Exchanging Description Logic Knowledge Bases

Marcelo Arenas
Department of Computer Science
PUC Chile
marenas@ing.puc.cl

Elena Botoeva
Diego Calvanese
Vladislav Ryzhikov
KRDB Research Centre for Knowledge and Data
Free University of Bozen-Bolzano, Italy
lastname@inf.unibz.it

Evgeny Sherkhonov
ISLA Intelligent Systems Lab
University of Amsterdam, Netherlands
e.sherkhonov@uva.nl

Abstract

In this paper, we study the problem of exchanging knowledge between a source and a target knowledge base (KB), connected through mappings. Differently from the traditional database exchange setting, which considers only the exchange of data, we are interested in exchanging implicit knowledge. As representation formalism we use Description Logics (DLs), thus assuming that the source and target KBs are given as a DL TBox+ABox, while the mappings have the form of DL TBox assertions. We study the problem of translating the knowledge in the source KB according to these mappings. We define a general framework of KB exchange, and address the problems of representing implicit source information in the target, and of computing different kinds of solutions, i.e., target KBs with specified properties, given a source KB and a mapping. We develop first results and study the complexity of KB exchange for *DL-Lite*$_{RDFS}$, a DL corresponding to the FOL fragment of RDFS, and for *DL-Lite*$_{\mathcal{R}}$.

1 Introduction

In data exchange, data structured under one schema (called source schema) must be restructured and translated into an instance of a different schema (called target schema) as it is specified by a mapping from the source schema to the target schema (Fagin et al. 2005). Such a problem has been studied extensively in recent years, under various choices for the languages used to specify the source and target schema, and the mappings (Barceló 2009). While incomplete information in this setting is introduced by the mapping layer (see also (Libkin and Sirangelo 2011)), one fundamental assumption in the works on data exchange is that the source is a (completely specified) database.

In this paper, we go beyond this setting by following the line of the work in (Arenas, Pérez, and Reutter 2011), where a general framework for data exchange is proposed, in which the source data may be incompletely specified, and thus (possibly infinitely) many source instances are implicitly represented. We refine that framework to the case where source and target are represented by description logic (DL) knowledge bases (KBs) constituted by a TBox (implicit information) and an ABox (explicit information), and where mappings are sets of DL inclusions. In such a setting, in order to minimize the exchange (and hence transfer and materialization) of explicit (i.e., ABox) information, we are interested in computing translations, from now on referred to as solutions, that contain as much implicit knowledge as possible. This leads us to define the novel notion of *representability*, which helps us in understanding the capacity of solutions to transfer implicit knowledge. Checking representability and computing a representation of a source TBox under a mapping turn out to be crucial problems in the context of knowledge base exchange.

Furthermore, we argue that the right notion of solution, on which to base our investigations, should not be the standard one based on the correspondence between models of source and target KBs. Indeed, we show that such solutions present severe limitations since, on the one hand, they do not allow for the use of implicit target information to represent implicit source information, and on the other hand, they may lead to exponentially large target ABoxes. To overcome these drawbacks, we introduce the weaker notion of $\mathcal{Q}$*-solution*, for a query language $\mathcal{Q}$, which is based on the correspondence between answers to queries in $\mathcal{Q}$ over source and target KBs. Notice that such a notion, though weaker, is in line with the objective of (data and) knowledge base exchange of providing in the target sufficient information to answer queries in $\mathcal{Q}$ that could also be posed over the source.

We then develop results and techniques for KB exchange and for the $\mathcal{Q}$-representability problem in the case where $\mathcal{Q}$ are unions of conjunctive queries (UCQs), and where KBs are expressed in *DL-Lite*$_{RDFS}$, a member of the *DL-Lite* family (Calvanese et al. 2007) that corresponds to the FOL fragment of RDFS (Brickley and Guha 2004), the widely adopted standard Semantic Web language.

2 Preliminaries

The DLs of the *DL-Lite* family (Calvanese et al. 2007) are characterized by the fact that reasoning can be done in polynomial time, and that data complexity of reasoning and conjunctive query answering is in AC^0. Here, we adopt *DL-Lite*$_{\mathcal{R}}$, a prominent member of the *DL-Lite* family, where *roles* R are either atomic roles, denoted P, or their inverses P^-, and *concepts* B are either atomic, denoted A, or of the form $\exists R$. In the following, we use N to denote either a concept or a role. A *DL-Lite*$_{\mathcal{R}}$ TBox is constituted by a finite set of concept and role inclusions $N_1 \sqsubseteq N_2$, and of concept and role disjointness assertions $N_1 \sqsubseteq \neg N_2$. We call *DL-Lite*$_{RDFS}$ the fragment of *DL-Lite*$_{\mathcal{R}}$ in which there are no disjointness assertions and only atomic concepts and atomic roles in the right-hand side of inclusions. ABoxes, KBs, and their semantics are defined as usual, see (Calvanese et al. 2007) for details. We just remark that we adopt the *standard name assumption*, that is, we assume given a fixed infinite set $\mathbf{U}$ of individuals, and we assume that for every interpretation $\mathcal{I}$, it holds that $\Delta^{\mathcal{I}} \subseteq \mathbf{U}$ and $a^{\mathcal{I}} = a$ for every individual a. This also implies that interpretations satisfy the *unique name assumption* over individuals.

