Conjunctive Query Answering in the Description Logic $\mathcal{EL}$ using a Relational Database System

Carsten Lutz
Universität Bremen, Germany
clu@informatik.uni-bremen.de

David Toman
University of Waterloo, Canada
david@uwaterloo.ca

Frank Wolter
University of Liverpool, UK
wolter@liverpool.ac.uk

Abstract

Conjunctive queries (CQ) are fundamental for accessing description logic (DL) knowledge bases. We study CQ answering in (extensions of) the DL $\mathcal{EL}$, which is popular for large-scale ontologies and underlies the designated OWL2-EL profile of OWL2. Our main contribution is a novel approach to CQ answering that enables the use of standard relational database systems as the basis for query execution. We evaluate our approach using the IBM DB2 system, with encouraging results.

1 Introduction

One of the main applications of ontologies in computer science is in data access, where an ontology formalizes conceptual information about data that is stored in one or multiple data sources, and this information is used to derive answers to queries over such sources. This general setup plays a central role in, e.g., ontology-based information integration and peer-to-peer data management. In these and similar applications, Description Logics (DLs) are popular ontology languages and conjunctive queries (CQs) are used as a fundamental querying tool. Hence, efficient and scalable approaches to CQ answering over DL ontologies are of great interest.

Calvanese et al. have argued that, in the short run, true scalability of conjunctive query answering over DL ontologies can only be achieved by making use of standard relational database management systems (RDBMSs) [2007b]. Alas, this is not straightforward as RDBMSs are unaware of TBoxes (the DL mechanism for storing conceptual information) and adopt the closed-world semantics. In contrast, ABoxes (the DL mechanism for storing data) and the associated ontologies employ the open-world semantics. Existing approaches to overcome these differences have serious limitations. For example, the approach of [Calvanese *et al.*, 2007b] applies only to DLs with data complexity of CQ answering in LOGSPACE whereas for many DLs, this problem is complete for PTIME or co-NP. In particular, the above approach cannot be directly used for DLs that admit qualified existential restrictions, which play an important role in many ontologies. This limitation is shared by the rule-based approach presented in [Wu *et al.*, 2008].

In this paper, we present a novel approach to using RDBMSs for CQ answering over DL ontologies that, in particular, accommodates qualified existential restrictions. We apply it to the $\mathcal{EL}$ family of DLs [Baader *et al.*, 2008], whose members are widely used as ontology languages for large-scale bio-medical ontologies such as SNOMED CT and (early versions of) NCI. Our main result shows that CQ answering in $\mathcal{ELH}_{\bot}^{dr}$, the extension of basic $\mathcal{EL}$ with the bottom concept, role inclusions, and domain and range restrictions, can be implemented using an RDBMS. This result is of particular relevance as $\mathcal{ELH}_{\bot}^{dr}$ can be viewed as the core of the designated OWL-EL profile of the upcoming OWL Version 2 ontology language. We evaluate our approach using the IBM DB2 RDBMS and show that it scales to TBoxes with more than 50.000 axioms where answer times typically range from a fraction of a second to a few seconds.

The central idea of our approach is to incorporate the consequences of the TBox $\mathcal{T}$ into the relational instance corresponding to the given ABox $\mathcal{A}$. To capture this formally, we introduce the notion of *combined first-order (FO) rewritability*. A DL enjoys combined FO rewritability if it is possible to effectively rewrite (i) $\mathcal{A}$ and $\mathcal{T}$ into an FO structure (independently of q) and (ii) q and (possibly) $\mathcal{T}$ into an FO query q^* (independently of $\mathcal{A}$) such that query answers are preserved, i.e, the answer to q^* over the FO structure is the same as the answer to q over $\mathcal{A}$ and $\mathcal{T}$. The connection to RDBMSs then relies on the well-known equivalence between FO structures and relational databases, and FO queries and SQL queries. The notion of combined FO rewritability generalizes the notion of *FO reducibility*, where the TBox is incorporated into the query q rather than into the ABox $\mathcal{A}$ while the ABox itself is used as a relational instance without any modification [Calvanese *et al.*, 2007b]. Notable properties of our approach include:

1. It applies to DLs for which data complexity of CQ answering is PTIME-complete, such as $\mathcal{ELH}_{\bot}^{dr}$.
2. For $\mathcal{ELH}_{\bot}^{dr}$, both rewriting steps can be carried out in polynomial time and produce only a polynomial blowup; moreover, the query rewriting only depends on the input CQ and the role inclusions in $\mathcal{T}$ (usually only few), but not on $\mathcal{T}$'s concept inclusions (usually very many).

In contrast and to the best of our knowledge, all existing approaches to (non-combined) FO reducibility generate a

rewritten query of size m^n, with m the number of symbols in the query and the TBox and n the size of the query.

In addition, we analyze the limitations of our approach and show that DLs with data complexity exceeding PTIME cannot enjoy *polynomial* combined FO rewritability, where the expansion of the ABox due to rewriting (but not necessarily of query rewriting) is bounded by a polynomial. We believe that this is a significant result as combined FO rewritability involving an exponential blowup of the data does not seem to be practical. In particular, the result implies that expressive DLs such as $\mathcal{ALC}$ cannot enjoy polynomial FO rewritability. The same holds even for $\mathcal{EL}$ enriched with negated ABox assertions and negated query atoms. For the latter case, we sketch an approach to query answering that involves only a polynomial blowup of the ABox, but is incomplete (in a way precisely characterized by an alternative semantics).

The remainder of this paper is organized as follows. In Section 2, we introduce some preliminaries. Section 3 describes ABox rewriting, which is based on the notion of a canonical model for an $\mathcal{ELH}_\bot^{dr}$knowledge base. Section 4 is concerned with query rewriting and establishes the main (theoretical) result of this paper. In Section 5, we sketch the actual implementation in an RDBMS and give experimental results. Section 6 is concerned with limitations of our approach and with treating negation. We conclude in Section 7. Proofs are deferred to the appendix.

2 Preliminaries

In $\mathcal{ELH}_\bot^{dr}$, *concepts* are built according to the syntax rule

$$C ::= A \mid \top \mid \bot \mid C \sqcap D \mid \exists r.C$$

where, here and in the remaining paper, A ranges over *concept names* taken from a countably infinite set $\mathsf{N_C}$, r ranges over *role names* taken from a countably infinite set $\mathsf{N_R}$, and C, D range over concepts. A *TBox* is a finite set of *concept inclusions* $C \sqsubseteq D$, *role inclusions* $r \sqsubseteq s$, *domain restrictions* $\mathsf{dom}(r) \sqsubseteq C$, and *range restrictions* $\mathsf{ran}(r) \sqsubseteq C$. An ABox is a finite set of *concept assertions* $A(a)$ and *role assertions* $r(a, b)$, where a, b range over a countably infinite set $\mathsf{N_I}$ of *individual names*. A *knowledge base* is a pair $(\mathcal{T}, \mathcal{A})$ with $\mathcal{T}$ a TBox and $\mathcal{A}$ an ABox.

As usual, an *interpretation* is a pair $(\Delta^\mathcal{I}, \cdot^\mathcal{I})$ with $\Delta^\mathcal{I}$ a non-empty *domain* and $\cdot^\mathcal{I}$ an *interpretation function* that maps each concept name A to a subset $A^\mathcal{I} \subseteq \Delta^\mathcal{I}$, each role name r to a binary relation $r^\mathcal{I} \subseteq \Delta^\mathcal{I} \times \Delta^\mathcal{I}$, and each individual name a to an element $a^\mathcal{I} \in \Delta^\mathcal{I}$ The interpretation function is extended to composite concepts by setting

$$\begin{aligned} \top^\mathcal{I} &= \Delta^\mathcal{I} \\ \bot^\mathcal{I} &= \emptyset \\ (C \sqcap D)^\mathcal{I} &= C^\mathcal{I} \cap D^\mathcal{I} \\ (\exists r.C)^\mathcal{I} &= \{d \in \Delta^\mathcal{I} \mid \exists e \in \Delta^\mathcal{I} : (d, e) \in r^\mathcal{I} \wedge e \in C^\mathcal{I}\}. \end{aligned}$$

An interpretation $\mathcal{I}$ *satisfies* a concept inclusion $C \sqsubseteq D$ if $C^\mathcal{I} \subseteq D^\mathcal{I}$, a role inclusion $r \sqsubseteq s$ if $r^\mathcal{I} \subseteq s^\mathcal{I}$, a domain restriction $\mathsf{dom}(r) \sqsubseteq C$ if $(d, e) \in r^\mathcal{I}$ implies $d \in C^\mathcal{I}$, a range restriction $\mathsf{ran}(r) \sqsubseteq C$ if $(d, e) \in r^\mathcal{I}$ implies $e \in C^\mathcal{I}$, a concept assertion $A(a)$ if $a^\mathcal{I} \in A^\mathcal{I}$, and a role assertion $r(a, b)$ if $(a^\mathcal{I}, b^\mathcal{I}) \in r^\mathcal{I}$. $\mathcal{I}$ is a *model* of a TBox $\mathcal{T}$ (ABox $\mathcal{A}$) if it satisfies all inclusions and restrictions in $\mathcal{T}$ (assertions in $\mathcal{A}$). It is a model of a knowledge base $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ if it is a model of $\mathcal{T}$ and $\mathcal{A}$. A knowledge base that has a model is called *consistent*. For a concept inclusion, role inclusions, or assertion α, we write $\mathcal{K} \models \alpha$ if α is satisfied in all models of $\mathcal{K}$. If empty, $\mathcal{A}$ is simply omitted.

