Query Rewriting Beyond DL-Lite

Carsten Lutz

Fachbereich Informatik, Universität Bremen, Germany

1 Abstract of Invited Talk

Query rewriting has become a very prominent tool for efficiently implementing ontology-mediated querying in practice. The technique was originally introduced in the context of DL-Lite [4], but is now increasingly being used also for more expressive DLs. While rewritings are not guaranteed to exist beyond DL-Lite, the simple structure of ontologies that emerge from practical applications gives hope that non-existence of rewritings is a rare case.

The aim of the talk is to survey FO- and Datalog-rewriting of ontology-mediated queries in description logics beyond DL-Lite. It is structured into three parts. The first part is concerned with FO-rewritings in Horn-DLs such as $\mathcal{EL}$, $\mathcal{ELI}$, and Horn-$\mathcal{SHI}$, the second part considers FO-rewritings in non-Horn-DLs such as $\mathcal{ALC}$ and $\mathcal{ALCI}$, and the third part is about Datalog-rewritings in non-Horn DLs. In all three parts, I will try to emphasize useful characterizations of FO-rewritability, practically efficient algorithms for constructing rewritings, and relevant computational complexity results.

The presentation is based on joint work with Meghyn Bienvenu, Balder ten Cate, Peter Hansen, İnanc Seylan, and Frank Wolter. The subsequent section provides some supplementary material that is featured in the talk, but has not yet been published elsewhere. It establishes a link between the first and the second part of the talk.

2 Supplementary Material

In [3, 7], we have proposed an approach to deciding the FO-rewritability of OMQs for the case where the ontology/TBox is formulated in a Horn DL such as $\mathcal{EL}$, $\mathcal{ELI}$, and Horn-$\mathcal{ALCI}$. The approach has led to efficient (yet complete) practical implementations, and it relies on a characterization of FO-rewritability in terms of tree-shaped ABoxes. Intuitively, the characterization relies on a property of TBoxes that is called 'unraveling tolerance' and which typically is enjoyed by Horn DLs, but not by DLs that include forms of disjunction. In contrast, the only known complete approach to deciding FO-rewritability of OMQs in which the TBox is formulated in full (non-Horn) $\mathcal{ALC}$ and $\mathcal{ALCI}$ is via the CSP connection in [9, 2]. Since $\mathcal{ALC}$- and $\mathcal{ALCI}$-TBoxes are typically not unraveling tolerant, it might seems that these two world are largely unrelated. In the following, though, we point out a characterization of FO-rewritability in full $\mathcal{ALCI}$ that establishes an interesting connection to tree-shaped ABoxes and thus to the Horn case. We

consider *Boolean atomic queries (BAQs)*, that is, queries of the form $\exists x\, A(x)$ with A a concept name.

An *ontology-mediated query (OMQ)* is a triple $Q = (\mathcal{T}, \Sigma, q)$ with $\mathcal{T}$ a TBox, Σ an ABox signature (set of concept and role names), and q a query. An *OBDA language* is a set of OMQs. We use $(\mathcal{ALCI}, \text{BAQ})$ to denote the OBDA language that consists of all OMQs $(\mathcal{T}, \Sigma, q)$ with $\mathcal{T}$ an $\mathcal{ALCI}$-TBox and q a BAQ, and likewise for other combinations of a DL and a query language. An OMQ $Q = (\mathcal{T}, \Sigma, q)$ is *FO-rewritable* if there is an FO-sentence φ such that for every Σ-ABox $\mathcal{A}$ that is consistent w.r.t. $\mathcal{T}$, we have $\mathcal{A} \models Q$ iff $\mathcal{A} \models \varphi$.

As usual in OBDA, an *ABox* is a finite set of assertions of the form $A(a)$ or $r(a, b)$ with A a concept name and r a role name. We write $r^-(a, b) \in \mathcal{A}$ to mean $r(b, a) \in \mathcal{A}$ and use $\mathsf{Ind}(\mathcal{A})$ to denote the set of individuals used in $\mathcal{A}$. An ABox $\mathcal{A}$ is *tree-shaped* if the undirected graph $(\mathsf{Ind}(\mathcal{A}), \{\{a, b\} \mid r(a, b) \in \mathcal{A}\})$ is a tree and whenever $r(a, b) \in \mathcal{A}$, then (i) $s(a, b) \in \mathcal{A}$ implies $r = s$ and (ii) $\mathcal{A}$ contains no assertion of the form $s(b, a)$. Tree-shapedness of conjunctive queries (CQs) is defined accordingly. Note that, in both cases, our trees allow upwards- and downwards-directed edges, but no multi-edges.

