Mapping Analysis in Ontology-based Data Access: Algorithms and Complexity (Extended Abstract)

Domenico Lembo[1], José Mora[1], Riccardo Rosati[1],
Domenico Fabio Savo[1], Evgenij Thorstensen[2]

[1] DIAG, Sapienza Università di Roma
lastname@dis.uniroma1.it

[2] Dept. of Informatics, University of Oslo
evgenit@ifi.uio.no

1 Introduction

Ontology-based data access (OBDA) is a recent paradigm for accessing *data sources* through an *ontology* that acts as a conceptual, integrated view of the data, and declarative *mappings* that connect the ontology to the data sources [13, 6, 14, 5, 2] .

One important aspect in OBDA concerns the construction of a system specification, i.e., defining the ontology and the mappings over an existing set of data sources. Mappings are indeed the most complex part of an OBDA specification, since they have to capture the semantics of the data sources and express such semantics in terms of the ontology. The first experiences in the application of the OBDA framework in real-world scenarios (e.g., [2, 9]) have shown that the semantic distance between the conceptual and the data layer is often very large, because data sources are mostly application-oriented: this often makes the definition, debugging, and maintenance of mappings a hard and complex task. Such experiences have clearly shown the need of tools for supporting the management of mappings.

The recent work [11] has started providing a theoretical basis for mapping management support in OBDA, focusing on the formal analysis of mappings in ontology-based data access. In particular, the two most important semantic anomalies of mappings have been analyzed: inconsistency and redundancy. Roughly speaking, an inconsistent mapping for an ontology and a source schema is a specification that gives rise to logical contradictions with the ontology and/or the source schema. Then, a mapping $\mathcal{M}$ is redundant with respect to an OBDA specification if adding the mapping $\mathcal{M}$ to the specification does not change its semantics. The work presented in [11] has defined both a *local* notion of mapping inconsistency and redundancy, which focuses on single mapping assertions, and a *global* notion, where inconsistency and redundancy is considered with respect to a whole mapping specification (set of mapping assertions).

In this paper, we study the computational properties of verifying both local and global mapping inconsistency and redundancy in an OBDA specification. We consider a wide range of ontology languages that comprises the description logics underlying OWL 2 and all its profiles (OWL 2 EL, OWL 2 QL, and OWL 2 RL),[1] and examine mapping languages of different expressiveness (the so-called GAV and GLAV mappings [7]) over sources corresponding to relational databases. We provide algorithms and establish tight complexity bounds for the decision problems associated with both local and global mapping inconsistency and mapping redundancy, and for both combined complexity and TBox complexity (which only considers the size of the TBox).

[1] http://www.w3.org/TR/owl2-profiles/

The outcome of our analysis is twofold:

- in our framework, it is possible to define *modular techniques* that are able to reduce the analysis of mappings to the composition of standard reasoning tasks over the ontology (inconsistency, instance checking, query answering) and over the data sources (query answering and containment). This is a non-trivial result, because mappings are formulas combining both ontology and data source elements;
- the above forms of mapping analysis enjoy *nice computational properties*, in the sense that they are not harder than the above mentioned standard reasoning tasks over the ontology and the data sources (see Figure 1 and Figure 2) .

According to the above results, in our OBDA framework, the analysis of mappings is feasible for languages with nice computational properties, like the three OWL profiles.

