored in Relational Databases (RDB)
- Integrate data from different RDBs
- Add semantics to the existing relational data



RDB2RDF mappings

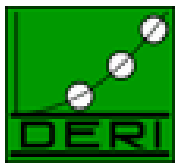
Two main ways of exposing the data:



RDB2RDF mappings

Two main ways of exposing the data:

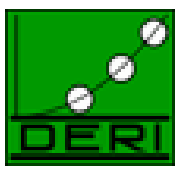
- Translate relational data to RDF (loadable into an RDF store)
- Generate a mapping of the RDB that can be queried using SPARQL



RDB2RDF mappings

Two main ways of exposing the data:

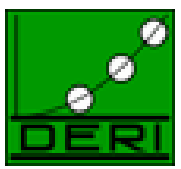
- Translate relational data to RDF (loadable into an RDF store)
  - Directly queriable RDF dump
- Generate a mapping of the RDB that can be queried using SPARQL
  - The SPARQL query is translated into SQL



RDB2RDF mappings

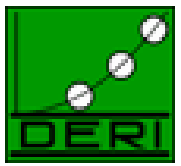
Two main ways of exposing the data:

- Translate relational data to RDF (loadable into an RDF store)
  - Directly queriable RDF dump
  - Harder to maintain consistency (e.g. constantly updated DB)
- Generate a mapping of the RDB that can be queried using SPARQL
  - The SPARQL query is translated into SQL
  - May lead to longer query times (e.g. inferencing)



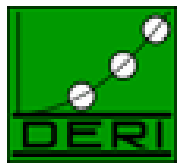
RDB2RDF mapping language goals

- Complete language (when compared to relational algebra). Subset for the first release.



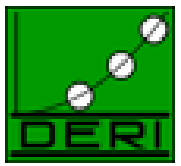
RDB2RDF mapping language goals

- Complete language (when compared to relational algebra). Subset for the first release.
- Human readable syntax and XML and RDF representations



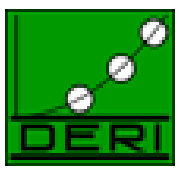
RDB2RDF mapping language goals

- Complete language (when compared to relational algebra). Subset for the first release.
- Human readable syntax and XML and RDF representations
- Mappings must be expressed using the Rule Interchange Format (RIF) [4] syntax



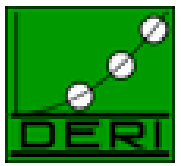
RDB2RDF mapping language goals

- Complete language (when compared to relational algebra). Subset for the first release.
- Human readable syntax and XML and RDF representations
- Mappings must be expressed using the Rule Interchange Format (RIF) [4] syntax
- Support for exposing vendor specific SQL features (e.g. spatial support)



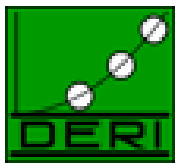
RDB2RDF mapping language goals

- Complete language (when compared to relational algebra). Subset for the first release.
- Human readable syntax and XML and RDF representations
- Mappings must be expressed using the Rule Interchange Format (RIF) [4] syntax
- Support for exposing vendor specific SQL features (e.g. spatial support)
- Mechanism to reuse public identifiers (or if necessary create new ones) for database entities



