Semantic Technologies for Data Access and Integration

Diego Calvanese
Free University of Bozen-Bolzano, IT
calvanese@inf.unibz.it

Guohui Xiao
Free University of Bozen-Bolzano, IT
xiao@inf.unibz.it

KEYWORDS

Ontology-based data access; data integration; query reformulation; mappings

ACM Reference Format:
Diego Calvanese and Guohui Xiao. 2018. Semantic Technologies for Data Access and Integration. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3269206.3274272

1 INTRODUCTION

One of the main difficulties in managing information assets in large organizations is the seamless access to the complex forms of data stored in a variety of different kinds of legacy sources. Novel technologies based on an explicit use of semantics about the domain of interest have recently been proposed to address the challenges arising in this setting. In particular, *knowledge graphs* are being used as a mechanism to provide a uniform representation of the heterogeneous information stored in the sources. Such graphs are based on a simple and general format for representing data, namely *RDF triples*. This extensional information is complemented by an *ontology*, which represents general knowledge about the domain, and the data together with the ontology can be queried using the standard SPARQL language. The RDF graph can be constructed explicitly (i.e., *materialized*) from the data at the sources, by following an approach that resembles the traditional extract-transform-load (ETL) workflow. However, this approach has several drawbacks related to data duplication, freshness, and potential conflicts with data management policies and privacy requirements.

To overcome these problems, a different approach has been proposed, which is based on declaratively mapping the sources to the ontology, and maintaining the RDF graph virtual. Such an approach is known as *ontology-based data access/integration* (OBDA/I), to emphasize the prominent role played by ontologies in managing the information assets of an organization. The OBDA/I approach has been deployed successfully in several industrial projects and use cases and in the public administration, e.g., at Statoil (now Equinor), Siemens, the Italian Ministry of Economy and Finance, in projects on Smart Cities, Electronic Health Records, Maritime Security, and Manufacturing.

However, answering queries posed over the (virtual) RDF graph requires sophisticated techniques that are based on query rewriting. The quality and efficiency of the whole query answering process is highly sensitive to the form of the ontology and notably the mappings. In particular, if an OBDA/I scenario is not set-up properly, queries might be excessively slow, and they might miss answers and/or return unwanted data. Setting up a high-quality OBDA/I scenario is a non-trivial task that requires a deep understanding of the underlying principles and a good knowledge of the involved technologies and tools.

2 OVERVIEW OF THE TUTORIAL

In this tutorial, we cover such principles and the main (semantic) technologies underlying OBDA/I. In addition to the theoretical underpinning, we provide participants also with a practical hands-on experience on state-of-the-art tools. In this way, industry practitioners and researchers gain a good understanding of semantic technologies for OBDA/I, and are able to deploy them for their data access and data integration needs in practical use-cases.

Target Audience. The target audience of the tutorial are researchers, PhD students, and practitioners, who are interested in deepening their theoretical understanding of semantic technologies for accessing and integrating data, and in getting insight into recent developments in this area.

Prerequisite Knowledge. We assume from participants basic knowledge on relational database foundations (relational model, relational algebra and SQL) and technologies (use of relational engines, JDBC). Some knowledge of first-order logic might be useful, but is not required to follow the tutorial. Similarly, background in Semantic Web standards, such as RDF and SPARQL can be of help, although these notions are an integral part of the tutorial, and are introduced and discussed in the first part.

3 OUTLINE OF THE TUTORIAL

As described, we cover in the tutorial both theoretical and practical aspects related to the use of semantic technologies in OBDA/I. Specifically, the tutorial is structured in the following 6 sessions of 60 minutes each, with a total duration of 6 hours:

1. **Semantic Web Standards** (60 mins)
   In this part, we introduce and discuss the main semantic technologies that underlie OBDA/I, as they have been standardized by the W3C. Specifically, we present the *RDF data model*, and the ontology language *OWL 2 QL*, on which OBDA/I systems are based. We then introduce *SPARQL*, which is the standard query language for the Semantic Web, covering both syntax and semantics of the main language constructs, and how queries are answered taking into account ontological knowledge (which is known as *entailment regimes*). Finally, we discuss the *R2RML mapping language*, which is a language that has been tailored towards exposing relational data sources as RDF graphs.

CIKM '18, October 22–26, 2018, Torino, Italy
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6014-2/18/10.
https://doi.org/10.1145/3269206.3274272

2. **Introduction to OBDA and OBDI** (60 mins)

   In this part, we provide an introduction to the principles of OBDA/I, introducing the general OBDA/I framework. We present the ideas behind query processing, based on query rewriting with respect to the ontology, and transformation to SQL using mappings. We discuss the architecture of a typical OBDA/I system, and the external software with which such a system interacts (i.e., the OBDA/I ecosystem). Further, we present several significant use-cases in which OBDA/I has been successfully deployed.

3. **Hands-on Session 1: Basics of OBDA System Modeling and Usage** (60 mins)

   In this first hands-on session, we let participants practice the main semantic technologies introduced before, making use of the widely used Protégé ontology editor, together with its plugin for the OBDA system *Ontop*. We also practice how to deploy an OBDA system as a SPARQL endpoint, i.e., a system providing SPARQL query answering functionality via the standard HTTP protocol.

4. **Query Processing in OBDA** (60 mins)

   In this part, we provide more insights into some key theoretical aspects of OBDA, by analyzing more in depth the query answering process based on rewriting. We discuss how the ontology constructs affect query transformation, and how they impact the computational efficiency of the whole query answering process. We dedicate also some time to presenting novel optimization techniques, e.g., those relying on the use of constraints on the data, and those applied in the presence of specific SPARQL constructs such as OPTIONAL. To make this part accessible to a wide audience, it is mostly example driven.