2 Theoretical background

An OBDA specification is a triple $\mathcal{J} = \langle \mathcal{T}, \mathcal{S}, \mathcal{M} \rangle$, where $\mathcal{T}$ is a DL TBox, $\mathcal{S}$ is a source schema, and $\mathcal{M}$ is a mapping between the two. In this paper, we consider TBoxes specified through DLs that are the logical basis of the W3C standard OWL and of its profiles, i.e., $\mathcal{SROIQ}$ [8], which underpins OWL 2, *DL-Lite*$_R$ [4], which is the basis of OWL 2 QL, RL [10], a simplified version of OWL 2 RL, and $\mathcal{EL}_\bot$, a slight extension of the DL $\mathcal{EL}$ [3], which is the basis of OWL 2 EL. The source schema is assumed to be relational, and we consider both *simple schemas*, i.e., without integrity constraints, and *FD schemas*, i.e., simple schemas with functional dependencies [1]. The *mapping* is a set of assertions m of the form $\phi(\boldsymbol{x}) \rightsquigarrow \psi(\boldsymbol{x})$, where $\phi(\boldsymbol{x})$, called the *body of* m, and $\psi(\boldsymbol{x})$, called the *head of* m, are conjunctive queries (CQs) over $\mathcal{S}$ and $\mathcal{T}$, respectively. We use $head(m)$ and $body(m)$ to denote the head and the body of m.

Mappings of the form above are called *GLAV*, and are the most expressive commonly studied mappings [12, 7]. Besides them, we refer also to *GLAV*$_{BE}$ mappings, which are GLAV mappings where $\psi(\boldsymbol{x})$ is a CQ with a bounded number of occurrences of existential variables, and to *GAV* mappings, which are GLAV mappings where $head(m)$ does not admit existential variables.

The semantics of an OBDA specification $\mathcal{J} = \langle \mathcal{T}, \mathcal{S}, \mathcal{M} \rangle$ is given in terms of first-order interpretations that satisfy both $\mathcal{T}$ and $\mathcal{M}$, given a source instance D legal for $\mathcal{S}$, i.e., an instance for $\mathcal{S}$ that satisfies the constraints of $\mathcal{S}$. We denote with $Mod(\mathcal{J}, D)$ the set of models of $\mathcal{J}$ w.r.t. D We also say that a mapping assertion m *is active on a source instance* D if the evaluation of the query $body(m)$ over D is non-empty. A mapping $\mathcal{M}$ is active on D if all its mapping assertions $m \in \mathcal{M}$ are active on D.

Below we recall the definitions given in [11] that formalize the mapping analysis services that we study in this paper. Given a TBox $\mathcal{T}$, a source schema $\mathcal{S}$, a mapping assertion m, a mapping $\mathcal{M}$, and an OBDA specification $\mathcal{J} = \langle \mathcal{T}, \mathcal{S}, \mathcal{M} \rangle$, we have that

- m *is (locally) inconsistent for* $\langle \mathcal{T}, \mathcal{S} \rangle$ if m is head-inconsistent for $\mathcal{T}$, i.e., $\mathcal{T} \models \forall \boldsymbol{x}.(\neg\psi(\boldsymbol{x}))$, or m is body-inconsistent for $\mathcal{S}$, i.e., $\mathcal{S} \models \forall \boldsymbol{x}.(\neg\phi(\boldsymbol{x}))$.
- $\mathcal{M}$ *is globally inconsistent for* $\langle \mathcal{T}, \mathcal{S} \rangle$ if there does not exist a source instance D legal for $\mathcal{S}$ such that $\mathcal{M}$ is active on D and $Mod(\mathcal{J}, D) \neq \emptyset$.
- A mapping $\mathcal{M}'$ *is globally redundant for* $\mathcal{J}$ if, for every source instance D that is legal for $\mathcal{S}$, $Mod(\langle \mathcal{T}, \mathcal{S}, \mathcal{M} \rangle, D) = Mod(\langle \mathcal{T}, \mathcal{S}, \mathcal{M} \cup \mathcal{M}' \rangle, D)$.

Local mapping redundancy is a special case of global mapping redundancy in which the mappings $\mathcal{M}$ and $\mathcal{M}'$ are both composed of a single assertion.