A *signature* Σ is a set of concept and role names. An interpretation $\mathcal{I}$ is said to be an interpretation of Σ if it is defined exactly on the concept and role names in Σ. Given a KB $\mathcal{K}$, the *signature* $\Sigma(\mathcal{K})$ *of* $\mathcal{K}$ is the alphabet of concept and role names occurring in $\mathcal{K}$, and $\mathcal{K}$ is said to be *defined over* (or simply, *over*) a signature Σ if $\Sigma(\mathcal{K}) \subseteq \Sigma$ (and likewise for a TBox $\mathcal{T}$, an ABox $\mathcal{A}$, and a concept or role inclusion $N_1 \sqsubseteq N_2$). A k-ary query q over a signature Σ, with $k \geq 0$, is a function that maps every interpretation $\langle \Delta^{\mathcal{I}}, \cdot^{\mathcal{I}} \rangle$ of Σ into a k-ary relation $q^{\mathcal{I}} \subseteq \Delta^k$.

As our main query formalism we adopt unions of conjunctive queries (UCQs), with certain answer semantics. In particular, given a query q and a knowledge base $\mathcal{K}$, term $cert(q, \mathcal{K})$ is used to denote the set of certain answers for q over $\mathcal{K}$. Note that in *DL-Lite*$_{\mathcal{R}}$ certain answers can be characterized through the notion of chase (for more details see, e.g., (Calvanese et al. 2007)).

3 Exchanging Knowledge Bases

In this section, we introduce the knowledge exchange framework used in the paper. The starting point to define this framework is the notion of mapping. Assume that Σ_1, Σ_2 are signatures with no concepts or roles in common. Then we say that an inclusion $N_1 \sqsubseteq N_2$ is an *inclusion from* Σ_1 *to* Σ_2, if N_1 is a concept or a role over Σ_1 and N_2 is a concept or a role over Σ_2. For a DL $\mathcal{L}$ (e.g., *DL-Lite*$_{\mathcal{R}}$), we define an $\mathcal{L}$*-mapping* (or just *mapping*, when $\mathcal{L}$ is clear from the context) as a tuple $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$, where $\mathcal{T}_{12}$ is a TBox in $\mathcal{L}$ consisting of concept and role inclusions from Σ_1 to Σ_2.

Let $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$ be a mapping. Intuitively, $\mathcal{M}$ specifies how a KB over the vocabulary Σ_1 should be translated into a KB over the vocabulary Σ_2. More specifically, given an interpretation $\mathcal{I}_1$ of Σ_1 and an interpretation $\mathcal{I}_2$ of Σ_2, pair $(\mathcal{I}_1, \mathcal{I}_2)$ *satisfies* TBox $\mathcal{T}_{12}$, denoted by $(\mathcal{I}_1, \mathcal{I}_2) \models \mathcal{T}_{12}$, if for each concept inclusion $C_1 \sqsubseteq C_2 \in \mathcal{T}_{12}$, it holds that ${C_1}^{\mathcal{I}_1} \subseteq {C_2}^{\mathcal{I}_2}$, and for each role inclusion $Q_1 \sqsubseteq Q_2 \in \mathcal{T}_{12}$, it holds that ${Q_1}^{\mathcal{I}_1} \subseteq {Q_2}^{\mathcal{I}_2}$. Moreover, given an interpretation $\mathcal{I}$ of Σ_1, $\textsc{Sat}_{\mathcal{M}}(\mathcal{I})$ is defined as the set of interpretations $\mathcal{J}$ of Σ_2 such that $(\mathcal{I}, \mathcal{J}) \models \mathcal{T}_{12}$, and given a set $\mathcal{X}$ of interpretations of Σ_1, $\textsc{Sat}_{\mathcal{M}}(\mathcal{X})$ is defined as $\textsc{Sat}_{\mathcal{M}}(\mathcal{X}) = \bigcup_{\mathcal{I} \in \mathcal{X}} \textsc{Sat}_{\mathcal{M}}(\mathcal{I})$. This notion of satisfaction is the key ingredient in the definition of the notion of solution under a mapping, which is a reformulation of the concept of solution for representation systems proposed in (Arenas, Pérez, and Reutter 2011).

Definition 1. *Let* $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$ *be a mapping,* $\mathcal{K}_1$ *a KB over* Σ_1*, and* $\mathcal{K}_2$ *a KB over* Σ_2*. Then* $\mathcal{K}_2$ *is a* solution *for* $\mathcal{K}_1$ *under* $\mathcal{M}$ *if* $\textsc{Mod}(\mathcal{K}_2) \subseteq \textsc{Sat}_{\mathcal{M}}(\textsc{Mod}(\mathcal{K}_1))$.

In other words, a target KB $\mathcal{K}_2$ is a solution for a source KB $\mathcal{K}_1$ under a mapping $\mathcal{M}$ if for every model $\mathcal{I}_2$ of $\mathcal{K}_2$, there exists a model $\mathcal{I}_1$ of $\mathcal{K}_1$ such that $(\mathcal{I}_1, \mathcal{I}_2) \models \mathcal{T}_{12}$.

Next we introduce the notion of *universal solution*, which is a simple extension of the concept of solution introduced in Definition 1, based on the notion of universal solution proposed in (Arenas, Pérez, and Reutter 2011).

Definition 2. *Let* $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$ *be a mapping,* $\mathcal{K}_1$ *a KB over* Σ_1*, and* $\mathcal{K}_2$ *a KB over* Σ_2*. Then* $\mathcal{K}_2$ *is a* universal solution *for* $\mathcal{K}_1$ *under* $\mathcal{M}$ *if* $\textsc{Mod}(\mathcal{K}_2) = \textsc{Sat}_{\mathcal{M}}(\textsc{Mod}(\mathcal{K}_1))$.