We now introduce unravelings of ABoxes and the notion of unraveling tolerance [9]. Let $\mathcal{A}$ be an ABox and $a \in \mathsf{Ind}(\mathcal{A})$. The *unraveling* $\mathcal{A}^u_a$ *of* $\mathcal{A}$ *at* a is the following (possibly infinite) ABox:

- $\mathsf{Ind}(\mathcal{A}^u_a)$ is the set of sequences $b_0 r_0 b_1 \cdots r_{n-1} b_n$, $n \geq 0$, such that $b_0 = a$, $b_0, \ldots, b_n \in \mathsf{Ind}(\mathcal{A})$ and $r_0, \ldots, r_{n-1}$ are (potentially inverse) roles;
- for each $C(b) \in \mathcal{A}$ and $\alpha = b_0 \cdots b_n \in \mathsf{Ind}(\mathcal{A}^u_a)$ with $b_n = b$: $C(\alpha) \in \mathcal{A}^u_a$;
- for each $\alpha = b_0 r_0 \cdots r_{n-1} b_n \in \mathsf{Ind}(\mathcal{A}^u_a)$ with $n > 0$: $r_{n-1}(b_0 \cdots b_{n-1}, \alpha) \in \mathcal{A}^u_a$.

For all $\alpha = b_0 \cdots b_n \in \mathsf{Ind}(\mathcal{A}^u_a)$, we write $\mathsf{tail}(\alpha)$ to denote b_n. Note that $\mathcal{A}^u_a$ is tree-shaped. An OMQ $Q = (\mathcal{T}, \Sigma, q)$ is *unraveling tolerant* if for every Σ-ABox $\mathcal{A}$, $\mathcal{A} \models Q$ implies $\mathcal{A}^u_a \models Q$ for some $a \in \mathsf{Ind}(\mathcal{A})$. Note that this is essentially the same notion of unraveling tolerance as introduced in [9].

It can be shown as in [9] that in OBDA languages where the TBoxes are formulated in Horn DLs such as $\mathcal{EL}$, $\mathcal{ELI}$, and Horn-$\mathcal{ALC}$ and where queries are BAQs or *atomic queries* (*AQs*, queries of the form $A(x)$ with A a concept name), all OMQs are unraveling tolerant. This underlies the following characterization from [3].

Theorem 1 ([3]). *A BAQ $Q = (\mathcal{T}, \Sigma, q)$ from (Horn-$\mathcal{ALCI}$, AQ) is FO-rewritable iff there exists a $k \geq 0$ such that for all tree-shaped Σ-ABoxes $\mathcal{A}$ which are consistent with $\mathcal{T}$, $\mathcal{A} \models Q$ implies $\mathcal{A}|_k \models Q$ where $\mathcal{A}_k$ is $\mathcal{A}$ with all nodes on level exceeding k removed.*

Using a pumping argument, it can be shown that if there is any bound k as Theorem 1, then we can choose $k = 2^{2|\mathcal{T}|}$. Based on this, worst-case optimal (ExpTime) decision procedures for FO-rewritability in (Horn-$\mathcal{ALCI}$, AQ) can be devised using automata methods. Efficiently computing rewritings in practice requires further algorithm engineering [7].