	GAV				GLAV			
task	*DL-Lite*$_R$	RL	$\mathcal{EL}_\bot$	$\mathcal{SROIQ}$	*DL-Lite*$_R$	RL	$\mathcal{EL}_\bot$	$\mathcal{SROIQ}$
local inc.	=NLOGSPACE	=P	=P	=N2EXPTIME	=NLOGSPACE	=P	=P	=N2EXPTIME
global inc.	=NLOGSPACE	=P	=P	=N2EXPTIME	=NLOGSPACE	=P	=P	=N2EXPTIME
local red.	=NLOGSPACE	=P	=P	=N2EXPTIME	=NP	=NP	=NP	open
global red.	=NLOGSPACE	=P	=P	=N2EXPTIME	=NP	=NP	=NP	open

Fig. 1. TBox compl. of mapping inconsistency and redundancy (for both simple and FD schemas).

	GAV				GLAV$_{\text{BE}}$			
task	*DL-Lite*$_R$	RL	$\mathcal{EL}_\bot$	$\mathcal{SROIQ}$	*DL-Lite*$_R$	RL	$\mathcal{EL}_\bot$	$\mathcal{SROIQ}$
local inc.	=NLOGSPACE (SI) =P (FD)	=P	=P	=N2EXPTIME	=NLOGSPACE (SI)* =P (FD)*	=P*	=P*	=N2EXPTIME*
global inc.	=NP	=NP	=NP	=N2EXPTIME	=NP	=NP	=NP	=N2EXPTIME
local red.	=NP	=NP	=NP	=N2EXPTIME	=NP	=NP	=NP	open
global red.	=NP	=NP	=NP	=N2EXPTIME	=NP	=NP	=NP	open

Fig. 2. Combined complexity of mapping inconsistency and redundancy (SI = simple schemas, FD = FD schemas). * The result also holds for arbitrary GLAV mappings.

3 Complexity Results

We summarize below our complexity results. We consider both TBox complexity, i.e., the complexity computed w.r.t. the size of the TBox only, and combined complexity.

For both simple and FD schemas, and for both GAV and GLAV mappings, the TBox complexity of *local mapping inconsistency* turns out to be the same as the TBox complexity of ontology inconsistency. As for the combined complexity, simple and FD schemas behave differently. For simple schemas, it is not necessary to check body-inconsistency (since there are no constraints in $\mathcal{S}$), and thus the combined complexity is the same as for mapping head-inconsistency, which in turn is the same as the combined complexity of ontology inconsistency. For FD schemas we further need to check whether the mapping assertion is body-consistent, which can be done in PTIME. Combining together this result with the above bounds for simple schemas, we obtain the exact bounds for combined complexity shown in Fig. 2.

Global inconsistency can be reduced to checking the consistency of an OBDA specification w.r.t. a (minimal) source database that activates $\mathcal{M}$. In particular we have that for both simple and FD schemas, for both GAV and GLAV mappings, the TBox complexity of global mapping inconsistency is the same as the TBox complexity of ontology inconsistency. As for combined complexity, we devise a non-deterministic algorithm exploiting the above mentioned correspondence of global mapping inconsistency and OBDA inconsistency. This algorithm allows us to prove that for both simple and FD schemas, and for both GAV and GLAV$_{\text{BE}}$ mappings, it holds that: (i) if the ontology language is *DL-Lite*$_R$, RL, or $\mathcal{EL}_\bot$, then the combined complexity of global mapping inconsistency is in NP; (ii) if the ontology language is $\mathcal{SROIQ}$, then it is in N2EXPTIME. We also prove that these bounds are in fact exact.

As for *redundancy*, our investigation shows that both local and global redundancy have the same computational behaviour. The complexity results are obtained with techniques that resemble those used for establishing complexity of global inconsistency. All our complexity results are reported in the tables in Fig. 1 and Fig. 2.

Acknowledgments. This research has been partially supported by the EU under FP7 Large-scale integrating project Optique (grant n. FP7-318338).