In this definition, KB $\mathcal{K}_2$ is considered a good solution for KB $\mathcal{K}_1$ under mapping $\mathcal{M}$ as the models of $\mathcal{K}_2$ exactly correspond to the valid translations of the models of $\mathcal{K}_1$ according to $\mathcal{M}$. We illustrate Definitions 1 and 2 in an example.

Example 1. *Let* $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$*, where* $\Sigma_1 = \{A_1(\cdot), B_1(\cdot)\}$*,* $\Sigma_2 = \{A_2(\cdot), B_2(\cdot)\}$*, and* $\mathcal{T}_{12} = \{A_1 \sqsubseteq A_2,\ B_1 \sqsubseteq B_2\}$*. Furthermore, assume that* $\mathcal{K}_1 = \langle \mathcal{T}_1, \mathcal{A}_1 \rangle$*, where* $\mathcal{T}_1 = \{B_1 \sqsubseteq A_1\}$ *and* $\mathcal{A}_1 = \{B_1(a)\}$*. Then the KB* $\mathcal{K}_2 = \langle \mathcal{T}_2, \mathcal{A}_2 \rangle$*, where* $\mathcal{T}_2 = \emptyset$ *and* $\mathcal{A}_2 = \{B_2(a), A_2(a)\}$*, is a universal solution for* $\mathcal{K}_1$ *under* $\mathcal{M}$*.* ■

In the data exchange scenario (Fagin et al. 2005; Barceló 2009), as well as in the knowledge exchange scenario (Arenas, Pérez, and Reutter 2011), the problems of defining a notion of good solution and of finding algorithms to materialize them are arguably the most important problems to solve. When finding a notion of good solution in the context of knowledge exchange, one has to take into consideration the fact that good solutions should transfer as much implicit knowledge as possible in this scenario. To this end, next we define two properties that will help us to understand the capacity of universal solutions, and also of the query-languages based notions of solutions that will be introduced in Section 5, to transfer implicit knowledge. In what follows, we use *chase*$_{\mathcal{T}}(\mathcal{A})$ to denote the chase of $\mathcal{A}$ w.r.t. $\mathcal{T}$ (as defined in (Calvanese et al. 2007)), and we use *chase*$_{\mathcal{T},\Sigma}(\mathcal{A})$ to denote the projection of *chase*$_{\mathcal{T}}(\mathcal{A})$ on the signature Σ.

Definition 3. *Let* $\mathcal{L}$ *be a DL,* $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$ *an* $\mathcal{L}$*-mapping, and* $\mathcal{T}_1$ *an* $\mathcal{L}$*-TBox over* Σ_1*. Then,*

- $\mathcal{T}_1$ *is* representable under $\mathcal{M}$ *if there exists an* $\mathcal{L}$*-TBox* $\mathcal{T}_2$ *over* Σ_2*, called a* representation of $\mathcal{T}_1$ under $\mathcal{M}$*, such that for every ABox* $\mathcal{A}_1$ *over* Σ_1*, if* $\langle \mathcal{T}_1 \cup \mathcal{T}_{12}, \mathcal{A}_1 \rangle$ *is consistent, then* $\langle \mathcal{T}_2, \textit{chase}_{\mathcal{T}_{12},\Sigma_2}(\mathcal{A}_1) \rangle$ *is a universal solution for* $\langle \mathcal{T}_1, \mathcal{A}_1 \rangle$ *under* $\mathcal{M}$*.*
- $\mathcal{T}_1$ *is* weakly representable under $\mathcal{M}$ *if there exists a mapping* $\mathcal{M}^{\star} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12}^{\star})$ *such that* $\mathcal{T}_{12} \subseteq \mathcal{T}_{12}^{\star}$*,* $\mathcal{T}_1 \cup \mathcal{T}_{12} \models \mathcal{T}_{12}^{\star}$*, and* $\mathcal{T}_1$ *is representable under* $\mathcal{M}^{\star}$*.*

Some remarks about the definition of representability need to be made. First, notice that in the definition of representability, target TBox $\mathcal{T}_2$ depends only on the source TBox $\mathcal{T}_1$ and the mapping $\mathcal{M}$. Thus, representability of $\mathcal{T}_1$ would mean that we can construct the TBox of a solution by considering only $\mathcal{T}_1$ and $\mathcal{M}$, independently of the source ABox. Second, our definition of representability takes into account that the implicit knowledge of the source TBox $\mathcal{T}_1$ is represented "entirely" in the TBox of the solution, so that the only knowledge that remains to be transferred via the assertions in the mapping $\mathcal{M}$ is the explicit knowledge of the source ABox. Third, notice that in the computation of the universal solution $\langle \mathcal{T}_2, \mathit{chase}_{\mathcal{T}_{12},\Sigma_2}(\mathcal{A}_1)\rangle$ from $\langle \mathcal{T}_1, \mathcal{A}_1\rangle$, the chase is used to translate the input ABox $\mathcal{A}_1$ according to the assertions in $\mathcal{M}$. The reason for this is that the chase has been shown to be the right tool to compute solutions in the scenario where one is given explicit source data (a relational database or an ABox) and a mapping (Fagin et al. 2005; Barceló 2009). Fourth, notice that we only consider in the definition of representability source ABoxes $\mathcal{A}_1$ such that $\langle \mathcal{T}_1 \cup \mathcal{T}_{12}, \mathcal{A}_1\rangle$ is consistent, that is, we consider only knowledge bases that can be effectively translated through the mapping $\mathcal{M}$. Finally, the main goal of this paper is to find polynomial time algorithms to solve the representability problem, as the existence of such algorithms would allow one to materialize good solutions of polynomial size, that is, solutions that can be effectively used in practice.