Let $\mathsf{N_V}$ be a countably infinite set of *variables*. Together, the sets $\mathsf{N_V}$ (of variables) and $\mathsf{N_I}$ (of individual names) form the set $\mathsf{N_T}$ of *terms*. A *first-order (FO) query* q is a first-order formula built from $\mathsf{N_T}$ and the unary and binary predicates from $\mathsf{N_C}$ and $\mathsf{N_R}$. We write $q = \varphi(\vec{v})$ to indicate that q is the FO formula φ whose free variables are among the variables $\vec{v} = v_1, \ldots, v_k$. Variables in $\vec{v}$ are the *answer variables* of q and q is *k-ary* if there are k answer variables. A *conjunctive query* is an FO query q of the form $\exists \vec{u}.\psi(\vec{u}, \vec{v})$, where ψ is a conjunction of *concept atoms* $A(t)$ and *role atoms* $r(t, t')$ with t, t' ranging over $\mathsf{N_T}$. The variables in $\vec{u}$ are called *quantified variables* of q. We use $\mathsf{var}(q)$ to denote the set of all variables in $\vec{u}$ and $\vec{v}$, $\mathsf{qvar}(q)$ for the set of quantified variables, $\mathsf{avar}(q)$ for the set of answer variables, and $\mathsf{term}(q)$ for the terms in q. Slightly abusing notation, we write $\alpha \in q$ if the concept or role atom α occurs in q.

Let $\mathcal{I}$ be an interpretation and $\pi : \mathsf{N_T} \to \Delta^\mathcal{I}$ a partial function such that $\pi(a) = a^\mathcal{I}$ for all $a \in \mathsf{dom}(\pi)$. We inductively define the relation $\mathcal{I} \models^\pi \varphi$ for quantifier-free first-order formulas $\varphi(\vec{v})$ with $\vec{v} \subseteq \mathsf{dom}(\pi)$:

- $\mathcal{I} \models^\pi A(t)$ iff $\pi(t) \in A^\mathcal{I}$;
- $\mathcal{I} \models^\pi r(t, t')$ iff $(\pi(t), \pi(t')) \in r^\mathcal{I}$;
- $\mathcal{I} \models^\pi \neg\varphi$ iff $\mathcal{I} \not\models^\pi \varphi$;
- $\mathcal{I} \models^\pi \varphi_1 \wedge \varphi_2$ iff $\mathcal{I} \models^\pi \varphi_i$ for all $i = 1, 2$;
- $\mathcal{I} \models^\pi \varphi_1 \vee \varphi_2$ iff $\mathcal{I} \models^\pi \varphi_i$ for some $i = 1, 2$.

Now let $q = \exists \vec{u}.\varphi(\vec{u}, \vec{v})$ be a first-order query with φ is quantifier-free. A *match* for $\mathcal{I}$ and q is a mapping $\pi : \mathsf{term}(q) \to \Delta^\mathcal{I}$ such that $\pi(a) = a^\mathcal{I}$ for all $a \in \mathsf{term}(q) \cap \mathsf{N_I}$ and $\mathcal{I} \models^\pi \varphi$. If $\vec{v} = v_1, \ldots, v_k$ with $\pi(v_i) = a_i^\mathcal{I}$ for $1 \leq i \leq k$, then π is called an $(a_1, \ldots, a_k)$-match for $\mathcal{I}$ and q. If such a match exists, we write $\mathcal{I} \models q[a_1, \ldots, a_k]$. A *certain answer* for a k-ary conjunctive query q and a knowledge base $\mathcal{K}$ is a tuple $(a_1, \ldots, a_k)$ of individual names such that $a_1, \ldots, a_k$ occur in $\mathcal{K}$ and $\mathcal{I} \models q[a_1, \ldots, a_k]$ for each model $\mathcal{I}$ of $\mathcal{K}$. We use $\mathsf{cert}(q, \mathcal{K})$ to denote the set of all certain answers for q and $\mathcal{K}$. This defines the querying problem studied in this paper: to compute $\mathsf{cert}(q, \mathcal{K})$ given an $\mathcal{ELH}_\bot^{dr}$ knowledge base $\mathcal{K}$ and a CQ q.

Note that CQ answering generalizes *instance checking*, the problem of deciding whether $\mathcal{K} \models C(a)$, for C a $\mathcal{ELH}_\bot^{dr}$-concept: any such C can be easily unfolded into a conjunctive query q_C such that $a \in \mathsf{cert}(q_C, \mathcal{K})$ iff $\mathcal{K} \models C(a)$. For example, the instance query $A \sqcap \exists s.B(a)$ can be unfolded into $\exists u.(A(a) \wedge s(a, u) \wedge B(u))$.

In the remainder of this paper we use the *unique name assumption*: $a^\mathcal{I} \neq b^\mathcal{I}$ for all interpretations $\mathcal{I}$ and all $a, b \in \mathsf{N_I}$ with $a \neq b$. It is not hard to see that this has no impact on certain answers.

We also assume, w.l.o.g., that (i) queries contain only individual names that occur in the KB against which they

are asked, (ii) TBoxes do not contain domain restrictions, (iii) TBoxes contain exactly one range restriction per role name, (iv) if $\mathcal{K} \models r \sqsubseteq s$ and $\mathsf{ran}(r) \sqsubseteq C, \mathsf{ran}(s) \sqsubseteq D$ are in $\mathcal{T}$, then $C \sqsubseteq_{\mathcal{T}} D$, and (v) there are no $r, s \in \mathsf{N_R}$ with $r \neq s$, $\mathcal{K} \models r \sqsubseteq s$, and $\mathcal{K} \models s \sqsubseteq r$. These five assumptions can be made w.l.o.g. This is true for assumption (i) because one can extend the KB with tautological assertions to introduce additional individual names before querying; for assumption (ii) because $\mathsf{dom}(r) \sqsubseteq C$ is equivalent to $\exists r.\top \sqsubseteq C$; for assumption (iii) because two range restrictions $\mathsf{ran}(r) \sqsubseteq C$ and $\mathsf{ran}(r) \sqsubseteq C'$ are equivalent to $\mathsf{ran}(r) \sqsubseteq C \sqcap C'$ and we can always introduce a range restriction $\mathsf{ran}(r) \sqsubseteq \top$ for each role name r; for assumption (iv) because $\{r \sqsubseteq s, \mathsf{ran}(r) \sqsubseteq C, \mathsf{ran}(s) \sqsubseteq D\}$ with $C \not\sqsubseteq_{\mathcal{T}} D$ is equivalent to $\{r \sqsubseteq s, \mathsf{ran}(r) \sqsubseteq C \sqcap D, \mathsf{ran}(s) \sqsubseteq D\}$; and for assumption (v) because if $\mathcal{K} \models r \sqsubseteq s$ and $\mathcal{K} \models s \sqsubseteq r$ with $s \neq r$, we can simply substitute r with s in $\mathcal{K}$ and q.