We will now establish a characterization of FO-rewritability in the non-Horn OBDA language $(\mathcal{ALCI}, \text{BAQ})$. It is shown in [2] that for every OMQ $Q = (\mathcal{T}, \Sigma, q)$ from $(\mathcal{ALCI}, \text{BAQ})$, there is a CSP template (a finite relational structure) T_Q over signature Σ such that for all Σ-ABoxes $\mathcal{A}$, we have $\mathcal{A}, \mathcal{T} \models q$ iff $\mathcal{A} \not\to T_Q$, that is, iff there is no homomorphism from $\mathcal{A}$ to T_Q (in the standard sense of labeled directed graphs). We say that a CSP template T is *FO-definable* if there is an FO-sentence φ such that for all finite Σ-structures S, we have $S \to T$ iff $S \models \varphi$. The complement of T is *definable in monadic Datalog* if there is a monadic Datalog program Π such that for all finite Σ-structures S, we have $S \not\to T$ iff $S \models \Pi$. Note that a CSP template is FO-definable iff its complement is (just take the negation of the defining sentence), but this is not true for monadic Datalog definability.

It is easy to see that an OMQ Q is FO-rewritable if and only if the complement of T_Q is FO-definable, and likewise for rewritability into monadic Datalog. In [2], this observation is used together with results on the FO-definability of CSPs [8] to show the following.

Theorem 2 ([2]). *FO-rewritability in $(\mathcal{ALCI}, BAQ)$ and $(\mathcal{ALCI}, AQ)$ is decidable and* NExpTime*-complete.*

This approach is also capable of producing actual rewritings, but unfortunately it is best-case exponential. This calls for a better understanding of FO-rewritability in $(\mathcal{ALCI}, \text{BAQ})$ and related languages, as a basis for more practical (yet complete) approaches.

As a preliminary, we show that unraveling tolerance is equivalent to rewritability into monadic Datalog. This actually follows straightforwardly from known results about CSPs.

Theorem 3. *An OMQ from $(\mathcal{ALCI}, BAQ)$ is unraveling tolerant iff it is rewritable into monadic Datalog.*

Proof. A CSP template T over signature Σ has *tree duality* iff there is a set $\mathcal{O}$ of tree-shaped Σ-structures (called *obstructions* and where tree-shapedness is defined as for ABoxes and CQs above) such that for all finite Σ-structures S, we have $T \leftarrow S$ iff $S \not\leftarrow O$ for all $O \in \mathcal{O}$. It was shown in [5] that T has tree duality iff the complement of T is definable in monadic Datalog. It thus remains to show that an OMQ from $(\mathcal{ALCI}, \text{BAQ})$ is unraveling tolerant iff T_Q has tree duality.

"if". Assume that $Q = (\mathcal{T}, \Sigma, q)$ is unraveling tolerant. Let $\mathcal{O}$ be the set of all tree-shaped Σ-ABoxes $\mathcal{A}$ with $\mathcal{A} \models Q$. Then $\mathcal{O}$ witnesses tree duality: if $T_Q \leftarrow \mathcal{A}$ for some Σ-ABox $\mathcal{A}$, then $\mathcal{A} \not\models Q$; since $\mathcal{B} \models Q$ and $\mathcal{B} \to \mathcal{A}$ implies $\mathcal{A} \models Q$ [2], we thus have $\mathcal{A} \not\leftarrow \mathcal{B}$ for all $\mathcal{B} \in \mathcal{O}$ as required. Conversely, assume that $\mathcal{A}$ is a Σ-ABox with $\mathcal{A} \not\leftarrow \mathcal{B}$ for all $\mathcal{B} \in \mathcal{O}$. Clearly, $\mathcal{A}^u_a \to \mathcal{A}$ for all $a \in \mathsf{Ind}(\mathcal{A})$. Thus, no such $\mathcal{A}^u_a$ is in $\mathcal{O}$, implying that $\mathcal{A}^u_a \not\models Q$. Since Q is unraveling tolerant, $\mathcal{A} \not\models Q$ which implies $T_Q \leftarrow \mathcal{A}$ as required.

"only if". Assume that T_Q has tree duality with set of obstructions $\mathcal{O}$. Let $\mathcal{A}$ be a Σ-ABox with $\mathcal{A} \models Q$. Then $T_Q \not\leftarrow \mathcal{A}$ and thus $\mathcal{A} \leftarrow \mathcal{B}$ for some $\mathcal{B} \in \mathcal{O}$. Since

$\mathcal{B}$ is tree-shaped, $\mathcal{A} \leftarrow \mathcal{B}$ implies $\mathcal{A}^u_a \leftarrow \mathcal{B}$ for some $a \in \mathsf{Ind}(\mathcal{A})$. Consequently $T_Q \not\leftarrow \mathcal{A}^u_a$ which yields $\mathcal{A}^u_a \models Q$ as required. □

We now establish the announced characterization.