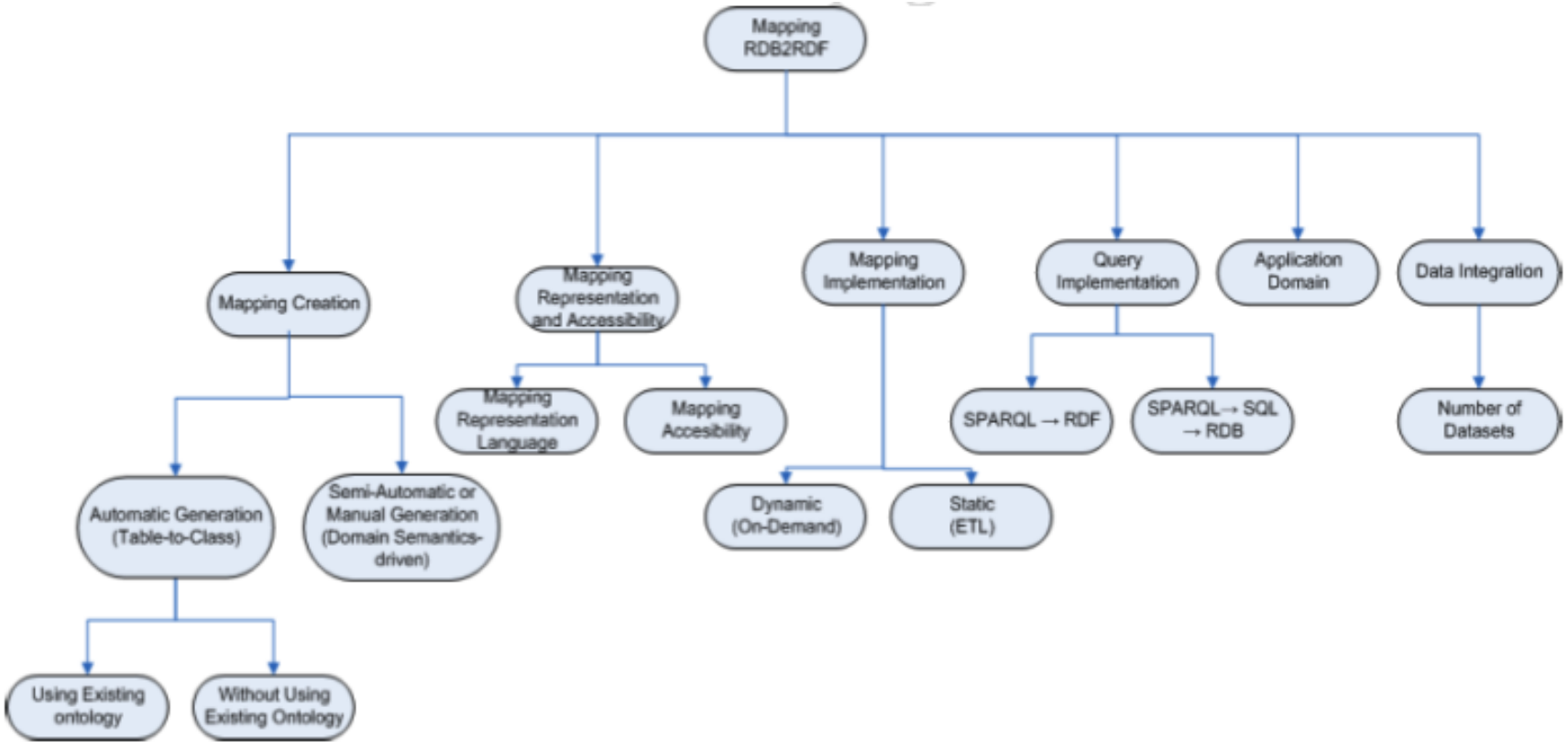
Reference Framework

"A Survey of Current Approaches for Mapping of Relational Databases to RDF" [3] defines a reference framework to compare different mapping approaches and presents a state of the art overview.



Reference Framework

"A Survey of Current Approaches for Mapping of Relational Databases to RDF" [3] defines a reference framework to compare different mapping approaches and presents a state of the art overview.
Characterizes each tool over the following aspects:



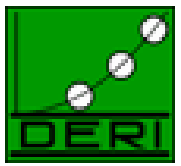
Mapping tools (I)

Overview of generic Tools and Applications:

Virtuoso RDF View: Table to class and column as predicate approach and user generated. Can use either type of Mapping and Query implementation.

D2RQ: Can use automatic and user generated mappings. Allows for either type of Mapping and Query implementation.

Triplify: Maps HTTP-URI requests to relational database queries expressed in SQL. No SPARQL support.

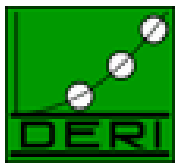


Mapping tools (II)

R2O: Can use automatic and user generated mappings. Allows for either type of Mapping and Query implementation.

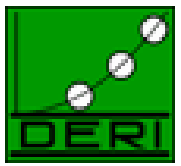
RDBToOnto: Creates automatically generated mappings (Table to Class). Static mapping (RDF dump). SPARQL on generated ontology.

SBDR, Automapper: Automatic mapping creation using the Table to Class and Columns as Properties approach. Allows for both types of Mapping implementation. SPARQL queries are rewritten to SQL.



Note worthy:

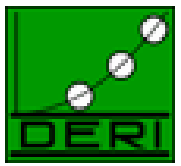
- No standard method for representation of mappings between RDB and RDF.
- Mappings should be available for re-use.



Note worthy:

- No standard method for representation of mappings between RDB and RDF.
- Mappings should be available for re-use.

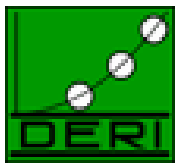
- The representation of mappings in a standardized form is necessary to enable their reuse.



Astronomy Data

Astronomy data is available as large volumes of distributed data [1]

- Use case for data integration
- Data stored in RDBs

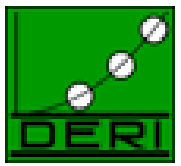


Astronomy Data

Astronomy data is available as large volumes of distributed data [1]

- Use case for data integration
- Data stored in RDBs

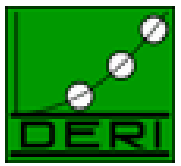
- Can SPARQL be used to express the types of query required in scientific applications?
- Can the performance of RDF triple stores and RDB2RDF tools meet the requirements of the large data sets encountered in the scientific domain?