4 Are Universal Solutions Appropriate?

Universal solutions are the preferred solutions to materialize when exchanging relational databases (Fagin et al. 2005; Fagin, Kolaitis, and Popa 2005; Barceló 2009), also in the case of relational databases with incomplete information (Arenas, Pérez, and Reutter 2011). However, universal solutions were not thought to take into consideration source data including implicit knowledge (in the form of TBoxes), so it is natural to ask whether they are appropriate for transferring this type of knowledge. In this section, we provide evidence that universal solutions, as defined in Section 3 and (Arenas, Pérez, and Reutter 2011), might not be appropriate in this scenario because of their limited capacity to represent implicit knowledge, and the high cost of computing them. We start with a motivating example exhibiting these limitations.

Example 2. *Let $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$ and $\mathcal{K}_1 = \langle \mathcal{T}_1, \mathcal{A}_1\rangle$ be as in Example 1. Furthermore, assume that $\mathcal{K}_2' = \langle \mathcal{T}_2', \mathcal{A}_2'\rangle$, where $\mathcal{T}_2' = \{B_2 \sqsubseteq A_2\}$ and $\mathcal{A}_2' = \{B_2(a)\}$. Then we have that $\mathcal{K}_2'$ is a solution for $\mathcal{K}_1$ under $\mathcal{M}$. However, we also have that $\mathcal{K}_2'$ is not a universal solution for $\mathcal{K}_1$ under $\mathcal{M}$. In fact, if $\mathcal{I}_1$ is an interpretation of Σ_1 such that $\Delta^{\mathcal{I}_1} = \{a\}$, ${A_1}^{\mathcal{I}_1} = \{a\}$ and ${B_1}^{\mathcal{I}_1} = \{a\}$, and $\mathcal{I}_2$ is an interpretation of Σ_2 such that $\Delta^{\mathcal{I}_2} = \{a, b\}$, ${A_2}^{\mathcal{I}_2} = \{a\}$ and ${B_2}^{\mathcal{I}_2} = \{a, b\}$, then we have that $\mathcal{I}_1$ is a model of $\mathcal{K}_1$ and $(\mathcal{I}_1, \mathcal{I}_2) \models \mathcal{T}_{12}$ and, therefore, $\mathcal{I}_2 \in \mathrm{SAT}_{\mathcal{M}}(\mathrm{MOD}(\mathcal{K}_1))$. Thus, we conclude that $\mathrm{SAT}_{\mathcal{M}}(\mathrm{MOD}(\mathcal{K}_1)) \neq \mathrm{MOD}(\mathcal{K}_2')$ as $\mathcal{I}_2$ is not a model of $\mathcal{K}_2'$ since it does not satisfy inclusion $B_2 \sqsubseteq A_2$.* ■

In Examples 1 and 2, a case is shown where universal solutions are not appropriate to represent the implicit source knowledge, as we are only able to construct a universal solution with an empty TBox. In the following proposition, we prove that this is not an isolated phenomenon. In this proposition, we say that a TBox $\mathcal{T}$ over a signature Σ is trivial if for every interpretation $\mathcal{I}$ of Σ, it holds that $\mathcal{I} \models \mathcal{T}$ (or, in other words, if $\mathcal{T}$ is equivalent to the empty set of formulas).

Proposition 4.1. *Let $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$ be a DL-Lite$_{\mathcal{R}}$-mapping, $\mathcal{K}_1 = \langle \mathcal{T}_1, \mathcal{A}_1\rangle$ a DL-Lite$_{\mathcal{R}}$ KB over Σ_1, and $\mathcal{K}_2 = \langle \mathcal{T}_2, \mathcal{A}_2\rangle$ a DL-Lite$_{\mathcal{R}}$ KB over Σ_2. If $\langle \mathcal{T}_1 \cup \mathcal{T}_{12}, \mathcal{A}_1\rangle$ is consistent and $\mathcal{K}_2$ is a universal solution for $\mathcal{K}_1$ under $\mathcal{M}$, then $\mathcal{T}_2$ is a trivial TBox.*

This proposition shows that universal solutions are not appropriate to transfer implicit knowledge if *DL-Lite*$_{\mathcal{R}}$ KBs and mappings are considered, as the TBox in the generated universal solutions is trivial.

We now turn to the problem of computing universal solutions in the context of knowledge exchange, that is, we consider the problem of computing, given an $\mathcal{L}$-mapping $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$ and an $\mathcal{L}$ KB $\mathcal{K}_1$ over Σ_1, an $\mathcal{L}$ KB $\mathcal{K}_2$ over Σ_2 such that $\mathcal{K}_2$ is a universal solution for $\mathcal{K}_1$ under $\mathcal{M}$. Our main goal here is to show that this problem cannot be solved efficiently for *DL-Lite*$_{\mathcal{R}}$ KBs and mappings, essentially because no implicit knowledge can be used in universal solutions (see Proposition 4.1).