References

1. Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison Wesley Publ. Co., 1995.
2. Natalia Antonioli, Francesco Castanò, Spartaco Coletta, Stefano Grossi, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, Emanuela Virardi, and Patrizia Castracane. Ontology-based data management for the Italian public debt. In *Proc. of FOIS 2014*, pages 372–385, 2014.
3. Franz Baader, Sebastian Brandt, and Carsten Lutz. Pushing the $\mathcal{EL}$ envelope. In *Proc. of IJCAI 2005*, pages 364–369, 2005.
4. Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family. *J. of Automated Reasoning*, 39(3):385–429, 2007.
5. Diego Calvanese, Martin Giese, Peter Haase, Ian Horrocks, Thomas Hubauer, Yannis E. Ioannidis, Ernesto Jiménez-Ruiz, Evgeny Kharlamov, Herald Kllapi, Johan W. Klüwer, Manolis Koubarakis, Steffen Lamparter, Ralf Möller, Christian Neuenstadt, T. Nordtveit, Özgür L. Özçep, Mariano Rodriguez-Muro, Mikhail Roshchin, Domenico Fabio Savo, Michael Schmidt, Ahmet Soylu, Arild Waaler, and Dmitriy Zheleznyakov. Optique: OBDA solution for big data. In *Proc. of ESWC 2013 Satellite Events*, volume 7955 of *LNCS*, pages 293–295. Springer, 2013.
6. Cristina Civili, Marco Console, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Lorenzo Lepore, Riccardo Mancini, Antonella Poggi, Riccardo Rosati, Marco Ruzzi, Valerio Santarelli, and Domenico Fabio Savo. MASTRO STUDIO: Managing ontology-based data access applications. *PVLDB*, 6:1314–1317, 2013.
7. AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. *Principles of Data Integration*. Morgan Kaufmann, 2012.
8. Ian Horrocks, Oliver Kutz, and Ulrike Sattler. The even more irresistible $\mathcal{SROIQ}$. In *Proc. of KR 2006*, pages 57–67, 2006.
9. Evgeny Kharlamov, Martin Giese, Ernesto Jimnez-Ruiz, Martin G. Skjveland, Ahmet Soylu, Dmitriy Zheleznyakov, Timea Bagosi, Marco Console, Peter Haase, Ian Horrocks, Sarunas Marciuska, Christoph Pinkel, Mariano Rodriguez-Muro, Marco Ruzzi, Valerio Santarelli, Domenico Fabio Savo, Kunal Sengupta, Michael Schmidt, Evgenij Thorstensen, Johannes Trame, and Arild Waaler. Optique 1.0: Semantic access to big data: The case of Norwegian petroleum directorate's factpages. In *Proc. of ISWC 2013 Posters & Demos Track*, pages 65–68, 2013.
10. Roman Kontchakov and Michael Zakharyaschev. An introduction to Description Logics and query rewriting. In Manolis Koubarakis, Giorgos B. Stamou, Giorgos Stoilos, Ian Horrocks, Phokion G. Kolaitis, Georg Lausen, and Gerhard Weikum, editors, *RW 2014 Tutorial Lectures*, volume 8714 of *LNCS*, pages 195–244. Springer, 2014.
11. Domenico Lembo, José Mora, Riccardo Rosati, Domenico Fabio Savo, and Evgenij Thorstensen. Towards mapping analysis in ontology-based data access. In *Proc. of RR 2014*, volume 8741 of *LNCS*, pages 108–123. Springer, 2014.
12. Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proc. of PODS 2002*, pages 233–246, 2002.
13. Antonella Poggi, Domenico Lembo, Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, and Riccardo Rosati. Linking data to ontologies. *J. on Data Semantics*, X:133–173, 2008.
14. Mariano Rodriguez-Muro, Roman Kontchakov, and Michael Zakharyaschev. Ontology-based data access: Ontop of databases. In *Proc. of ISWC 2013*, volume 8218 of *LNCS*, pages 558–573. Springer, 2013.