3 ABox Rewriting / Canonical Models

The rewriting of the ABox consists of an extension of the ABox to a canonical model of the knowledge base. For the remainder of this section, we fix an $\mathcal{ELH}^{dr}_{\bot}$-KB $\mathcal{K} = (\mathcal{T}, \mathcal{A})$. We use $\mathsf{sub}(\mathcal{T})$ to denote the set of all subconcepts of concepts that occur in $\mathcal{T}$, $\mathsf{rol}(\mathcal{T})$ for the set of role names that occur in $\mathcal{T}$, and $\mathsf{Ind}(\mathcal{A})$ for the set of individual names that occur in $\mathcal{A}$. We also use $\mathsf{ran}_{\mathcal{T}}(r)$ to denote the (unique) concept C with $\mathsf{ran}(r) \sqsubseteq C \in \mathcal{T}$, and set

$$\begin{array}{rcl} \mathsf{ran}(\mathcal{T}) & := & \{\mathsf{ran}_{\mathcal{T}}(r) \mid r \in \mathsf{rol}(\mathcal{T})\} \\ \mathsf{NI}_{\mathsf{aux}} & := & \{x_{C,D} \mid C \in \mathsf{ran}(\mathcal{T}) \text{ and } D \in \mathsf{sub}(\mathcal{T})\}, \end{array}$$

assuming $\mathsf{N_I} \cap \mathsf{N}_{\mathsf{Iaux}} = \emptyset$. The canonical model $\mathcal{I}_{\mathcal{K}}$ of $\mathcal{K}$ is defined in Figure 1. It is standard to show the following (similar to proofs in [Baader *et al.*, 2005b; Lutz and Wolter, 2007]):

Proposition 1. *If $\mathcal{K}$ is consistent, then $\mathcal{I}_{\mathcal{K}}$ is a model of $\mathcal{K}$.*

Note that the cardinality of $\Delta^{\mathcal{I}_{\mathcal{K}}}$ is only quadratic in the size of $\mathcal{K}$, and linear if $\mathcal{T}$ does not contain range restrictions. The model $\mathcal{I}_{\mathcal{K}}$ can be computed in polynomial time since subsumption and instance checking in $\mathcal{ELH}^{dr}_{\bot}$ can be decided in poly-time [Baader *et al.*, 2008]. In fact, there is an easy rule-based procedure for computing the canonical model and the rules can be implemented as standard database operations; see the workshop publication [Lutz *et al.*, 2008] for details. Consistency of $\mathcal{K}$ can also be checked in polynomial time [Baader *et al.*, 2008]. $\mathcal{I}_{\mathcal{K}}$ can be used for instance checking: it can be shown that $\mathcal{K} \models C(a)$ iff $\mathcal{I}_{\mathcal{K}} \models C(a)$, for $\mathcal{K}$ consistent, C an $\mathcal{ELH}^{dr}_{\bot}$-concept, and $a \in \mathsf{Ind}(\mathcal{A})$. Unfortunately, an analogous statement for conjunctive queries, namely $(a_1, \ldots, a_k) \in \mathsf{cert}(q, \mathcal{K})$ iff $\mathcal{I}_{\mathcal{K}} \models q[a_1, \ldots, q_k]$, does not hold. This is due to two reasons, the first one being that $\mathcal{I}_{\mathcal{K}}$ may contain unnecessary elements:

Example 2. Take $\mathcal{K}_1 = (\mathcal{T}_1, \mathcal{A}_1)$ with $\mathcal{T}_1 = \{A \sqsubseteq A\}$ and $\mathcal{A}_1 = \{B(a)\}$, and $q_1 = \exists u.(B(v) \wedge A(u))$. Then $x_{\top,A} \in A^{\mathcal{I}_{\mathcal{K}_1}}$ and so $\mathcal{I}_{\mathcal{K}_1} \models q_1[a]$, but clearly $a \notin \mathsf{cert}(q_1, \mathcal{K}_1)$.

This deficiency of $\mathcal{I}_{\mathcal{K}}$ is easily repaired by restricting it to elements reachable from some $a^{\mathcal{I}_{\mathcal{K}}}$ with $a \in \mathsf{Ind}(\mathcal{A})$. Formally, we define $\mathsf{Ind}(\mathcal{A})^{\mathcal{I}} = \{a^{\mathcal{I}} \mid a \in \mathsf{Ind}(\mathcal{A})\}$ for each interpretation $\mathcal{I}$. A *path* in $\mathcal{I}$ is a finite sequence $d_0 r_1 d_1 \cdots r_n d_n$, $n \geq 0$, where $d_0 \in \mathsf{Ind}(\mathcal{A})^{\mathcal{I}}$ and $(d_i, d_{i+1}) \in r^{\mathcal{I}}_{i+1}$ for all $i < n$. We use $\mathsf{paths}_{\mathcal{A}}(\mathcal{I})$ to denote the set of all paths in $\mathcal{I}$ and for all $p \in \mathsf{paths}_{\mathcal{A}}(\mathcal{I})$, $\mathsf{tail}(p)$ to denote the last element d_n in p.

Now $\mathcal{I}^r_{\mathcal{K}}$ denotes the restriction of $\mathcal{I}_{\mathcal{K}}$ to those $d \in \Delta^{\mathcal{I}_{\mathcal{K}}}$ for which there exists a $p \in \mathsf{paths}_{\mathcal{A}}(\mathcal{I}_{\mathcal{K}})$ such that $d = \mathsf{tail}(p)$. Then $\mathcal{I}^r_{\mathcal{K}}$ provides the correct certain answers to the query q_1 from Example 2. A much more severe deficiency of $\mathcal{I}_{\mathcal{K}}$ (and $\mathcal{I}^r_{\mathcal{K}}$) is the following:

Example 3. Take $\mathcal{K}_2 = (\mathcal{T}_2, \mathcal{A}_2)$ with $\mathcal{T}_2 = \{A \sqsubseteq \exists r.B \sqcap \exists s.B\}$ and $\mathcal{A}_2 = \{A(a)\}$, and $q_2 = \exists u.(r(v,u) \wedge s(v,u))$. Then $(a, x_{\top,B}) \in r^{\mathcal{I}_{\mathcal{K}_2}}$ and $(a, x_{\top,B}) \in s^{\mathcal{I}_{\mathcal{K}_2}}$ and therefore $\mathcal{I}_{\mathcal{K}_2} \models q_2[a]$, but clearly $a \notin \mathsf{cert}(q_2, \mathcal{K}_2)$.

In principle, this problem can be overcome by replacing $\mathcal{I}^r_{\mathcal{K}}$ with its unraveling into a less constrained, tree-like model. In the following, we introduce unraveling as a general operation on models. Let $\mathcal{R}$ be the set of role inclusions in $\mathcal{K}$. The $(\mathcal{A}, \mathcal{R})$-*unraveling* $\mathcal{J}$ of $\mathcal{I}$ is defined as follows:

$$\begin{array}{rcl} \Delta^{\mathcal{J}} & := & \mathsf{paths}_{\mathcal{A}}(\mathcal{I}) \text{ and } a^{\mathcal{J}} := a^{\mathcal{I}} \text{ for all } a \in \mathsf{Ind}(\mathcal{A}) \\ A^{\mathcal{J}} & := & \{p \mid \mathsf{tail}(p) \in A^{\mathcal{I}}\} \\ r^{\mathcal{J}} & := & \{(d,e) \mid d, e \in \mathsf{Ind}(\mathcal{A})^{\mathcal{I}} \wedge (d,e) \in r^{\mathcal{I}}\} \cup \\ & & \{(p, p \cdot se) \mid p, p \cdot se \in \Delta^{\mathcal{J}} \text{ and } \mathcal{R} \models s \sqsubseteq r\} \end{array}$$

where "$\cdot$" denotes concatenation. Denote by $\mathcal{U}_{\mathcal{K}}$ the $(\mathcal{A}, \mathcal{R})$-unraveling of $\mathcal{I}^r_{\mathcal{K}}$. Notice that the construction of $\mathcal{U}_{\mathcal{K}}$ from $\mathcal{I}^r_{\mathcal{K}}$ does not depend on the concept inclusions in $\mathcal{T}$, but only on $\mathcal{R}$. The following result is proved similarly to the analogous result for the DL $\mathcal{ELI}^f$ in [Krisnadhi and Lutz, 2007].

Proposition 4. *If $\mathcal{K}$ is consistent, then for all k-ary conjunctive queries q and all $a_1, \ldots, a_k \in \mathsf{Ind}(\mathcal{A})$, we have $(a_1, \ldots, a_k) \in \mathsf{cert}(q, \mathcal{K})$ iff $\mathcal{U}_{\mathcal{K}} \models q[a_1, \ldots, a_k]$.*

By Proposition 4, $\mathcal{U}_{\mathcal{K}}$ gives the correct answers to conjunctive queries, but in contrast to $\mathcal{I}^r_{\mathcal{K}}$, it is typically infinite. Thus, we do not use it as a target for ABox rewriting and work with $\mathcal{I}^r_{\mathcal{K}}$. To overcome the problem indicated in Example 3 and similar ones (see [Lutz *et al.*, 2008]), we use query rewriting.

4 Query Rewriting

Our aim is to rewrite the original CQ q into an FO query $q^*_{\mathcal{R}}$ such that $\mathcal{U}_{\mathcal{K}} \models q[a_1, \ldots, a_k]$ iff $\mathcal{I}^r_{\mathcal{K}} \models q^*_{\mathcal{R}}[a_1, \ldots, a_k]$ for all $a_1, \ldots, a_k \in \mathsf{Ind}(\mathcal{A})$. By Proposition 4, we obtain the desired answers $\mathsf{cert}(q, \mathcal{K})$ by using $\mathcal{I}^r_{\mathcal{K}}$ (the rewriting of $\mathcal{A}$) as a relational database instance and replacing q with $q^*_{\mathcal{R}}$. We now formulate the main result of this paper.

