

ESWC 20011, Tutorial 8:

Publishing Relational Data on the Semantic Web

Oktie Hassanzadeh (University of Toronto), Anastasios Kementsietsidis (IBM T.J. Watson), Yannis Velegrakis (University of Trento)

The Resource Description Framework (RDF) is becoming the de-facto standard for the representation and exchange of information on the web. Evidence of this can clearly be found in initiatives like the Linking Open Data (LOD) project, where web data from a diverse set of domains like geographic locations, films and music, scientific data, and the like, are interlinked to provide one large data cloud. The adoption of RDF is even more evident as governments (most notably the ones from US and UK), and large corporations (e.g., Pfizer, Best Buy, New York Times, and the BBC) increasingly make their data available (internally or externally) as RDF.

Yet, in spite of the increased adoption, one cannot escape the fact that a vast amount of data nowadays resides in relational databases. The relational model has largely monopolized data management for the last 30 years, to a large extent due to the significant advances in efficient relational storage representation and querying. It is thus of no surprise that a considerable number of sources even in the LOD cloud of data sets offer but an RDF view of their relational back-ends.

But how exactly one makes this transition from the "old" relational world to the "new" RDF world? How, does one transfer and preserve the semantics of well-structured and distinct tables, sitting in different relational silos, into a single integrated set of RDF triples? How do the resulted RDF data look like and how can we query them? And, how do we make the RDF data part of the *web of data* by properly linking to existing sources, as opposed to creating *isolated islands* of RDF data? These are some of the key questions that this tutorial will try to answer, while reviewing the current state-of-the-art components of the bridge between the relational and the RDF worlds.

The objective of this tutorial is twofold. At a very practical level, we want to present an overview of the current work (and the related research challenges) of publishing relational data on the semantic web. At the same time, we want to superimpose the works from two communities, namely, the data management and the semantic web communities, in an effort to (a) highlight common areas of interest; and (b) promote the exchange of ideas.

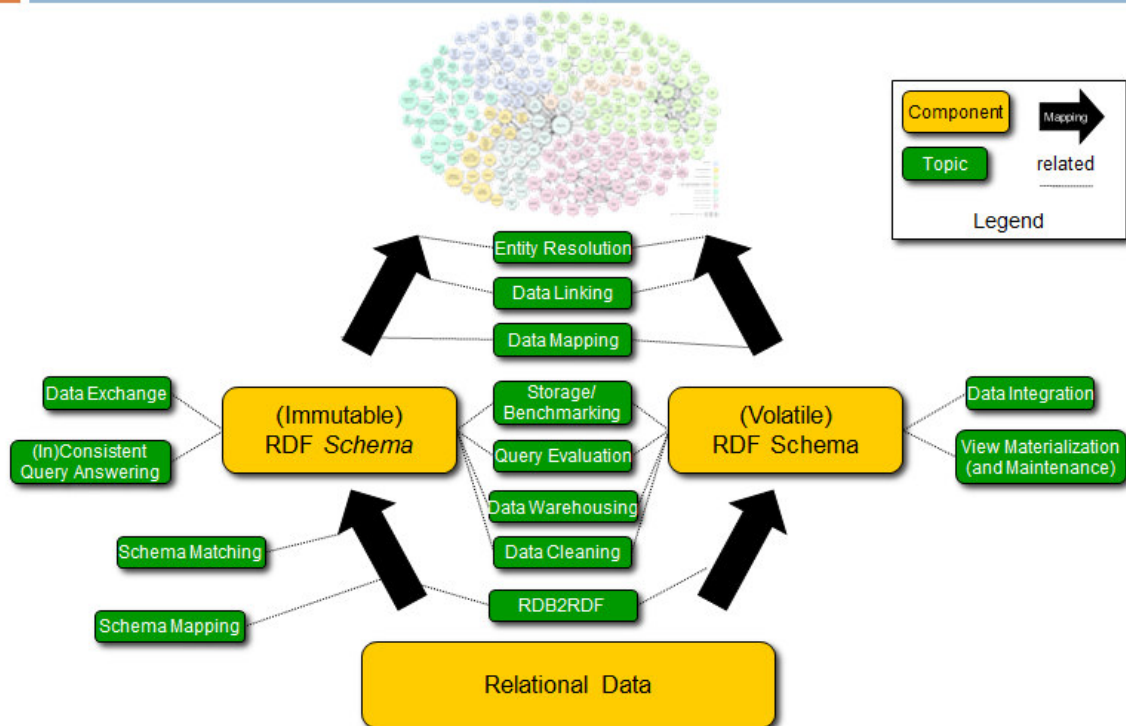
The general task starts with a relational database instance and ends up with a set of RDF data. The task is not a simple data model translation, but involves many challenging steps.