Dataset & Queries

Dataset:

- Original dataset stored as a RDB containing 6.4 million objects
- A sample of this data is publicly available for download (14 relations, approx 1 250 000 rows)



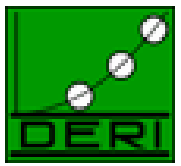
Dataset & Queries

Dataset:

- Original dataset stored as a RDB containing 6.4 million objects
- A sample of this data is publicly available for download (14 relations, approx 1 250 000 rows)

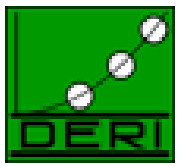
20 queries (scientific application domain)

- Large number involve mathematical functions, aggregates, ordering



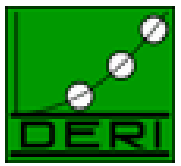
SPARQL

- Some SQL syntactic shortcuts can be expressed as SPARQL filters (e.g. *between*, *abs*, string matching, . . .)



SPARQL

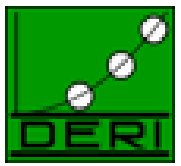
- Some SQL syntactic shortcuts can be expressed as SPARQL filters (e.g. *between*, *abs*, string matching, . . .)
- No mathematical functions
- No aggregate or grouping



SPARQL

- Some SQL syntactic shortcuts can be expressed as SPARQL filters (e.g. *between*, *abs*, string matching, . . .)
- No mathematical functions
- No aggregate or grouping

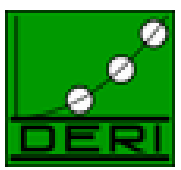
- Of 18 queries (2 were not considered due to depending on a server function), 9 are expressible in SPARQL.



Compared Systems

5 systems were compared (1 RDB, 2 RDB2RDF and 2 triple stores):

- **MySQL:** Used with the publicly available data and queries.
- **D2RQ:** Uses Jena for RDF access and data was stored in the MySQL database.
- **SquirrelRDF:** Exposes the RDB data as an RDF view and allows to query using SPARQL. Similar conditions to D2RQ.
- **Jena:** Triple store with a relational database backend. (Not the MySQL relational data).
- **Sesame:** Triple store with RDF native backend.

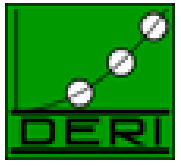


Performance results

	MySQL (ms)		D2RQ (ms)		SquirrelRDF (ms)		Jena (ms)		Sesame (ms)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Query 1	34	0.5	352	17.2	613	13.5	3,450	775.5	39	24.2
Query 2	38	1.0	5,339	36.5	21,492	259.7	485,932	169,800.6	83	12.0
Query 3	33	1.2	2,733	34.4	837	11.2	7,229	2,549.8	69	26.4
Query 5	34	2.5	4,090	43.0	1,307	10.0	17,793	8,849.5	65	13.4
Query 6	1	0.4	7,468	224.5	19,984	87.8	372,561	60,204.8	56	32.7

Queries: `http://surveys.roe.ac.uk/ssa/sqlcookbook.html#Examples`

- best results: MySQL
- best SPARQL-enabled: Sesame



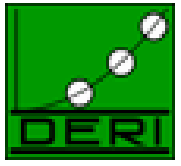
Performance results

	MySQL (ms)		D2RQ (ms)		SquirrelRDF (ms)		Jena (ms)		Sesame (ms)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Query 1	34	0.5	352	17.2	613	13.5	3,450	775.5	39	24.2
Query 2	38	1.0	5,339	36.5	21,492	259.7	485,932	169,800.6	83	12.0
Query 3	33	1.2	2,733	34.4	837	11.2	7,229	2,549.8	69	26.4
Query 5	34	2.5	4,090	43.0	1,307	10.0	17,793	8,849.5	65	13.4
Query 6	1	0.4	7,468	224.5	19,984	87.8	372,561	60,204.8	56	32.7

Queries: `http://surveys.roe.ac.uk/ssa/sqlcookbook.html#Examples`

- best results: MySQL
- best SPARQL-enabled: Sesame

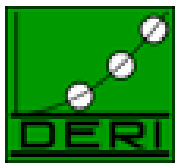
- RDF and SPARQL still can not compete with relational databases in the scientific domain



Conclusions

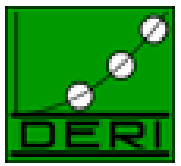
Steps forward:

- New SPARQL (1.1) features include aggregates, SPARQL functions, which provide some of the missing features
- More research on query translation and SPARQL optimisation is needed



Thank you!

Questions?



Bibliography

Alasdair J. G. Gray, Norman Gray, and Iadh Ounis.
Can RDB2RDF Tools Feasibily Expose Large Science Archives for Data Integration?
In *ESWC 2009*, volume 5554 of *LNCS*, pp 491–505. Springer.

Ashok Malhotra.
W3C RDB2RDF Incubator Group Report.
http://www.w3.org/2005/Incubator/rdb2rdf/XGR-rdb2rdf/

Satya S. Sahoo *et al.*
A Survey of Current Approaches for Mapping of Relational Databases to RDF.
W3C RDB2RDF XG Report, W3C, 2009.
http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf

W3C Rule Interchange Format Working Group.
http://www.w3.org/2005/rules/wg