Universal Solutions in *DL-Lite*$_{RDFS}$. The chase has been shown to be a powerful tool to compute universal solutions in data exchange and knowledge exchange (Fagin et al. 2005; Arenas, Pérez, and Reutter 2011). For the case of *DL-Lite*$_{RDFS}$, the chase can also be used to compute universal solutions. More specifically, given a *DL-Lite*$_{RDFS}$-mapping $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$ and a *DL-Lite*$_{RDFS}$ KB $\mathcal{K}_1 = \langle \mathcal{T}_1, \mathcal{A}_1\rangle$, we have that $\mathit{chase}_{\mathcal{T}_1}(\mathcal{A}_1)$ and $\mathit{chase}_{\mathcal{T}_{12},\Sigma_2}(\mathit{chase}_{\mathcal{T}_1}(\mathcal{A}_1))$ are finite sets of assertions, and, thus, they can be considered as ABoxes[1]. As a corollary of Theorem 8.11 in (Arenas, Pérez, and Reutter 2011) we obtain the following result.

Proposition 4.2. *Let $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$ be a DL-Lite$_{RDFS}$-mapping and $\mathcal{K}_1 = \langle \mathcal{T}_1, \mathcal{A}_1\rangle$ a DL-Lite$_{RDFS}$ KB over Σ_1. Then $\langle \emptyset, \mathit{chase}_{\mathcal{T}_{12},\Sigma_2}(\mathit{chase}_{\mathcal{T}_1}(\mathcal{A}_1))\rangle$ is a universal solution for $\mathcal{K}_1$ under $\mathcal{M}$.*

Thus, given that the chase can be computed in polynomial time for *DL-Lite*$_{RDFS}$ (Calvanese et al. 2007), we obtain as a corollary of Proposition 4.2 that the problem of computing universal solutions for *DL-Lite*$_{RDFS}$-mappings can be solved in polynomial time. Moreover, we also obtain as a corollary of Proposition 4.2 that every *DL-Lite*$_{RDFS}$ KB has a polynomial-size universal solution under a *DL-Lite*$_{RDFS}$-mapping, which is a desirable condition in practice.

Universal Solutions in *DL-Lite*$_{\mathcal{R}}$. Unfortunately, allowing for existentials on the right-hand side of concept inclusions ruins the nice computational properties holding for *DL-Lite*$_{RDFS}$-mappings. In fact, for *DL-Lite*$_{\mathcal{R}}$, infinite sets of assertions can be generated by the chase, which leads to cases of mappings having source KBs without universal solutions (recall that a solution is a KB, which by definition must be finite). This is shown by the following example.

Example 3. *Let $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$, where $\Sigma_1 = \{A(\cdot), P(\cdot,\cdot)\}$, $\Sigma_2 = \{T(\cdot,\cdot)\}$, and $\mathcal{T}_{12} = \{P \sqsubseteq T\}$. Furthermore, assume that $\mathcal{K}_1 = \langle \mathcal{T}_1, \mathcal{A}_1\rangle$, where $\mathcal{A}_1 = \{A(a)\}$ and $\mathcal{T}_1 = \{A \sqsubseteq \exists P, \exists P^- \sqsubseteq \exists P\}$. In this case, we have that $\mathit{chase}_{\mathcal{T}_1}(\mathcal{A}_1)$ contains an infinite path of the form $P(a, n_1), P(n_1, n_2), P(n_2, n_3), \ldots$, where $n_1, n_2, \ldots$ is an infinite sequence of pairwise distinct existentially implied*

[1] Recall that ABoxes are assumed to be finite, while interpretations can be infinite.

objects. Thus, $chase_{\mathcal{T}_{12},\Sigma_2}(chase_{\mathcal{T}_1}(\mathcal{A}_1))$ is the infinite path $T(a, n_1), T(n_1, n_2), T(n_2, n_3), \ldots$. In this case, it follows that if $\mathcal{K}_2$ is a DL-Lite$_\mathcal{R}$ KB over Σ_2, then $\mathcal{K}_2$ cannot be a universal solution for $\mathcal{K}_1$ under $\mathcal{M}$. ■

At this point it is natural to ask how expensive it is to compute universal solutions. We show that universal solutions can be of exponential size for the case of *DL-Lite*$_\mathcal{R}$-mappings, thus indicating that it can be difficult to deal with them in practice. In this proposition, $|\mathcal{M}|$ and $|\mathcal{K}|$ are used to denote the sizes of a mapping $\mathcal{M}$ and a KB $\mathcal{K}$, respectively.

Proposition 4.3. *There exists a family of DL-Lite$_\mathcal{R}$-mappings $\{\mathcal{M}_n = (\Sigma_1^n, \Sigma_2^n, \mathcal{T}_{12}^n)\}_{n\geq 1}$ and a family of DL-Lite$_\mathcal{R}$ KBs $\{\mathcal{K}_n\}_{n\geq 1}$ such that every $\mathcal{K}_n$ is defined over Σ_1^n ($n \geq 1$), and the smallest universal solution for $\mathcal{K}_n$ under $\mathcal{M}_n$ is of size $2^{\Omega(|\mathcal{M}_n|+|\mathcal{K}_n|)}$.*

The exact complexity of computing universal solutions for *DL-Lite*$_\mathcal{R}$ mappings and KBs remains open.

5 Transferring Implicit Knowledge: Query Languages to the Rescue

In Section 4, we have provided strong evidence that universal solutions are not appropriate in the context of knowledge exchange, mainly because of their poor capacity to represent implicit knowledge. We show now that this limitation can be overcome by simply parametrizing the notion of universal solution by a query language. In fact, we show that the notion of representability defined in Section 3 can be easily adapted to the new setting based on a query language, thus providing a natural and useful notion of solution in the context of knowledge exchange.