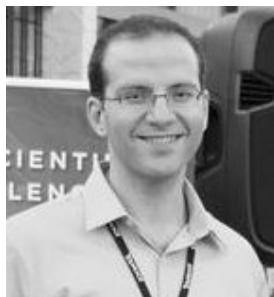
1. Decide a target (RDF) schema for your data
 - ✓ Decide between a volatile or an immutable schema
2. Decide a mapping between relational schema/data and RDF schema/data
 - ✓ Decide on a mapping language
 - ✓ Decide between manual vs. automatic mapping generation

3. Publish the relational data as RDF
 - ✓ Decide whether to materialize
 - ✓ Decide on a materialization/maintenance policy
 - ✓ Decide on a storage strategy
4. Link your published RDF data to other RDF data
 - ✓ Perform *record linkage* (also known as *entity resolution* or *duplicate detection*)
 - ✓ Perform semantic linking
5. Benchmarking: Evaluate whether the task was successful
 - ✓ Performance
 - ✓ Quality of mappings/links

During the tutorial, we will first provide a brief introduction to relational and RDF data management, and will explain the benefits that the publication of relational data in RDF has to offer. We will provide a description of the existing solutions and then we will investigate the above issues. A special focus will be put on research directions and open problems.

The following figure illustrates the research sub-topics that will be discussed during the presentation.





Oktie Hassanzadeh is a PhD candidate and an IBM PhD Fellow in the Department of Computer Science at the University of Toronto. His research interests are in the areas of data cleaning and integration, text data management and web information retrieval. He has received the 2010 Yahoo! Key Scientific Challenges award, and is a two-time recipient of the first prize at the Triplication Challenge, an annual contest that awards prizes to the most promising projects in the area of Linked Data. He holds an MSc in computer science from University of Toronto and a dual bachelors degree in Software Engineering and Hardware Engineering from Sharif University of Technology. For more information, refer to his home page: <http://www.cs.toronto.edu/~oktie>



Dr. Anastasios Kementsietsidis is a Research Staff Member at IBM T.J. Watson Research Center at Hawthorne, NY. Anastasios has a PhD in computer science from the University of Toronto. He is currently interested in various aspects of RDF data management (including, querying, storing and benchmarking RDF data). In the past, he worked (and is still interested in continuing working) on data integration, cleaning, provenance and annotation, security, as well as (distributed) query evaluation and optimization on relational or semi-structured data. He has several publications in the leading database conferences, including a best paper award in ICDE 2007, a best demo award in EDBT 2006, and his CIKM 2009 paper was a runner-up for a best paper award. He has served on the programme committee of several leading conferences and workshops and the Proceedings Chair for SIGMOD 2011. Anastasios is the Chair of the Data Management Professional Interest Community (DM PIC) at IBM T.J. Watson.



Yannis Velegrakis is a faculty member of the Department of Information Engineering and Computer Science of the University of Trento. His research area of expertise includes tool building for schema and ontology mapping, interoperability, keyword searching, semantic web, social networking, view and meta-data management. Prior to joining the University of Trento, he held a researcher position at AT&T Research Labs in the US. He has also spent time as a visitor at the University of California, Santa-Cruz, the IBM Almaden Research Center, and the Center of Advanced Studies of the IBM Toronto Lab. He was a member of the committee for the CIMI cultural profile of the ANSI/NISO Z39.50 standard. He has served in program committees of many national and international conferences, has also served as a reviewer for numerous international journals. He has been a PC and organization chair of the DEXA workshop on "Semantic Web Architectures for Enterprises" (2007), and of the ICDE Workshop "Data Engineering meets the Semantic Web" (2010 & 2011). He is also a general chair for VLDB 2013. He holds 2 US patents and has also been a Marie Curie Fellow for the period 2006-2008.