Theorem 5. *For every finite set of role inclusions $\mathcal{R}$ and k-ary CQ q, one can construct in polynomial time a k-ary FO query $q^*_{\mathcal{R}}$ such that for all $\mathcal{ELH}^{dr}_{\bot}$-KBs $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ with $\mathcal{R}$ the set of role inclusions in $\mathcal{T}$ and all $a_1, \ldots, a_k \in \mathsf{Ind}(\mathcal{A})$, we have $\mathcal{I}^r_{\mathcal{K}} \models q^*_{\mathcal{R}}[a_1, \ldots, a_k]$ iff $\mathcal{U}_{\mathcal{K}} \models q[a_1, \ldots, a_k]$.*

In the remainder of this section, we show how to construct $q^*_{\mathcal{R}}$. The query $q^*_{\mathcal{R}}$ contains one additional unary predicate $\mathsf{Aux}(x)$; we assume that Aux is always interpreted as $\Delta^{\mathcal{I}^r_{\mathcal{K}}} \setminus \mathsf{Ind}(\mathcal{A})^{\mathcal{I}^r_{\mathcal{K}}}$ in $\mathcal{I}^r_{\mathcal{K}}$. Fix a finite set $\mathcal{R}$ of role inclusions and a k-ary conjunctive query q. To construct $q^*_{\mathcal{R}}$, we use several auxiliary definitions. Let $\sim_q$ denote the smallest relation on $\mathsf{term}(q)$ that includes the identity relation, is transitive, and satisfies the following closure condition:

$$
\begin{array}{rcl}
\Delta^{\mathcal{I}_\mathcal{K}} & := & \mathsf{Ind}(\mathcal{A}) \uplus \mathsf{NI}_{\mathsf{aux}} \text{ and } a^{\mathcal{I}_\mathcal{K}} := a \text{ for all } a \in \mathsf{Ind}(\mathcal{A}) \\
A^{\mathcal{I}_\mathcal{K}} & := & \{a \in \mathsf{Ind}(\mathcal{A}) \mid \mathcal{K} \models A(a)\} \cup \{x_{C,D} \in \mathsf{NI}_{\mathsf{aux}} \mid \mathcal{K} \models C \sqcap D \sqsubseteq A\} \\
r^{\mathcal{I}_\mathcal{K}} & := & \{(a,b) \in \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A}) \mid s(a,b) \in \mathcal{A} \text{ and } \mathcal{K} \models s \sqsubseteq r\} \cup \\
 & & \{(a, x_{C,D}) \in \mathsf{Ind}(\mathcal{A}) \times \mathsf{NI}_{\mathsf{aux}} \mid \mathcal{K} \models \exists s.D(a), \mathsf{ran}_\mathcal{T}(s) = C, \text{ and } \mathcal{K} \models s \sqsubseteq r\} \cup \\
 & & \{(x_{C,D}, x_{C',D'}) \in \mathsf{NI}_{\mathsf{aux}} \times \mathsf{NI}_{\mathsf{aux}} \mid \mathcal{K} \models C \sqcap D \sqsubseteq \exists r.D', \mathsf{ran}_\mathcal{T}(s) = C', \text{ and } \mathcal{K} \models s \sqsubseteq r\}
\end{array}
$$

Figure 1: The canonical model $\mathcal{I}_\mathcal{K}$.

$(*)$ if $r_1(s,t), r_2(s',t') \in q$ with $t \sim_q t'$, then $s \sim_q s'$.

The relation $\sim_q$ is central to our rewriting procedure. To see this, let $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ be such that the role inclusions of $\mathcal{T}$ coincide with $\mathcal{R}$. Intuitively, $\mathcal{U}_\mathcal{K}$ is produced from $\mathcal{I}^r_\mathcal{K}$ by keeping the $\mathsf{Ind}(\mathcal{A})$-part intact and relaxing the Aux-part into a collection of trees. To understand $(*)$, first assume $t = t'$. Then $(*)$ describes a non-tree situation in the query since $t = t'$ has two predecessors s and s'. Therefore, any match of the query in $\mathcal{I}^r_\mathcal{K}$ that maps t to the Aux-part should map s and s' to the same element; otherwise such a match cannot be reproduced in $\mathcal{U}_\mathcal{K}$. The case where $t \sim_q t'$ instead of $t = t'$ can be understood inductively.

It is not hard to verify that $\sim_q$ is an equivalence relation and can be computed in time polynomial in the size of q. For $t \in \mathsf{term}(q)$, we use $[t]$ to denote the equivalence class of t w.r.t. $\sim_q$ and define, for any equivalence class ζ of $\sim_q$, the sets:

$$
\begin{array}{l}
\mathsf{pre}(\zeta) := \{t \mid r(t,t') \in q \text{ for some } r \in \mathsf{N_R} \text{ and } t' \in \zeta\}, \text{and} \\
\mathsf{in}(\zeta) := \{r \mid r(t,t') \in q \text{ for some } t \in \mathsf{term}(q) \text{ and } t' \in \zeta\}.
\end{array}
$$

For $R \subseteq \mathsf{N_R}$ and $r \in \mathsf{N_R}$, r is called an *implicant* of R if $\mathcal{R} \models r \sqsubseteq s$ for all $s \in R$. It is called a *prime implicant* if, additionally, $\mathcal{R} \not\models r \sqsubseteq r'$ for all implicants r' of R with $r \neq r'$. By assumption (v) in Section 2, there is a prime implicant for any set $R \subseteq \mathsf{N_R}$ for which there is an implicant. We define:

- $\mathsf{Fork}_=$ is the set of pairs $(\mathsf{pre}(\zeta), \zeta)$ with $\mathsf{pre}(\zeta)$ of cardinality at least two;
- $\mathsf{Fork}_\neq$ is the set of variables $v \in \mathsf{qvar}(q)$ such that there is no implicant of $\mathsf{in}([v])$;
- $\mathsf{Fork}_\mathcal{H}$ is the set of of pairs (I, ζ) such that $\mathsf{pre}(\zeta) \neq \emptyset$, there is a prime implicant of $\mathsf{in}(\zeta)$ that is not contained in $\mathsf{in}(\zeta)$, and I is the set of all prime implicants of $\mathsf{in}(\zeta)$;
- Cyc is the set of variables $v \in \mathsf{qvar}(q)$ such that there are $r_0(t_0, t_0'), \ldots, r_m(t_m, t_m'), \ldots, r_n(t_n, t_n') \in q$, $n, m \geq 0$, with $v \sim_q t_i$ for some $i \leq n$, $t_i' \sim_q t_{i+1}$ for all $i < n$, and $t_n' \sim_q t_m$.

It is not hard to see that $\mathsf{Fork}_=$, $\mathsf{Fork}_\neq$, $\mathsf{Fork}_\mathcal{H}$, and Cyc can be computed in time polynomial in the size of q. For each equivalence class ζ of $\sim_q$, choose a representative $t_\zeta \in \zeta$ and if $\mathsf{pre}(\zeta) \neq \emptyset$, choose a $t^{\mathsf{pre}}_\zeta \in \mathsf{pre}(\zeta)$. For $q = \exists \vec{u}.\psi$, the rewritten query $q^*_\mathcal{R}$ is now defined as $\exists \vec{u}.(\psi \wedge \varphi_1 \wedge \varphi_2 \wedge \varphi_3)$, where φ_1, φ_2, and φ_3 are as follows:

$$
\begin{array}{rcl}
\varphi_1 & := & \displaystyle\bigwedge_{v \in \mathsf{avar}(q) \cup \mathsf{Fork}_\neq \cup \mathsf{Cyc}} \neg \mathsf{Aux}(v) \\
\varphi_2 & := & \displaystyle\bigwedge_{(\{t_1, \ldots, t_k\}, \zeta) \in \mathsf{Fork}_=} (\, \mathsf{Aux}(t_\zeta) \rightarrow \bigwedge_{1 \leq i < k} t_i = t_{i+1} \,) \\
\varphi_3 & := & \displaystyle\bigwedge_{(I, \zeta) \in \mathsf{Fork}_\mathcal{H}} (\, \mathsf{Aux}(t_\zeta) \rightarrow \bigvee_{r \in I} r(t^{\mathsf{pre}}_\zeta, t_\zeta) \,)
\end{array}
$$