Let $\mathcal{M}$ be the mapping and $\mathcal{K}_1$, $\mathcal{K}_2$, $\mathcal{K}_2'$ the KBs shown in Examples 1 and 2. In these examples, $\mathcal{K}_2'$ is not a universal solution for $\mathcal{K}_1$ under $\mathcal{M}$ since inclusion $B_2 \sqsubseteq A_2$ cannot be deduced from the information in $\mathcal{K}_1$ and $\mathcal{M}$. Or, more formally, $\mathcal{K}_2'$ is not a universal solution as $B_2 \sqsubseteq A_2$ is not implied by $\langle \mathcal{T}_1 \cup \mathcal{T}_{12}, \mathcal{A}_1 \rangle$. However, $\mathcal{K}_2'$ can also be considered as a solution of $\mathcal{K}_1$ that is desirable to materialize, as the implicit knowledge in $\mathcal{K}_2'$ (i.e., TBox $\mathcal{T}_2'$) represents the implicit knowledge in $\mathcal{K}_1$ (i.e., TBox $\mathcal{T}_1$), given the way that concepts A_1 and B_1 have to be translated according to mapping $\mathcal{M}$. In fact, if one focuses on a particular query language to compare the information in these two solutions, as it has been done to solve some fundamental problems in data exchange (Madhavan and Halevy 2003; Fagin et al. 2008; Arenas et al. 2009), then one discovers that $\mathcal{K}_2'$ is as good as $\mathcal{K}_2$ but with the advantage that $\mathcal{K}_2'$ represents knowledge in a more compact way. In what follows, we introduce a new class of good solutions that captures this intuition.

Definition 4. *Let $\mathcal{Q}$ be a class of queries, $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$ a mapping, $\mathcal{K}_1 = \langle \mathcal{T}_1, \mathcal{A}_1 \rangle$ a KB over Σ_1, and $\mathcal{K}_2$ a KB over Σ_2. Then $\mathcal{K}_2$ is said to be a* $\mathcal{Q}$-solution *for $\mathcal{K}_1$ under $\mathcal{M}$ if for every query $q \in \mathcal{Q}$ over Σ_2, $cert(q, \langle \mathcal{T}_1 \cup \mathcal{T}_{12}, \mathcal{A}_1 \rangle) \subseteq cert(q, \mathcal{K}_2)$. Moreover, $\mathcal{K}_2$ is said to be a* universal $\mathcal{Q}$-solution *for $\mathcal{K}_1$ under $\mathcal{M}$ if for every query $q \in \mathcal{Q}$ over Σ_2, $cert(q, \langle \mathcal{T}_1 \cup \mathcal{T}_{12}, \mathcal{A}_1 \rangle) = cert(q, \mathcal{K}_2)$.*

Notably, for the widely used class UCQ of unions conjunctive queries, we have in Examples 1 and 2 that both $\mathcal{K}_2$ and $\mathcal{K}_2'$ are universal UCQ-solutions for $\mathcal{K}_1$ under $\mathcal{M}$.

The main goal when introducing universal $\mathcal{Q}$-solutions is to propose a natural notion of solution that overcomes the limitations of universal solutions reported in Section 4. The relationship between the different notions of solution that we introduced is established in the following proposition.

Proposition 5.1. *Let $\mathcal{Q}$ be a class of queries, $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$ a mapping, $\mathcal{K}_1$ a KB over Σ_1, and $\mathcal{K}_2$ a KB over Σ_2. If $\mathcal{K}_2$ is a (universal) solution for $\mathcal{K}_1$ under $\mathcal{M}$, then $\mathcal{K}_2$ is a (universal) $\mathcal{Q}$-solution for $\mathcal{K}_1$ under $\mathcal{M}$.*

We can now adapt the notions of representability and weak representability given in Definition 3 to the parameterized notion of solution introduced above.

Definition 5. *Let $\mathcal{Q}$ be a class of queries, $\mathcal{L}$ a DL, $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$ an $\mathcal{L}$-mapping, and $\mathcal{T}_1$ an $\mathcal{L}$-TBox over Σ_1.*

- *$\mathcal{T}_1$ is* $\mathcal{Q}$-representable under $\mathcal{M}$ *if there exists an $\mathcal{L}$-TBox $\mathcal{T}_2$ over Σ_2, called a* $\mathcal{Q}$-representation *of $\mathcal{T}_1$ under $\mathcal{M}$, such that for every ABox $\mathcal{A}_1$ over Σ_1, if $\langle \mathcal{T}_1 \cup \mathcal{T}_{12}, \mathcal{A}_1 \rangle$ is consistent, then $\langle \mathcal{T}_2, chase_{\mathcal{T}_{12},\Sigma_2}(\mathcal{A}_1) \rangle$ is a universal $\mathcal{Q}$-solution for $\langle \mathcal{T}_1, \mathcal{A}_1 \rangle$ under $\mathcal{M}$.*
- *$\mathcal{T}_1$ is* weakly $\mathcal{Q}$-representable under $\mathcal{M}$ *if there exists a mapping $\mathcal{M}^\star = (\Sigma_1, \Sigma_2, \mathcal{T}_{12}^\star)$ such that $\mathcal{T}_{12} \subseteq \mathcal{T}_{12}^\star$, $\mathcal{T}_1 \cup \mathcal{T}_{12} \models \mathcal{T}_{12}^\star$, and $\mathcal{T}_1$ is $\mathcal{Q}$-representable under $\mathcal{M}^\star$.*

We illustrate these notions in the following example.

