

MASTRO: Ontology-Based Data Access at Work (Extended Abstract)

Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi,
Riccardo Rosati, Marco Ruzzi, Domenico Fabio Savo

Sapienza Università di Roma
Dipartimento di Ingegneria Informatica, Automatica e Sistemistica
lastname@dis.uniroma1.it

1 Introduction

In this paper we present the current version of MASTRO, a system for ontology-based data access (OBDA) developed at Sapienza Università di Roma. MASTRO allows users for accessing external data sources by querying an ontology expressed in a fragment of the W3C Web Ontology Language (OWL). As in data integration [6], mappings are used in OBDA to specify the correspondence between a unified view of the domain (called global schema in data integration terminology) and the data stored at the sources. The distinguishing feature of the OBDA approach, however, is the fact that the global schema is specified using an ontology language, which typically allows to provide a rich conceptualization of the domain of interest that is independent from the representation adopted for the data stored at the sources.

In the current version of MASTRO, ontologies are specified in $DL-Lite_{A,id,den}$, a logic of the $DL-Lite$ family of tractable Description Logics, which is specifically tailored for OBDA and is at the basis of OWL 2 QL, one of the profiles of OWL 2, the current W3C standard language for specifying ontologies. $DL-Lite_{A,id,den}$ captures the main modeling features of a variety of representation languages, such as basic ontology languages and conceptual data models [7]. Furthermore, it allows for specifying advanced forms of identification assertions [3] and denial assertions [5], which are not part of OWL 2, but are very useful in practice.

Answering unions of conjunctive queries (UCQs) in MASTRO can be done through a very efficient technique that reduces this task, via query rewriting, to standard SQL query evaluation. One key feature of the current version of MASTRO, with respect to previous ones [1], is that it adopts the Presto algorithm [8] for rewriting user queries with respect to the ontology (cf. Section 2).

In what follows we briefly describe the query answering algorithm of the MASTRO system and report on some real-world OBDA applications where MASTRO has been successfully trialed.

An extended version of this abstract can be found in [4].

2 Query Answering

In this section we describe the query answering process of the MASTRO system. In such a process, the user query, which is specified in terms of the ontology, is first rewritten into an SQL query expressed over the source relations, and then evaluated over the data sources, which are managed by a relational database management (or data federation) system (see also [1, 4]).

The rewriting algorithm, which is at the core of the answering process, is performed in three main steps (see Figure 1):

1. *TBox rewriting*: The first step rewrites the input UCQ according to the ontology. The rewriting, performed using the Presto algorithm [8], produces as output a non-recursive Datalog program, which encodes the intensional knowledge expressed by the ontology and the user query. The output Datalog program contains the definition of auxiliary predicates, not belonging to the alphabet of the ontology.
2. *Datalog Unfolding*: The output of the first step is then unfolded into a new UCQ by means of the *Datalog Unfolding* algorithm. It consists of a classic rule unfolding technique which eliminates all the auxiliary predicate symbols introduced by the Presto algorithm and produces a final UCQ expressed in terms of ontology concepts, roles, and attributes.
3. *Mapping Unfolding*: The last step takes the unfolded UCQ and the mapping assertions as input and produces an SQL query which can be directly evaluated over the data sources. In particular, the mapping assertions are first *split* into assertions of a simpler form, in which the head of every mapping assertion contains only a single ontology predicate; then, the final reformulation is produced through a mapping unfolding step, as described in [7].