In Appendix B, we show that $q^*_\mathcal{R}$ is as required. The following examples illustrate the definition of $q^*_\mathcal{R}$. Since φ_3 is simply true when $\mathcal{R} = \emptyset$, we omit it in this case.
(1) Let $\mathcal{R} = \emptyset$ and consider $q = \exists u.\psi$ with $\psi = r(v,u) \wedge r(v',u)$. This query illustrates the role of $\mathsf{Fork}_=$. $\sim_q$ consists of the equivalence classes $\{v, v'\}$ and $\{u\}$. We have $\mathsf{pre}(\{u\}) = \{v, v'\}$ and $\mathsf{in}(\{u\}) = \{r\}$. Hence $\mathsf{Fork}_= = \{(\{v, v'\}, \{u\})\}$ and $\mathsf{Fork}_\neq = \mathsf{Fork}_\mathcal{H} = \mathsf{Cyc} = \emptyset$. We obtain

$$q^*_\mathcal{R} = \exists u.(\psi \wedge \neg \mathsf{Aux}(v) \wedge \neg \mathsf{Aux}(v') \wedge (\mathsf{Aux}(u) \rightarrow v = v')).$$

(2) Let $\mathcal{R} = \emptyset$ and consider $q = \exists u.(r(v,u) \wedge s(u,u))$. This query illustrates the role of Cyc. We have $u \in \mathsf{Cyc}$ and so

$$q^*_\mathcal{R} = \exists u.(r(v,u) \wedge s(u,u) \wedge \neg \mathsf{Aux}(v) \wedge \neg \mathsf{Aux}(u)).$$

(3) Let $\mathcal{R} = \emptyset$ and consider $q = \exists u.(r(v,u) \wedge s(v,u))$ from Example 3. This query illustrates the role of $\mathsf{Fork}_\neq$. We have $\mathsf{in}(\{u\}) = \{r, s\}$. There is no implicant of $\mathsf{in}(\{u\})$ in $\mathsf{in}(\{u\})$ and thus $u \in \mathsf{Fork}_\neq$. We obtain

$$q^*_\mathcal{R} = \exists u.(r(v,u) \wedge s(v,u) \wedge \neg \mathsf{Aux}(v) \wedge \neg \mathsf{Aux}(u)).$$

For $\mathcal{R} = \{s \sqsubseteq r\}$ and the same query q, s is an implicant of $\mathsf{in}(\{u\})$ in $\mathsf{in}(\{u\})$. Thus $u \notin \mathsf{Fork}_\neq$. Observe that $\mathsf{Fork}_\mathcal{H} = \emptyset$ as the prime implicant s of $\mathsf{in}(\{u\})$ is contained in $\mathsf{in}(\{u\})$. We obtain

$$q^*_\mathcal{R} = \exists u.(r(v,u) \wedge s(v,u) \wedge \neg \mathsf{Aux}(v)).$$

Finally, assume $\mathcal{R} = \{r_0 \sqsubseteq r, r_0 \sqsubseteq s\}$. Again $u \notin \mathsf{Fork}_\neq$, but now the prime implicant r_0 of $\mathsf{in}(\{u\})$ is not contained in $\mathsf{in}(\{u\})$. Thus, $\mathsf{Fork}_\mathcal{H} = \{(\{r_0\}, \{u\})\}$ and we obtain

$$q^*_\mathcal{R} = \exists u.(r(v,u) \wedge s(v,u) \wedge \neg \mathsf{Aux}(v) \wedge (\mathsf{Aux}(u) \rightarrow r_0(v,u))).$$

(4) For queries $q_C = \exists \vec{v}.\psi$ that result from the unfolding of an $\mathcal{EL}$-concept C or have no quantified variables almost no query rewriting is needed: in both cases we have

$$q^*_\mathcal{R} = \exists \vec{v}.(\psi \wedge \textstyle\bigwedge_{v \in \mathsf{avar}(q)} \neg \mathsf{Aux}(v)).$$

(5) Let $\mathcal{R} = \emptyset$ and let $q = \exists v_0, \ldots, v_7.\psi$ be the query shown in Figure 2, where all variables are quantified. Then $\sim_q$

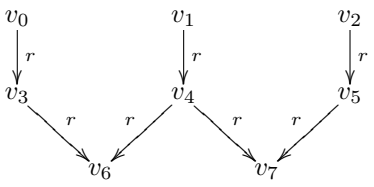

Figure 2: Query for Example (5).

consists of the equivalence classes $\{v_0, v_1, v_2\}$, $\{v_3, v_4, v_5\}$, $\{v_6\}$, and $\{v_7\}$. Assume that the chosen representative for $\{v_3, v_4, v_5\}$ is v_3. Then, we have

$$q^*_{\mathcal{R}} = \exists v_0, \ldots, v_7.(\ \psi \wedge \mathsf{Aux}(v_6) \rightarrow (v_3 = v_4) \wedge \mathsf{Aux}(v_7) \rightarrow (v_4 = v_5) \wedge \mathsf{Aux}(v_3) \rightarrow ((v_0 = v_1) \wedge (v_1 = v_2))$$

(6) Let $\mathcal{R} = \emptyset$ and $q_n = \exists v_0, \ldots, v_{n-1}.q'_n$ with $q'_n = \bigwedge_{i,j<n} r(v_i, v_j)$ an n-clique. Then $\sim_q$ consists of a single equivalence class $\{v_0, \ldots, v_{n-1}\}$. Assume that the representative is v_0. Then $(q^c_n)^*_{\mathcal{R}}$ is

$$\exists v_0, \ldots, v_{n-1}.(q'_n \wedge \neg\mathsf{Aux}(v_0) \wedge \cdots \wedge \neg\mathsf{Aux}(v_{n-1}) \wedge \mathsf{Aux}(v_0) \rightarrow ((v_0 = v_1) \wedge \cdots \wedge (v_{n-2} = v_{n-1})))$$

which can be simplified to the equivalent

$$\exists v_0, \ldots, v_{n-1}.(q'_n \wedge \neg\mathsf{Aux}(v_0) \wedge \cdots \wedge \neg\mathsf{Aux}(v_{n-1})).$$

As illustrated by the last example, we can drop a conjunct from vp_2 whenever the variable occurring in its precondition occurs in a conjunct of vp_1.

In the following, we comment on certain relevant properties of $q^*_{\mathcal{R}}$ and on consequences of our approach regarding the computational complexity of query answering in $\mathcal{ELH}^{dr}_{\bot}$. Firstly, the size of $q^*_{\mathcal{R}}$ is bounded by $\mathcal{O}(nm)$, where n is the size of q and m is $\mathsf{min}\{1, |\mathcal{R}|\}$. To see this, note that the number of conjuncts in of φ_1 is bounded by the number of variables in q. For φ_2, let $\mathsf{Fork}_= = \{(T_1, \zeta_1), \ldots, (T_\ell, \zeta_\ell)\}$. Then $|T_1| + \cdots + |T_\ell|$ is bounded by the number of role atoms in q, and thus the size of φ_2 is $\mathcal{O}(n)$. Finally, the number of conjuncts in φ_3 is bounded by the number of quantified variables in q and each conjunct has at most m disjuncts. Note that $q^*_{\mathcal{R}}$ is thus of size $\mathcal{O}(n)$ when $\mathcal{R} = \emptyset$.

Secondly, since ψ is a conjunct of the body of $q^*_{\mathcal{R}}$ it is readily checked that $q^*_{\mathcal{R}}$ is domain independent, and thus can be expressed as an SQL query. Moreover it is of the form $\exists \vec{u}.\psi$ with ψ quantifier-free. Thus, we obtain that the combined complexity of deciding whether $\mathcal{K} \models q[a_1, \ldots, a_k]$ is in NP: construct (in poly-time) $\mathcal{I}^r_{\mathcal{K}}$ and $q^*_{\mathcal{R}} = \exists \vec{u}.\psi$, then check whether $\mathcal{I}^r_{\mathcal{K}} \models \exists \vec{u}.\psi[a_1, \ldots, a_k]$ using the obvious NP algorithm for model checking this class of formulas. Since an NP lower bound is trivial, we obtain NP-completeness. See [Krötzsch *et al.*, 2007; Rosati, 2007] for similar results regarding other variants of $\mathcal{EL}$.

5 Implementation and Experiments

To validate the value of the proposed approach, we have conducted a series of experiments based on the NCI thesaurus (version 08.08d), which is a well-known ontology from the bio-medical domain [Sioutos *et al.*, 2006]. We have extracted an $\mathcal{EL}$-TBox that contains approximately 65 thousand (65K) primitive concept names, 70 primitive roles, and over 70K concept inclusions and concept definitions. The auxiliary part of canonical models, which is independent of the ABox, consists of 702K concept assertions and 171K role assertions (tuples). For the experiments reported below, we used the IBM DB2 DBMS (version 9.5.0 running on SUN Fire-280R server with two UltraSPARC III 1.2GHz CPUs, 4GB memory, and 1TB storage under Solaris 5.10).