Theorem 4. *Let $Q = (\mathcal{T}, q, \Sigma)$ be an OMQ from $(\mathcal{ALCI}, BAQ)$. Then Q is FO-rewritable iff*

1. *Q is FO-rewritable on tree-shaped Σ-ABoxes and*
2. *Q is unraveling tolerant.*

Proof. "if". Assume that Q is unraveling tolerant and FO-rewritable on tree-shaped ABoxes. By Theorem 3, the complement of the template T_Q is definable by a monadic Datalog program Π_Q. Let Π'_Q be obtained from Π_Q by identifying the variables in rule bodies in all possible ways and then retaining only those rules whose bodies are a tree-shaped CQ. It can be verified that Π'_Q is a rewriting of Q: $\mathcal{A} \models Q$ implies $\mathcal{A}^u_a \models Q$ for some $a \in \mathsf{Ind}(\mathcal{A})$ (since Q is unraveling tolerant) implies $\mathcal{A}^u_a \models \Pi_Q$ (since Π_Q is a rewriting of Q) implies $\mathcal{A}^u_a \models \Pi'_Q$ (since $\mathcal{A}^u_a$ is tree-shaped) implies $\mathcal{A} \models \Pi'_Q$ (since $\mathcal{A}^u_A \to \mathcal{A}$). Conversely, $\mathcal{A} \models \Pi'_Q$ implies $\mathcal{A} \models \Pi_Q$ (by construction of Π'_Q) implies $\mathcal{A} \models Q$. It is easy to further modify Π'_Q so that in addition to being tree shaped, every role body contains at most one EDB atom.

We now use the existence of Π'_Q to argue that Q has an FO-rewriting φ on tree-shaped ABoxes that takes the form of a union of tree-shaped CQs. Let ψ be an FO-rewriting of Q on tree-shaped ABoxes. By Gaifman's locality theorem, there is a number $d \geq 0$ such that for every Σ-ABox $\mathcal{A}$, we have $\mathcal{A} \models \psi$ iff $\mathcal{A}^*_d \models \psi$ where $\mathcal{A}^*_d$ is obtained by taking the disjoint union of all d-neighborhoods in $\mathcal{A}$; here, the *d-neighborhood* in $\mathcal{A}$ around $a \in \mathsf{Ind}(\mathcal{A})$ is the restriction of $\mathcal{A}$ to all individuals that can be reached from a on a role path in $\mathcal{A}$ of length at most d. Note that ψ is a rewriting of Q and every OMQ from $(\mathcal{ALCI}, \mathrm{BAQ})$ satisfies the property that if a Σ-ABox $\mathcal{A}$ is the disjoint union of ABoxes $\mathcal{A}_1, \ldots, \mathcal{A}_k$, then $\mathcal{A} \models Q$ iff $\mathcal{A}_i \models Q$ for at least one $\mathcal{A}_i$. We can thus strengthen the above obervation as follows: for every Σ-ABox $\mathcal{A}$, we have $\mathcal{A} \models \psi$ iff there is some d-neighborhood $\mathcal{N}$ in $\mathcal{A}$ such that $\mathcal{N} \models \psi$. Since both ψ and Π'_Q are rewritings of the same query Q, the same applies to the monadic Datalog program Π'_Q instead of to ψ. Moreover, we can find an $\ell \geq 0$ such that for every Σ-ABox $\mathcal{A}$ with $\mathcal{A} \models \Pi'_Q$, there is an $\mathcal{A}' \subseteq \mathcal{A}$ with $\mathcal{A}' \models \Pi'_Q$ and in which every individual has degree at most ℓ—due to the special shape of Π'_Q, we can in fact simply choose for ℓ the number of IDB relations in Π'_Q. Combining these two observations, we get the following: for every tree-shaped Σ-ABox $\mathcal{A}$ with $\mathcal{A} \models Q$, there is a tree-shaped ABox $\mathcal{A}' \subseteq \mathcal{A}$ of depth at most d and degree at most ℓ such that $\mathcal{A}' \models Q$. We can thus choose as the desired rewriting φ the UCQ that consists of all tree-shaped ABoxes $\mathcal{A}$ (viewed as a CQ) that satisfy $\mathcal{A} \models Q$ and are of depth at most d and of degree at most ℓ.