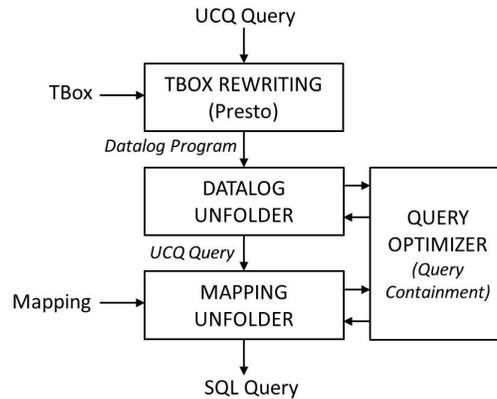


Fig. 1. The MASTRO rewriting process

Notice that the use of **Presto** is one of the key features of the current version of **MASTRO** with respect to previous ones [1]. **Presto** is an optimization of the well-known **PerfectRef** [2] algorithm. The latter may indeed produce rewritings consisting of a lot of UCQs, many of which are possibly redundant. **Presto** instead rewrites the user query into a Datalog program whose rules encode only necessary rewriting steps, thus preventing the generation of useless rules. It is important to note that after the Datalog unfolding program, one can have again an exponential number of queries, but **MASTRO** experiences on real-world applications showed a dramatic performance improvement with respect to the performance of **PerfectRef**. This is also by virtue of novel optimization techniques that exploit containment relationships between queries generated at the rewriting steps in order to reduce the size of the final SQL query.

3 The system at work: experiences on real cases

The usefulness of OBDA and the efficiency of the **MASTRO** system are showed by several industrial applications in which it has been experimented [4], e.g. in collaboration with **SELEX Sistemi Integrati**, **Banca Monte dei Paschi di Siena** and **Accenture**. As an example, we briefly report on the experiments carried out with the Italian Ministry of Economy and Finance (**MEF**), and refer the reader to [4] for descriptions on the other experiences. The main objectives of the project have been: the design and specification in $DL-Lite_{A,id,den}$ of an ontology for the domain of the Italian public debt; the realization of the mapping between the ontology and relational data sources that are part of the management accounting system currently in use at the ministry; the definition and execution of queries over the ontology aimed at extracting data of core interest for **MEF** users. In particular, the information returned by such queries relates to sales of bonds issued by the Italian government, maturities of bonds, monitoring of various financial products, and is at the basis of various reports on the overall trend of the national public debt. The Italian public dept ontology is over an alphabet containing 164 atomic concepts, 47 atomic roles, 86 attributes, and comprises around 1440 ontology assertions. The 300 mapping assertions involve around 60 relational tables managed by Microsoft **SQLServer**. We tested a very high number of queries and produced through **MASTRO** several reports of interest for the ministry.

References

1. D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, M. Rodriguez-Muro, R. Rosati, M. Ruzzi, and D. F. Savo. The Mastro system for ontology-based data access. *Semantic Web J.*, 2(1):43–53, 2011.
2. D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family. *J. of Automated Reasoning*, 39(3):385–429, 2007.

3. D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Path-based identification constraints in description logics. In *Proc. of the 11th Int. Conf. on the Principles of Knowledge Representation and Reasoning (KR 2008)*, pages 231–241, 2008.
4. G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, R. Rosati, M. Ruzzi, and D. F. Savo. Mastro: A reasoner for effective ontology-based data access. In *Proc. of the OWL Reasoner Evaluation Workshop (ORE 2012)*, volume 858 of *CEUR Electronic Workshop Proceedings*, <http://ceur-ws.org/>, 2012.
5. D. Lembo, M. Lenzerini, R. Rosati, M. Ruzzi, and D. F. Savo. Inconsistency-tolerant first-order rewritability of DL-Lite with identification and denial assertions. In *Proc. of the 25th Int. Workshop on Description Logic (DL 2012)*, volume 846 of *CEUR Electronic Workshop Proceedings*, <http://ceur-ws.org/>, 2012.
6. M. Lenzerini. Data integration: A theoretical perspective. In *Proc. of the 21st ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems (PODS 2002)*, pages 233–246, 2002.
7. A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati. Linking data to ontologies. *J. on Data Semantics*, X:133–173, 2008.
8. R. Rosati and A. Almatelli. Improving query answering over *DL-Lite* ontologies. In *Proc. of the 12th Int. Conf. on the Principles of Knowledge Representation and Reasoning (KR 2010)*, pages 290–300, 2010.