For rewriting ABoxes into canonical models, we have used a rule-based approach implemented via iterated querying; see [Lutz *et al.*, 2008] for more details. As relational systems are not optimized to handle tens of thousands of relatively small relations, one for each concept and role name in the ontology, we have used only two relations to represent $\mathcal{I}^r_{\mathcal{K}}$:

`acbox(conceptid,indid)` and
`arbox(roleid,domain-indid,range-indid)`,

where `conceptid` and `roleid` are numerical identifiers for concept names and roles names, `indid`, `domain-indid`, and `range-indid` are numerical identifiers for individuals from $\mathsf{N_I} \cup \mathsf{N_{I_{aux}}}$, `acbox` represents concept memberships, and `arbox` role memberships. Indexes were generated on the attributes `conceptid,indid` and on `roleid,domain-indid` and `roleid,range-indid`, using B+trees. We distinguish individuals from $\mathsf{N_I}$ and $\mathsf{N_{I_{aux}}}$ by positive and negative identifiers, and thus need not store Aux as a relation. As an example for this representation, take the query $q = \mathit{Nerve}(x) \wedge \neg\mathsf{Aux}(x)$, which translates into the SQL statement

```
select indid from  acbox
where  conceptid=141723 and indid > 0
```

Data sets (ABox instances) for our experiments were generated randomly. When generating concept assertions, we have focused on most specific concept names, i.e., concept names without any subsumees in the TBox. The generation of role assertions was guided by the domain and range restrictions in NCI. The numbers of concept and role assertions in the initial and rewritten ABoxes used in our experiments are reported in Figure 3 in thousands/millions. Due to implementation particularities of existing DB systems, ABox rewriting took up to several hours. We point out that (i) this high time consumption is not inherent to our approach and (ii) ABox rewriting can be implemented as an offline task in typical applications such as online analytical processing (OLAP) and data warehousing.

Figure 3 summarizes the running times (in seconds) for each of the test queries for varying sizes of data; for each query we list the number of concept and role atoms in parentheses; the structure of the queries ranges from simple chains (Q1) and star queries (Q2,Q3,Q4) to queries with cycles in their bodies (Q5). We show only a few representative samples here.

The experimental results can be interpreted as follows. Firstly, the rewriting of moderately-sized ABoxes into the

[1] This time is solely due to a large size of the result (60M tuples).

Number of assertions in ABox of $\mathcal{K}$									
Concept	100K	100K	100K	200K	200K	200K	400K	800K	1.6M
Role	25K	50K	75K	40K	65K	90K	360K	1.5M	5.8M
Number of assertions in $\mathcal{I}^r_{\mathcal{K}}$									
Concept	440K	440K	441K	683K	683K	684K	1.3M	2.6M	5.1M
Role	197K	237K	273K	323K	371K	414K	986K	2.7M	8.2M
Query Execution Time in seconds									
Q1 (2c1r)	0.19	0.19	0.20	0.23	0.25	0.24	0.27	0.46	0.59
Q2 (3c2r)	0.23	0.22	0.23	0.52	0.25	0.56	0.33	0.42	0.69
Q3 (3c2r)	0.25	0.27	0.26	0.31	0.31	0.31	0.42	0.86	1.13
Q4 (4c3r)	0.24	0.23	0.23	0.25	0.26	0.25	0.31	0.42	1.44
Q5 (5c5r)	0.36	0.36	0.30	0.60	0.34	0.45	2.24	7.93	128[1]

Figure 3: Summary of Experimental Results.

canonical model $\mathcal{I}^r_{\mathcal{K}}$ is well within the storage and query capabilities of existing relational technology. Secondly, query performance scales well even when the naive physical design described above is used. Thirdly, query rewriting does not increase the query processing times; the performance of rewritten queries is almost identical to the performance of the original queries over the completed ABox.

6 Limitations / Negation

Recall from Section 1 that a DL enjoys *polynomial* combined FO rewritability if it has combined FO rewritability such that the blowup of ABox rewriting (but not necessarily of query rewriting) is at most polynomial. Since ABoxes in realistic applications are large, combined FO rewritability that is not polynomial in this sense does not seem to be of much use. The following result gives a fundamental limitation of polynomial combined FO rewritability. Note that *ground* CQ answering, where a CQ may contain only individual names but not variables, is the decisional variant of CQ answering.

Theorem 6. *If the data complexity of ground CQ answering in a DL $\mathcal{L}$ is not in* PTIME*, then $\mathcal{L}$ does not enjoy polynomial combined FO rewritability.*

Proof. We show the contrapositive. Assume that $\mathcal{L}$ enjoys polynomial combined FO rewritability, i.e., there are effectively computable mappings δ that takes each $\mathcal{L}$-TBox $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ to a first-order structure $\delta(\mathcal{K})$ and γ that takes each pair $(q, \mathcal{T})$, with q a k-ary conjunctive query and $\mathcal{T}$ an $\mathcal{L}$-TBox, to a first-order formula $\gamma(q, \mathcal{T})$ with k free variables such that δ can be computed in polynomial time, the size of $\delta(\mathcal{K})$ is polynomial in the size of $\mathcal{K}$ for all $\mathcal{K}$, and the following condition holds: for all combined $\mathcal{L}$-KBs $\mathcal{K} = (\mathcal{T}, \mathcal{A})$, all k-ary conjunctive queries q, and all tuples $(a_1, \ldots, a_k) \in \mathsf{N_I}$, $\mathcal{K} \models q[a_1, \ldots, a_k]$ iff $\delta(\mathcal{K}) \models \gamma(q, \mathcal{T})[a_1, \ldots, a_k]$.

Then the data complexity of ground CQ answering in $\mathcal{L}$ is in PTIME: to check whether $\mathcal{K} \models q$ with q ground, we can compute $\delta(\mathcal{K})$ in polynomial time and $\gamma(q, \mathcal{T})$ in constant time (as the size of q and $\mathcal{T}$ is constant), and then use FO model checking to decide whether $\delta(\mathcal{K}) \models \gamma(q, \mathcal{T})$. The latter can be done in LOGSPACE. ❑

For expressive DLs such as $\mathcal{ALC}$ and $\mathcal{SHIQ}$, the data complexity of ground CQ answering is co-NP-hard, thus Theorem 6 implies that these DLs do not enjoy polynomial combined FO rewritability unless PTIME = co-NP.

In many applications, it is natural to admit also negated concept assertions in the ABox and to extend conjunctive queries with negated concept atoms in order to query the resulting ABoxes, see e.g. [Patel *et al.*, 2007]. Let us call ABoxes, queries, and KBs of this form *literal*. The result stated in Theorem 6 also applies to literal KBs and literal queries. Since the data complexity of literal CQ answering over literal $\mathcal{EL}$-KBs (where $\mathcal{EL}$ is $\mathcal{ELH}^{dr}_{\bot}$ without $\bot$, role hierarchies, and domain and range restrictions) is co-NP-hard [Schaerf, 1993], it follows that even this mild extension of $\mathcal{EL}$ does not enjoy polynomial combined FO rewritability.

However, there is still a (pragmatic, yet formal) way to add negation to our approach. In the following, we sketch an incomplete approach to literal CQ answering over literal $\mathcal{ELH}^{dr}_{\bot}$-KBs. Its incompleteness can be precisely characterized in terms of an epistemic semantics for negated concept atoms in literal CQs, inspired by [Calvanese *et al.*, 2007a].

As a preliminary, we assume *standard names*, i.e., interpretation domains $\Delta^{\mathcal{I}}$ have to be a subset of the (countably infinite) set $\mathsf{N_I}$ of individual names and $a^{\mathcal{I}} = a$ for all $\mathcal{I}$ and all $a \in \mathsf{N_I}$. For a literal $\mathcal{ELH}^{dr}_{\bot}$-KB $\mathcal{K}$, an interpretation $\mathcal{I}$, and a literal CQ $q = \exists \vec{u}.\psi(\vec{u}, \vec{v})$ with $\vec{v} = v_1, \ldots, v_k$, we set $\mathcal{I} \models_e q[a_1, \ldots, a_k]$ iff there exists a variable assignment π with $\pi(v_i) = a_i$ for $1 \leq i \leq k$ and

- $\mathcal{I} \models_\pi \alpha$ for all positive atoms α in ψ,
- $\mathcal{K} \models \neg A(a)$ for all $\neg A(a)$ in ψ with $a \in \mathsf{N_I}$, and
- $\mathcal{K} \models \neg A(\pi(v))$ for all $\neg A(v)$ in ψ with $v \in \mathsf{N_V}$.