It remains to note that, due to its syntactic shape, φ is an FO-rewriting not only on tree-shaped ABoxes, but also on unrestricted ones. First assume that $\mathcal{A}$ is a Σ-ABox with $\mathcal{A} \models Q$. Since Q is unraveling tolerant, there then is an

$a \in \mathsf{Ind}(\mathcal{A})$ with $\mathcal{A}_a^u \models Q$. Since φ is an FO-rewriting on tree-shaped ABoxes, we get $\mathcal{A}_a^u \models \varphi$. Since φ is a UCQ and $\mathcal{A}_a^u \to \mathcal{A}$, we obtain $\mathcal{A} \models \varphi$. Conversely, assume $\mathcal{A} \models \varphi$. Since φ is a union of tree-shaped CQs, this yields $\mathcal{A}_a^u \models \varphi$ for some $a \in \mathsf{Ind}(\mathcal{A})$, thus $\mathcal{A}_a^u \models Q$ and $\mathcal{A} \models Q$.

"only if". Assume that Q is FO-rewritable. Then it is clearly also FO-rewritable on tree-shaped ABoxes (the same rewriting works). It thus remains to show that Q is unraveling tolerant.

It is proved in [1] that a CSP template T over signature Σ is FO-rewritable iff it has *finite duality*, that is, iff there is a finite set of structures $\mathcal{O}$ such that for all finite Σ-structures S, we have $T \leftarrow S$ iff $S \not\leftarrow O$ for all $O \in \mathcal{O}$. It was shown in [10] that finite duality implies tree duality. In fact, as observed in [8], we can assume w.l.o.g. that the finitely many elements of $\mathcal{O}$ are finite and tree-shaped. One could call this *finite duality in terms of finite trees*.

Now back to our OMQ Q. Since Q is FO-rewritable, so is T_Q. By the above result on finite duality in terms of finite trees, there is thus a finite set Γ of tree-shaped ABoxes such that for all Σ-ABoxes $\mathcal{A}$, we have $\mathcal{A} \models Q$ iff $\mathcal{B} \to \mathcal{A}$ for some $\mathcal{B} \in \Gamma$. Consequently, the UCQ $\widehat{q} = \bigvee_{\mathcal{B} \in \Gamma} q_{\mathcal{B}}$ is an FO-rewriting of Q, where $q_{\mathcal{B}}$ is $\mathcal{B}$ viewed as a Boolean CQ in the obvious way. Note that $\widehat{q}$ is a disjunction of tree-shaped CQs. It is thus straightforward to show that for all Σ-ABoxes $\mathcal{A}$, we have $\mathcal{A} \models \widehat{q}$ iff $\mathcal{A}_a^u \models \widehat{q}$ for some $a \in \mathsf{Ind}(\mathcal{A})$. The unraveling tolerance of Q follows. □

The proof of Theorem 4 also yields the following corollary, which strengthens the observation from [2] that in $(\mathcal{ALCI}, \mathrm{BAQ})$, every FO-rewritable OMQ is UCQ-rewritable (essentially a consequence of Rossmann's homomorphism preservation theorem).

Corollary 1. *If an OMQ in $(\mathcal{ALCI}, BAQ)$ is FO-rewritable, then it is rewritable into a union of tree-shaped conjunctive queries.*

We remark that, even when switching to the OBDA language $(\mathcal{ALC}, \mathrm{BAQ})$, it is not possible to replace the undirected trees in Corollary 1 with directed trees.

