# Progress Report From the W3C RDB2RDF XG

Ashok Malhotra
Oracle Corporation
Oracle Parkway
Redwood Shores, CA 94065
001-914-271-6477
malhotrasahib@gmail.com

Jim Melton
Oracle Corporation
Oracle Parkway
Redwood Shores, CA 94065
001-801.942.3345
jim.melton@acm.org

## ABSTRACT

This paper discusses the progress of the W3C RDB2RDF Incubator Group [1] which is concerned with the mapping of Relational data to RDF and to OWL ontologies

**Categories and Subject Descriptors**

**D.3 PROGRAMMING LANGUAGES**

**E.1 DATA STRUCTURES**

**H.2 DATABASE MANAGEMENT**

**General Terms**

Algorithms, Design, Standardization, Languages, Theory.

**Keywords**

Semantic Web, Relational databases, RDF, SPARQL.

## 1. INTRODUCTION

Following a successful workshop on RDF Access to Relational Databases in late October 2007 [2], the W3C RDB2RDF Incubator Group (XG) was chartered in February 2008 to provide direction in the area of mapping Relational Data to RDF and to OWL ontologies. There is a significant amount of existing work in this area and, in the last six months the XG has concentrated on understanding and classifying this work enroute to developing a recommendation to the W3C on what they may wish to standardize.

## 2. RATIONALE

There are several good reasons for mapping Relational data to RDF and OWL:

Relational databases do not contain much semantic information and what there is, is often buried in table and column names. Also, information about a single entity may be stored in multiple databases.

Therefore, mapping data from several databases to RDF and OWL, possibly, with additional semantics, carries the promise of richer and more meaningful queries over the vast volume of existing Relational data.

## 3. DIMENSIONS OF THE PROBLEM

The requirements for RDB to RDF mapping cover a broad spectrum. For example, many websites store their data in Relational databases and all that is needed is a simple and automatic mapping to RDF. On the other hand, for example, users in the life sciences deal with many interrelated databases embellished with additional semantics and rules expressed in different formats.

Another dimension is whether the mapped data is stored as RDF, perhaps in a triple store, or whether it is merely surfaced as RDF i.e. a virtual mapping is provided. Both these approaches have benefits and drawbacks. If the mapped data is stored as RDF it is directly accessible. On the other hand, Relational databases can be very large and it may be impractical to map and store huge volumes of data and refresh it every time the underlying databases change.

## 4. LITERATURE SURVEY

As part of its work the RDB2RDF XG has compiled a survey of the literature in this area in the form of a Wiki [3]. The remainder of this paper discusses the classification metrics used in this literature survey. This work is ongoing and the survey can be used as a resource for research in this area.

### 4.1 Mapping Approach

We can classify the approach used to map Relational data to RDF as either direct conversion of database schemas or ER diagrams to RDF components or mapping from a domain ontology back to the underlying databases.

The first approach takes advantage of ER diagram semantics and (in most cases) maps the table name to a class and the column name to a predicate. An example of this approach is the Virtuoso

*Conference'04*, Month 1–2, 2004, City, State, Country.

RDF View [4] that uses the unique identifier of a record (row key) as the RDF object, the column of a table as RDF predicate and the column value as the RDF object.

The second approach makes use of a domain ontology as the reference knowledge model and defines transformation rules to map class information to Relational data or SQL queries. This approach is an ontology population technique where the transformed data are instances of the ontology schema concepts.

These two approaches work from different directions. In the first approach you start from the database schema or the ER diagram and (usually automatically) generate the RDF triples or the ontology. In the second approach you, ideally, create a domain ontology without looking at the data in the databases. Then you map the ontology classes to Relational data or SQL queries on the underlying database(s). As implemented in various tools, both approaches allow greater or lesser degrees of customization and injection of additional semantics. In general, the second approach is more oriented to customization and addition of semantics. As an example, the U.K. Ordnance Survey [5] approach uses their hydrology ontology enriched with rules as the reference knowledge model to define the database mapping.