Now set $\mathcal{K} \models_e q[a_1, \ldots, a_k]$ iff $\mathcal{I} \models_e q[a_1, \ldots, a_k]$ for all models $\mathcal{I}$ of $\mathcal{K}$ with standard names. Thus, answers to negated atoms do not depend on a concrete interpretation $\mathcal{I}$, but only on deducibility. Epistemic semantics is sound: every answer is also an answer under the standard semantics (but not vice versa); it is also conservative: it yields the same answers as standard semantics when either the KB or the query does not contain negation. An example for which the epistemic semantics is different from the standard semantics is given by $\mathcal{A} = \{A'(a), \neg A(a)\}$, $\mathcal{T} = \{A' \sqsubseteq \exists r.\top, \exists r.B \sqsubseteq A\}$, and $q = \exists u.r(a, u) \wedge \neg B(u)$. Then $\mathcal{K} \models q$ but $\mathcal{K} \not\models_e q$.

We give a simple (and poly-time) reduction of epistemic literal CQ answering over literal $\mathcal{ELH}^{dr}_{\bot}$-KBs to standard CQ answering over $\mathcal{ELH}^{dr}_{\bot}$-KBs. Via the rewritings presented in the main part of this paper, the reduction enables the use of RDBMSs also for the case of $\mathcal{ELH}^{dr}_{\bot}$ with negation (under the epistemic semantics).

Given a consistent literal $\mathcal{ELH}^{dr}_{\bot}$-KB $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ and a literal query q, replace every literal $\neg A$ in q with a fresh concept name $\overline{A}$. Additionally, add $\top \sqsubseteq \overline{A}$ to $\mathcal{T}$ whenever $\mathcal{T} \models A \sqsubseteq \bot$, and $\overline{A}(a)$ to $\mathcal{A}$ for all $a \in \mathsf{Ind}(\mathcal{A})$ with $\mathcal{K} \models \neg A(a)$, and then remove all negated assertions from $\mathcal{A}$. Call the result $\mathcal{K}' = (\mathcal{T}', \mathcal{A}')$ and q'.

Theorem 7. *Let $\mathcal{K}$ be a consistent literal $\mathcal{ELH}^{dr}_{\bot}$-KB. Then $\mathcal{K}'$ can be computed in polynomial time and $\mathcal{K} \models_e q[a_1, \ldots, a_k]$ iff $\mathcal{K}' \models q'[a_1, \ldots, a_k]$ for all literal CQs q and all $a_1, \ldots, a_k \in \mathsf{Ind}(\mathcal{A})$.*

7 Conclusion

We have proposed a novel approach to CQ answering in DLs using RDBMSs. Unlike previous approaches, it can be used also for DLs for which the data complexity is between LOGSPACE and PTIME. In particular, this includes DLs that fully admit existential restrictions (both on the left- and right-hand side of concept inclusions; see [Calvanese *et al.*, 2006]) and paves the way to using RDBMSs for CQ answering in the $\mathcal{EL}$ family of DLs. Our experiments exhibit a promising performance even without a sophisticated physical design. One drawback of our approach is the blowup of the data, which is polynomial but still considerable on large data sets. As future work, it might be interesting to reduce this blowup by incorporating the TBox partly into the data and partly into the query. We will also develop effective approaches to update the canonical model/auxiliary data when assertions are added to or deleted from the ABox.

References

[Baader *et al.*, 2005b] F. Baader, S. Brandt, and C. Lutz. Pushing the $\mathcal{EL}$ envelope. In *Proc. of IJCAI05*, pages 364–369. Professional Book Center, 2005.

[Baader *et al.*, 2008] F. Baader, S. Brandt, and C. Lutz. Pushing the $\mathcal{EL}$ envelope further. In *In Proc. of OWLED08*, 2008.

[Calvanese *et al.*, 2006] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Data complexity of query answering in description logics. *Proc. of KR06*, pages 260–270, 2006.

[Calvanese *et al.*, 2007a] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. EQL-Lite: Effective first-order query processing in DLs. In *Proc. of IJCAI07*, pages 274–279. AAAI press, 2007.

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

[Krisnadhi and Lutz, 2007] A. Krisnadhi and C. Lutz. Data complexity in the $\mathcal{EL}$ family of DLs. In *Proc. of LPAR07*, volume 4790 of *LNCS*, pages 333–347. Springer, 2007.

[Krötzsch *et al.*, 2007] M. Krötzsch, S. Rudolph, and P. Hitzler. Conjunctive queries for a tractable fragment of OWL 1.1. In *Proc. of ISWC07*, volume 4825 of *LNCS*, pages 310–323. Springer, 2007.

[Lutz and Wolter, 2007] C. Lutz and F. Wolter. Conservative extensions in the lightweight description logic $\mathcal{EL}$. In *Proc. of CADE21*, volume 4603 of *LNAI*, pages 84–99. Springer, 2007.

[Lutz *et al.*, 2008] C. Lutz, D. Toman, and F. Wolter Conjunctive query answering in $\mathcal{EL}$ using a database system. In *Proc. of OWLED08*, 2008.

[Lutz *et al.*, 2009] C. Lutz, D. Toman, and F. Wolter Conjunctive query answering in the DL $\mathcal{EL}$ using an RDBMS Technical Report, 2009. http://www.informatik.uni-bremen.de/∼clu/papers/

[Patel *et al.*, 2007] C. Patel, J. J. Cimino, J. Dolby, A. Fokoue, A. Kalyanpur, A. Kershenbaum, L. Ma, E. Schonberg, and K. Srinivas. Matching patient records to clinical trials using ontologies. In *Proc. of ISWC07*, volume 4825 of *LNCS*, pages 816–829. Springer, 2007.

[Rosati, 2007] Riccardo Rosati. On conjunctive query answering in $\mathcal{EL}$. In *Proc. of DL07*, volume 250 of *CEUR-WS*, 2007.

[Schaerf, 1993] A. Schaerf. On the complexity of the instance checking problem in concept languages with existential quantification. *J. of Intelligent Information Systems*, 2:265–278, 1993.

[Sioutos *et al.*, 2006] N. Sioutos, S. de Coronado, M.W. Haber, F.W. Hartel, W.L. Shaiu, and L.W. Wright. NCI thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. of Biomedical Informatics*, 40(1):30–43, 2006.

[Wu *et al.*, 2008] Z. Wu, G. Eadon, S. Das, E. Chong, V. Kolovski, M. Annamalai, and J. Srinivasan. Implementing an inference engine for RDFS/OWL constructs and user-defined rules in oracle. In *Proc. of ICDE08*, pages 1239–1248. IEEE, 2008.

A Proofs for Section 3

We start with a technical lemma stating two fundamental properties of canonical models.

Lemma 8. *Let $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ be a KB. For all concepts C, $a \in \mathsf{Ind}(\mathcal{A})$, and $x_{D,E} \in \mathsf{NI}_{\mathsf{aux}}$, we have*

1. *$a \in C^{\mathcal{I}_\mathcal{K}}$ iff $\mathcal{K} \models C(a)$;*
2. *$x_{D,E} \in C^{\mathcal{I}_\mathcal{K}}$ iff $\mathcal{K} \models D \sqcap E \sqsubseteq C$.*

Proof. We prove Points 1 and 2 simultaneously by induction on the structure of C. The induction start is trivial by definition of $\mathcal{I}_\mathcal{K}$. In the induction step, the case $C = F \sqcap F'$ is easy using the semantics and induction hyopthesis. Hence we concentrate on the case $C = \exists r.F$.

"$\Rightarrow$". (1) If $a \in C^{\mathcal{I}_\mathcal{K}}$, then there is a $d \in \Delta^{\mathcal{I}_\mathcal{K}}$ with $(a,d) \in r^{\mathcal{I}_\mathcal{K}}$ and $d \in F^{\mathcal{I}_\mathcal{K}}$. First assume that $d = b \in \mathsf{Ind}(\mathcal{A})$. By IH, $\mathcal{K} \models F(b)$. Since $(a,b) \in r^{\mathcal{I}_\mathcal{K}}$, we have $s(a,b) \in \mathcal{A}$ for some s with $\mathcal{K} \models s \sqsubseteq r$. It follows that $\mathcal{K} \models \exists r.F(a)$. Now assume that $d = x_{C',D'} \in \mathsf{NI}_{\mathsf{aux}}$. By IH, $\mathcal{K} \models C' \sqcap D' \sqsubseteq F$. Since $(a, x_{C',D'}) \in r^{\mathcal{I}_\mathcal{K}}$, there is a role name s such that $\mathcal{K} \models \exists s.D$, $\mathsf{ran}_\mathcal{T}(s) = C$, and $\mathcal{K} \models s \sqsubseteq r$. Thus, $\mathcal{K} \models \exists r.F(a)$.