Example 4. *Let $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$ and $\mathcal{K}_1 = \langle \mathcal{T}_1, \mathcal{A}_1 \rangle$ be as in Examples 1 and 2. Then we have that $\mathcal{T}_2 = \{B_2 \sqsubseteq A_2\}$ is a UCQ-representation of $\mathcal{T}_1$ under $\mathcal{M}$.*

On the other hand, if $\mathcal{M}' = (\Sigma_1, \Sigma_2, \mathcal{T}_{12}')$ with $\mathcal{T}_{12}' = \{A_1 \sqsubseteq A_2\}$, then $\mathcal{T}_1$ is not UCQ-representable under $\mathcal{M}'$: let $\mathcal{A}_1' = \{B_1(a)\}$, then $chase_{\mathcal{T}_{12}',\Sigma_2}(\mathcal{A}_1') = \emptyset$ and for no TBox $\mathcal{T}_2'$, $\langle \mathcal{T}_2', chase_{\mathcal{T}_{12}',\Sigma_2}(\mathcal{A}_1') \rangle$ is a universal UCQ-solution for $\langle \mathcal{T}_1, \mathcal{A}_1' \rangle$ under $\mathcal{M}'$. However, $\mathcal{T}_{12}^\star = \mathcal{T}_{12}' \cup \{B_1 \sqsubseteq A_2\}$ witnesses that $\mathcal{T}_1$ is weakly UCQ-representable under $\mathcal{M}'$, as $\mathcal{T}_{12}' \subseteq \mathcal{T}_{12}^\star$, $\mathcal{T}_1 \cup \mathcal{T}_{12}' \models \mathcal{T}_{12}^\star$, and $\mathcal{T}_1$ is UCQ-representable under $\mathcal{M}^\star = (\Sigma_1, \Sigma_2, \mathcal{T}_{12}^\star)$ (the empty TBox does the job). ■

6 UCQ-Representability for *DL-Lite*$_{RDFS}$

We present now the results for the UCQ-representability and weak UCQ-representability problems for the case where TBoxes and mappings are expressed in *DL-Lite*$_{RDFS}$.

We start by considering the decision problem associated with UCQ-representability: Given a *DL-Lite*$_{RDFS}$-mapping $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$, a *DL-Lite*$_{RDFS}$-TBox $\mathcal{T}_1$ over Σ_1, and a *DL-Lite*$_{RDFS}$-TBox $\mathcal{T}_2$ over Σ_2, check whether $\mathcal{T}_2$ is a UCQ-representation of $\mathcal{T}_1$ under $\mathcal{M}$, i.e., for each ABox $\mathcal{A}_1$ over Σ_1, $\langle \mathcal{T}_2, chase_{\mathcal{T}_{12},\Sigma_2}(\mathcal{A}_1) \rangle$ is a universal UCQ-solution for $\langle \mathcal{T}_1, \mathcal{A}_1 \rangle$ under $\mathcal{M}$.

For a *DL-Lite*$_{RDFS}$ TBox $\mathcal{T}$ and a concept or role N, we define the *upward closure of N with respect to $\mathcal{T}$* as the set $\mathbb{U}_\mathcal{T}(N) = \{N' \mid N' \text{ is concept or role and } \mathcal{T} \models N \sqsubseteq N'\}$, and the *strict* closure $\mathbb{U}_\mathcal{T}^s(N)$ as $\mathbb{U}_\mathcal{T}(N) \setminus \{N\}$. Then, for a set $\mathbf{N}$ of concepts and roles we define $\mathbb{U}_\mathcal{T}(\mathbf{N}) = \bigcup_{N\in\mathbf{N}} \mathbb{U}_\mathcal{T}(N)$, and its strict version $\mathbb{U}_\mathcal{T}^s(\mathbf{N})$. Notice that both $\mathbb{U}_{\mathcal{T}_{12}}^s(\mathbb{U}_{\mathcal{T}_1}(N))$ and $\mathbb{U}_{\mathcal{T}_2}(\mathbb{U}_\mathcal{M}^s(N))$ are sets over Σ_2, for each concept or role N over Σ_1. We are ready to provide a characterization of UCQ-representations.

Proposition 6.1. *Let $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$ be a DL-Lite$_{RDFS}$-mapping, $\mathcal{T}_1$ a DL-Lite$_{RDFS}$-TBox over Σ_1, and $\mathcal{T}_2$ a DL-Lite$_{RDFS}$-TBox over Σ_2. Then $\mathcal{T}_2$ is a UCQ-representation of $\mathcal{T}_1$ under $\mathcal{M}$ iff $\mathbb{U}^s_{\mathcal{T}_{12}}(\mathbb{U}_{\mathcal{T}_1}(N)) = \mathbb{U}_{\mathcal{T}_2}(\mathbb{U}^s_{\mathcal{T}_{12}}(N))$, for each concept or role N over Σ_1.*

The necessary and sufficient condition in Proposition 6.1 can be checked in polynomial time, as the implication problem for *DL-Lite*$_{RDFS}$ can be solved in polynomial time (Calvanese et al. 2007). This characterization gives us a way to construct an algorithm for checking representability of $\mathcal{T}_1$ under $\mathcal{M}$. The idea of this algorithm is to create the "maximum" representation candidate and then to check whether it is a UCQ-representation. If yes, then $\mathcal{T}_1$ is UCQ-representable under $\mathcal{M}$, otherwise it is not. The candidate $\mathcal{T}_2$ is obtained, by first collecting all inclusions over Σ_2 in order to satisfy the condition $\mathbb{U}^s_{\mathcal{T}_{12}}(\mathbb{U}_{\mathcal{T}_1}(N)) \subseteq \mathbb{U}_{\mathcal{T}_2}(\mathbb{U}^s_{\mathcal{T}_{12}}(N))$, and then removing some of them to satisfy the opposite condition $\mathbb{U}^s_{\mathcal{T}_{12}}(\mathbb{U}_{\mathcal{T}_1}(N)) \supseteq \mathbb{U}_{\mathcal{T}_2}(\mathbb{U}^s_{\mathcal{T}_{12}}(N))$.