## 4.2 Mapping Representation and Access

The mapping algorithm used for conversion of RDB to RDF may be represented in a XSLT stylesheet using XPath rules or in a XML-based declarative language such as R2O [6]. To encourage reuse and extension the mappings should be extensible in a modular fashion. This is especially true if the mappings allow the incorporation of rich domain semantics.

## 4.3 Mapping Implementation

The approaches to converting RDB data to RDF can be broadly classified as either static Extract Transform Load (ETL) or a query-driven dynamic implementation. The ETL implementation, also called “RDF dump”, uses a batch process to create the RDF repository from RDB. The query-driven approach implements the conversion dynamically in response to a query. There are multiple advantages and disadvantages associated with each of these approaches. For example, the ETL approach may not reflect the most current data, while the query-driven approach may have a performance penalty due to the on-demand conversion and the additional layers required in query execution.

## 4.4 Query Implementation

The query implementation can either be a direct execution of SPARQL over a RDF repository or the SPARQL query may be mapped to SQL queries which are subsequently executed over a RDB. Note that SQL queries can take advantage of the SQL optimizer which can result in a significant performance improvement.

## 4.5 Application Domain

As discussed above, an important aspect of RDB to RDF mapping is the incorporation of domain semantics. By studying the application domain, we may be able identify unique domain-specific and some cross-domain common mapping characteristics. If virtual mapping is used, these domain semantics need to be incorporated into the translation of SPARQL queries into SQL.

## 4.6 Data Integration

A primary objective of using RDF data model is to enable the integration of data from disparate, heterogeneous Relational and non-Relational data sources. While the primary focus of the RDB2RDF XG is not on integrating data from disparate sources, this requirement has been brought up several times and several implementations support it in one form or another. See [1], [7], [8], [9], [10].

# 5. ACKNOWLEDGMENTS

Many thanks to the participants in the RDB2RDF XG and, in particular to Wolfgang Halb and Satya Sahoo who helped compile the literature survey.

# 6. REFERENCES

[1] RDB2RDF XG http://www.w3.org/2005/Incubator/rdb2rdf/

[2] W3CWorkshop on RDF Access to Relational Databases http://www.w3.org/2007/03/RdfRDB/

[3] RDB2RDF Literature Survey Wiki http://esw.w3.org/topic/Rdb2RdfXG/StateOfTheArt

[4] C. Blakeley, OpenLink Software: 2007 RDF Views of SQL Data (Declarative SQL Schema to RDF Mapping) http://portal.acm.org/citation.cfm?id=1135777.1136019&coll=GUIDE&dl=GUIDE&type=series&idx=SERIES968&part=series&WantType=Proceedings&title=WWW

[5] J. Green and C. Dolbear,. UK Ordnance Survey, Linking Ontologies to Relational Databases: 2007 http://esw.w3.org/topic/Rdb2RdfXG?action=AttachFile&do=view&target=RDB2RDF_OS.pps

[6] DB2OWL: A Tool for Aotomatic Database to Ontology Mapping, N. Cullot, R. Ghawi and K. Yetongnon. In Proc. of 15th Italian Symposium on Advanced Database Systems (SEBD 2007), pages 491-494, Torre Canne, Italy, 17-20 June 2007.

[7] E. Mena, A. Illarramendi, V. Kashyap, and A. Sheth, “OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies,” http://knoesis.wright.edu/library/download/MKSI96.pdf

[8] S. Sahoo, O. Bodenreider, J. Rutter, K. Skinner and A. Sheth. Journal of Biomedical Informatics (Special Issue: Semantic Biomedical Mashups), (in press), 2008. http://mor.nlm.nih.gov/pubs/pdf/2008-jbi-ss.pdf

[9] Z. Wu, H. Chen, H. Wang, Y. Wang, Y. Mao, J. Tang and C. Zhou. Dartgrid: A Semantic Web Toolkit for Integrating Heterogeneous Relational databases, Semantic Web Challenge at 4th International Semantic Web Conference (ISWC 2006), Athens, USA, 5-9 November 2006. http://www.aifb.uni-karlsruhe.de/WBS/ywa/publications/wu06TCM_ISWC06.pdf

[10] Asio Tools from BBN: http://bbn.com/technology/data_indexing_and_mining/asio_tool_suite