5. **Hands-on Session 2: Advanced OBDA/I System Deployment** (60 mins)

   In this second hands-on session, we let participants practice and understand more complex forms of OBDA/I setups. The aim of this session is twofold: On the one hand, participants better understand the impact of mapping design and of optimization techniques that rely on constraints on the data. On the other hand, they learn how to connect to a federation layer to integrate multiple data sources via OBDI.

6. **Latest Advancements in OBDA/I** (60 mins)

   In this part, we provide an overview on the latest advancements in OBDA/I that concern various extension of the basic OBDA framework. Possible extensions that we consider are: *(i)* how to account for the access to non-relational (NoSQL) datasources; *(ii)* how to establish correspondences between data items in different sources, and how to access such cross-linked datasets in an integrated way; *(iii)* how to access via OBDA spatial and temporal data.

Tutorial Material. For the tutorial we use slides and we refer to a recent overview article [7]. The tutorial is freely available at http://ontop.inf.unibz.it/cikm-2018-tutorial/. For the two hands-on sessions, we expect participants to bring their own laptop, and to pre-install the software that we make available in advance from the tutorial web-page. To simplify the setup for participants, we prepare a Docker container in which all required tools and libraries are pre-installed.

4 PRESENTERS

Diego Calvanese (http://www.inf.unibz.it/~calvanese/) is a full professor at the KRDB Research Centre for Knowledge and Data, Free University of Bozen-Bolzano, Italy. His research interests include formalisms for knowledge representation and reasoning, ontology languages, description logics, conceptual data modeling, data integration, graph data management, data-aware process verification, and service modeling and synthesis. He is the author of more than 350 refereed publications, including ones in the most prestigious international journals and conferences in AI and Databases, and he is one of the editors of the Description Logic Handbook. He is associated editor of AIJ and JAIR, and member of the editorial board of JAR. He has been program chair of PODS'15 and general chair of ESSLLI'16. He is an EurAI Fellow since 2015. He has extensive experience in tutorials and keynotes, including tutorials at ISWC'07, ISWC'08, RW'09, ESSLLI'10, BigDat'15, AMW'16, EKAW'16, BigDat'17, ESSLLI'17, and BPM'17, and keynotes at BDA'12, JELIA'14, AIMSA'14, PODS'15, DL'16, AMW'16, and Ontobras'18.

Guohui Xiao (http:///www.inf.unibz.it/~gxiao/) is an assistant professor at the KRDB Research Centre for Knowledge and Data, Free University of Bozen-Bolzano, Italy. He is currently leading the development within the Ontop team. His main research interests are knowledge representation, description logics, semantic web, database theory, ontology-based data access, logic programming, and optimization and implementation of reasoning engines. His research resulted in more than 60 refereed publications. He received the Semantic Web Journal Outstanding Paper Award in 2016 and the Best In-Use Paper Award at the 16th International Semantic Web Conference (ISWC 2017). He has given tutorials about OBDA at the Summer School of the Chinese Semantic Web Conference in 2014 and 2017, at ISWC'15, EKAW'16, and IJCAI'18.

REFERENCES

[1] Diego Calvanese, Benjamin Cogrel, Sarah Komla-Ebri, Roman Kontchakov, Davide Lanti, Martin Rezk, Mariano Rodriguez-Muro, and Guohui Xiao. 2017. Ontop: Answering SPARQL Queries over Relational Databases. *Semantic Web J.* 8, 3 (2017), 471–487. https://doi.org/10.3233/SW-160217

[2] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, Mariano Rodriguez-Muro, Riccardo Rosati, Marco Ruzzi, and Domenico Fabio Savo. 2011. The Mastro System for Ontology-Based Data Access. *Semantic Web J.* 2, 1 (2011), 43–53.

[3] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. 2007. Tractable Reasoning and Efficient Query Answering in Description Logics: The *DL-Lite* Family. *J. of Automated Reasoning* 39, 3 (2007), 385–429.

[4] Antonella Poggi, Domenico Lembo, Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, and Riccardo Rosati. 2008. Linking Data to Ontologies. *J. on Data Semantics* 10 (2008), 133–173. https://doi.org/10.1007/978-3-540-77688-8_5

[5] Mariano Rodriguez-Muro and Martin Rezk. 2015. Efficient SPARQL-to-SQL with R2RML Mappings. *J. of Web Semantics* 33 (2015), 141–169. https://doi.org/10.1016/j.websem.2015.03.001

[6] Juan F. Sequeda, Marcelo Arenas, and Daniel P. Miranker. 2014. OBDA: Query Rewriting or Materialization? In Practice, Both!. In *Proc. of the 13th Int. Semantic Web Conf. (ISWC) (Lecture Notes in Computer Science)*, Vol. 8796. Springer, 535–551. https://doi.org/10.1007/978-3-319-11964-9_34

[7] Guohui Xiao, Diego Calvanese, Roman Kontchakov, Domenico Lembo, Antonella Poggi, Riccardo Rosati, and Michael Zakharyaschev. 2018. Ontology-Based Data Access: A Survey. In *Proc. of the 27th Int. Joint Conf. on Artificial Intelligence (IJCAI)*. AAAI Press, 5511–5519.

[8] Guohui Xiao, Dag Hovland, Dimitris Bilidas, Martin Rezk, Martin Giese, and Diego Calvanese. 2018. Efficient Ontology-Based Data Integration with Canonical IRIs. In *Proc. of the 15th Extended Semantic Web Conf. (ESWC) (Lecture Notes in Computer Science)*, Vol. 10843. Springer, 697–713. https://doi.org/10.1007/978-3-319-93417-4_45