Theorem 6.2. *There exists a polynomial time algorithm that, given a DL-Lite$_{RDFS}$-mapping $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$ and a DL-Lite$_{RDFS}$-TBox $\mathcal{T}_1$ over Σ_1, decides whether $\mathcal{T}_1$ is representable under $\mathcal{M}$.*

Interestingly, when source TBoxes are expressed in *DL-Lite*$_{RDFS}$, they are always weakly representable by enriching mappings as follows. Given a *DL-Lite*$_{RDFS}$-mapping $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$ and a *DL-Lite*$_{RDFS}$-TBox $\mathcal{T}_1$ over Σ_1, define the enriched mapping $\mathcal{M}^\star = (\Sigma_1, \Sigma_2, \mathcal{T}^\star_{12})$, where $\mathcal{T}^\star_{12} = \{N_1 \sqsubseteq N_2 \mid N_1 \text{ is over } \Sigma_1, N_2 \text{ is over } \Sigma_2, \text{ and } \mathcal{T}_1 \cup \mathcal{T}_{12} \models N_1 \sqsubseteq N_2\}$. It follows that $\mathcal{T}_{12}$ is contained in $\mathcal{T}^\star_{12}$ and $\mathcal{T}_1 \cup \mathcal{T}_{12} \models \mathcal{T}^\star_{12}$.

Theorem 6.3. *Let $\mathcal{M} = (\Sigma_1, \Sigma_2, \mathcal{T}_{12})$ be a DL-Lite$_{RDFS}$-mapping and $\mathcal{T}_1$ a DL-Lite$_{RDFS}$ TBox over Σ_1. Then $\mathcal{T}_1$ is UCQ-representable under $\mathcal{M}^\star$ and, thus, $\mathcal{T}_1$ is weakly UCQ-representable under $\mathcal{M}$.*

7 Conclusions

In this paper, we have specialized the framework of KB exchange proposed in (Arenas, Pérez, and Reutter 2011) to the case of DLs, and introduced the novel problems of representability and representability with respect to a query language. We have studied KB exchange for *DL-Lite*$_{\mathcal{R}}$ and developed techniques for UCQ-representability for *DL-Lite*$_{RDFS}$. We are currently working on extending our results to *DL-Lite*$_{\mathcal{R}}$, addressing also the problem of deciding the existence of universal solutions in this case. Interesting directions for future work are to study KB exchange and representability for other DLs, e.g., those of the $\mathcal{EL}$ family and (very) expressive DLs, and to address the problems studied in data exchange, such as composition and inversion of mappings, in the setting of KB exchange. The precise connection to conservative extensions remains also to be explored.

Acknowledgements

The authors were supported by the EU Marie Curie action FP7-PEOPLE-2009-IRSES (grant 24761), by Fondecyt (grant 1090565), by NWO (project 612.001.012 (DEX)), and by the EU ICT projects ACSI (grant FP7-257593), OntoRule (grant FP7-231875), and FOX (grant FP7-233599).

References

Arenas, M.; Pérez, J.; Reutter, J. L.; and Riveros, C. 2009. Composition and inversion of schema mappings. *SIGMOD Record* 38(3):17–28.

Arenas, M.; Pérez, J.; and Reutter, J. L. 2011. Data exchange beyond complete data. In *Proc. of the 30th ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems (PODS 2011)*, 83–94.

Barceló, P. 2009. Logical foundations of relational data exchange. *SIGMOD Record* 38(1):49–58.

Brickley, D., and Guha, R. V. 2004. RDF vocabulary description language 1.0: RDF Schema. W3C Recommendation, World Wide Web Consortium. Available at http://www.w3.org/TR/rdf-schema/.

Calvanese, D.; De Giacomo, G.; Lembo, D.; Lenzerini, M.; and Rosati, R. 2007. Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family. *J. of Automated Reasoning* 39(3):385–429.

Fagin, R.; Kolaitis, P. G.; Miller, R. J.; and Popa, L. 2005. Data exchange: Semantics and query answering. *Theoretical Computer Science* 336(1):89–124.

Fagin, R.; Kolaitis, P. G.; Nash, A.; and Popa, L. 2008. Towards a theory of schema-mapping optimization. In *Proc. of the 27th ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems (PODS 2008)*, 33–42.

Fagin, R.; Kolaitis, P. G.; and Popa, L. 2005. Data exchange: Getting to the core. *ACM Trans. on Database Systems* 30(1):174–210.

Libkin, L., and Sirangelo, C. 2011. Data exchange and schema mappings in open and closed worlds. *J. of Computer and System Sciences* 77(3):542–571.

Madhavan, J., and Halevy, A. Y. 2003. Composing mappings among data sources. In *Proc. of the 29th Int. Conf. on Very Large Data Bases (VLDB 2003)*, 572–583.