We close with some discussion of Theorem 4. As future work, we plan to adapt the result from $(\mathcal{ALCI}, \mathrm{BAQ})$ to $(\mathcal{ALCI}, \mathrm{AQ})$ and to use them as a basis for developing practically feasible algorithms that construct FO-rewritings. Dealing with $(\mathcal{ALCI}, \mathrm{AQ})$ seems to require more liberal definitions of tree-shaped ABoxes and of unraveling tolerance which allow for back-edges to the root as in the tree-model property for DLs with nominals. To obtain a first impression of the effect of answer variables, the reader might want to consider the following OMQ $Q = (\mathcal{T}, \Sigma, q)$ from $(\mathcal{ALCI}, \mathrm{BAQ})$:

$$\mathcal{T} = \{P \sqcap \exists r.P \sqsubseteq A, \neg P \sqcap \exists r.\neg P \sqsubseteq A\} \qquad \Sigma = \{r\} \qquad q = \exists x\, A(x)$$

and its variation Q' from $(\mathcal{ALCI}, \mathrm{BAQ})$ obtained by replacing q with the AQ $q' = A(x)$. Q is not unraveling tolerant as witnessed by the ABox $\mathcal{A} = \{r(a, a)\}$ which satisfies $\mathcal{A} \models Q$, but $\mathcal{A}_a^u \not\models Q$. The same is true for Q' if the notion of

unraveling tolerance is adapted in a naive way to non-Boolean OMQs. However, while Q is not FO-rewritable (by Theorem 4), it is not too hard to prove that the FO-formula $r(x, x)$ is an FO-rewriting of Q'.

Another interesting question concerns the complexity of deciding FO-rewritability in $(\mathcal{ALCI}, \text{BAQ})$ (and of course also $(\mathcal{ALCI}, \text{AQ})$) via Theorem 4. It is shown in [5] that unraveling tolerance is decidable (in 3-ExpTime) and using techniques from [2], it is possible to prove NExpTime-hardness. We speculate that the problem might actually be NExpTime-complete. Regarding FO-rewritability in $(\mathcal{ALCI}, \text{BAQ})$ on tree-shaped ABoxes, it seems likely that a 2-ExpTime lower bound can be established by combining reductions from [3] and [6]—thus FO-rewritability on tree-shaped ABoxes would be harder than on unrestricted ABoxes! However, if we already know that Q is unraveling tolerant, then FO-rewritability on trees is trivially in NExpTime, simply by Theorem [6].

Acknowledgements. I am grateful to Frank Wolter for, as always, very helpful and stimulating discussions.

References

1. Atserias, A.: On digraph coloring problems and treewidth duality. Eur. J. Comb. 29(4), 796–820 (2008)
2. Bienvenu, M., ten Cate, B., Lutz, C., Wolter, F.: Ontology-based data access: A study through disjunctive datalog, CSP, and MMSNP. ACM Trans. Database Syst. 39(4), 33:1–33:44 (2014)
3. Bienvenu, M., Lutz, C., Wolter, F.: First-order rewritability of atomic queries in Horn description logics. In: Proc. of IJCAI (2013)
4. Calvanese, D., Giacomo, G.D., Lembo, D., Lenzerini, M., Rosati, R.: Tractable reasoning and efficient query answering in description logics: The DL-Lite family. J. Autom. Reasoning 39(3), 385–429 (2007)
5. Feder, T., Vardi, M.Y.: The computational structure of monotone monadic SNP and constraint satisfaction: A study through datalog and group theory. SIAM J. Comput. 28(1), 57–104 (1998)
6. Ghilardi, S., Lutz, C., Wolter, F.: Did I damage my ontology? A case for conservative extensions in description logics. In: Proc. of KR (2006)
7. Hansen, P., Lutz, C., İnanç Seylan, Wolter, F.: Efficient query rewriting in the description logic $\mathcal{EL}$ and beyond. In: Proc. of IJCAI (2015)
8. Larose, B., Loten, C., Tardif, C.: A characterisation of first-order constraint satisfaction problems. Logical Methods in Computer Science 3(4) (2007)
9. Lutz, C., Wolter, F.: Non-uniform data complexity of query answering in description logics. In: Proc. of KR (2012)
10. Nesetril, J., Tardif, C.: Duality theorems for finite structures (characterising gaps and good characterisations). J. Comb. Theory, Ser. B 80(1), 80–97 (2000)