(2) If $x_{D,E} \in C^{\mathcal{I}_\mathcal{K}}$, then there is an $x_{C',D'} \in \Delta^{\mathcal{I}_\mathcal{K}}$ with $(x_{D,E}, x_{C',D'}) \in r^{\mathcal{I}_\mathcal{K}}$ and $x_{C',D'} \in F^{\mathcal{I}_\mathcal{K}}$. By IH, $\mathcal{K} \models C' \sqcap D' \sqsubseteq F$. Since $(x_{D,E}, x_{C',D'}) \in r^{\mathcal{I}_\mathcal{K}}$, there is a role name s such that $\mathcal{K} \models D \sqcap E \sqsubseteq \exists r.D'$, $\mathsf{ran}_\mathcal{T}(s) = C'$, and $\mathcal{K} \models s \sqsubseteq r$. Thus, $\mathcal{K} \models D \sqcap E \sqsubseteq \exists r.F$.

"$\Leftarrow$". (1) If $\mathcal{K} \models C(a)$, then $(a, x_{C',F}) \in r^{\mathcal{I}_\mathcal{K}}$, where $\mathsf{ran}_\mathcal{T}(r) = C'$. We have $\mathcal{K} \models C' \sqcap F \sqsubseteq F$, and thus IH yields $x_{C',F} \in F^{\mathcal{I}_\mathcal{K}}$. By the semantics, $a \in C^{\mathcal{I}_\mathcal{K}}$. Point (2) can be treated analogously. ❏

We now prove Proposition 1 from Section 3.

Proposition 1. If $\mathcal{K}$ is consistent, then $\mathcal{I}_\mathcal{K}$ is a model of $\mathcal{K}$.

Proof. Let $\mathcal{K} = (\mathcal{T}, \mathcal{A})$. By definition, $\mathcal{I}_\mathcal{K}$ is a model of $\mathcal{A}$ and all role inclusions in $\mathcal{T}$. Let $\mathsf{ran}(r) \sqsubseteq C \in \mathcal{T}$ and $(d,e) \in r^{\mathcal{I}_\mathcal{K}}$. If $e = a \in \mathsf{Ind}(\mathcal{A})$, then $d = b \in \mathsf{Ind}(\mathcal{A})$ and $(b,a) \in \mathcal{A}$ by definition of $\mathcal{I}_\mathcal{K}$. It follows that $\mathcal{K} \models C(a)$ and thus the Lemma 8 yields $a \in C^{\mathcal{I}_\mathcal{K}}$ as required. The case $e = x_{C',D'} \in \mathsf{NI}_{\mathsf{aux}}$ is similar. Finally, let $C \sqsubseteq D \in \mathcal{T}$ and $d \in C^{\mathcal{I}_\mathcal{K}}$. If $e = a \in \mathsf{Ind}(\mathcal{A})$, then Lemma 8 yields $\mathcal{K} \models C(a)$ and thus $\mathcal{K} \models D(a)$ and we can apply Lemma 8 once more to derive $d \in D^{\mathcal{I}_\mathcal{K}}$. The case $e = x_{C',D'} \in \mathsf{NI}_{\mathsf{aux}}$ is again similar. ❏

The proof of the following lemma is standard, and thus omitted.

Lemma 9. *Let $\mathcal{R}$ be a set of role inclusions, $\mathcal{A}$ an ABox, $\mathcal{I}$ an interpretation, and $\mathcal{J}$ the $(\mathcal{A},\mathcal{R})$-unraveling of $\mathcal{I}$. Then we have $d_0 r_1 d_1 \cdots r_n d_n \in C^\mathcal{J}$ iff $d_n \in C^\mathcal{I}$ for all concepts C and $d_0 r_1 d_1 \cdots r_n d_n \in \Delta^\mathcal{J}$.*

Using Lemma 9 and Proposition 1, it is straightfoward to establish the following.

Proposition 10. *If $\mathcal{K}$ is consistent, then $\mathcal{U}_\mathcal{K}$ is a model of $\mathcal{K}$.*

Proposition 4. If $\mathcal{K}$ is consistent, then for all k-ary conjunctive queries q and all $a_1, \ldots, a_k \in \mathsf{Ind}(\mathcal{A})$, we have $(a_1, \ldots, a_k) \in \mathsf{cert}(q, \mathcal{K})$ iff $\mathcal{U}_\mathcal{K} \models q[a_1, \ldots, a_k]$.

Proof. The "$\Rightarrow$" direction is trivial by Proposition 10. Hence, we concentrate on "$\Leftarrow$". Assume that $\mathcal{U}_\mathcal{K} \models^\pi q[a_1, \ldots, a_k]$ and let $\mathcal{I}$ be a model of $\mathcal{K}$. For each $d \in \Delta^{\mathcal{U}_\mathcal{K}}$, we use $\mathsf{dep}(d)$ to denote the length of the shortest sequence $d_0, \ldots, d_n$ such that $d_0 \in \mathsf{Ind}(\mathcal{A})^\mathcal{I}$, $(d_i, d_{i+1}) \in \bigcup_{r \in \mathsf{N_R}} r^\mathcal{I}$ for all $i < n$, and $d_n = d$. We define a mapping $\delta : \Delta^{\mathcal{U}_\mathcal{K}} \to \Delta^\mathcal{I}$ such that

(a) $\delta(a) = a^\mathcal{I}$ for all $a \in \mathsf{Ind}(\mathcal{A})$;

(b) $d \in C^{\mathcal{U}_\mathcal{K}}$ implies $\delta(d) \in C^\mathcal{I}$ for all $d \in \Delta^{\mathcal{U}_\mathcal{K}}$ and concepts C;

(c) $(d,d') \in r^{\mathcal{U}_\mathcal{K}}$ implies $(\delta(d), \delta(d')) \in r^\mathcal{I}$ for all $d, d' \in \Delta^{\mathcal{U}_\mathcal{K}}$ and $r \in \mathsf{N_R}$.

The definition of $\delta(d)$ is by induction on $\mathsf{dep}(d)$. For the case $\mathsf{dep}(d) = 1$, $\delta(d)$ is dictated by (a). Let $d \in \Delta^{\mathcal{U}_\mathcal{K}}$ with $\mathsf{dep}(d) = n > 1$, i.e., $d = d_0 r_1 d_1 \cdots r_n d_n$. By definition of $\mathcal{U}_\mathcal{K}$, $(d_{n-1}, d_n) \in r_n^{\mathcal{I}_\mathcal{K}}$, $d_n = x_{C,D}$ for some C, D, and either (i) $d_{n-1} = a \in \mathsf{Ind}(\mathcal{A})$ or (ii) $d_{n-1} = x_{C',D'}$ for some C', D'. Let $d' = d_0 r_1 d_1 \cdots r_{n-1} d_{n-1}$.

In Case (i), the definition of $\mathcal{I}_\mathcal{K}$ implies that there is a role name s with $\mathcal{K} \models \exists s.D(a)$, $\mathsf{ran}_\mathcal{T}(s) = C$, and $\mathcal{K} \models s \sqsubseteq r_n$. Thus, (a) implies that $\delta(d') \in (\exists s.D)^\mathcal{I}$ and we set $\delta(d_n)$ to some $e \in \Delta^\mathcal{I}$ with $(d_{n-1}, e) \in s^\mathcal{I}$ and $e \in (C \sqcap D)^\mathcal{I}$. In Case (ii), the definition of $\mathcal{I}_\mathcal{K}$ implies that there is a role name s with $\mathcal{K} \models C' \sqcap D' \sqsubseteq \exists r.D$, $\mathsf{ran}_\mathcal{T}(s) = C$, and $\mathcal{K} \models s \sqsubseteq r$. By Lemmas 8 and 9 and by (b), we have $\delta(d') \in (\exists r.D)^\mathcal{I}$ and set $\delta(d_n)$ to some $e \in \Delta^\mathcal{I}$ with $(d_{n-1}, e) \in s^\mathcal{I}$ and $e \in (C \sqcap D)^\mathcal{I}$.

Note that δ satisfies (a) and (c) by construction. To see that (b) is satisfied, assume $d = d_0 r_1 d_1 \cdots r_n d_n \in C^{\mathcal{U}_\mathcal{K}}$ with $d_n = x_{D,E}$. By Lemmas 9 and 8, $\mathcal{K} \models D \sqcap E \sqsubseteq C$. By construction of δ, $\delta(d) \in (D \sqcap E)^\mathcal{I}$. Thus, $\delta(d) \in C^\mathcal{I}$.

It remains to verify that the composition π' of π with δ (i.e., $\pi'(t) = \delta(\pi(t))$ for all $t \in \mathsf{term}(q)$) satisfies $\mathcal{I} \models^{\pi'} q[a_1, \ldots, a_k]$, which is straightforward using (a)-(c). ❏

B Proof of Theorem 5

To complete the proof of Theorem 5 given in Section 4, it remains to show that for all $\mathcal{ELH}_\bot^{dr}$-KBs $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ with $\mathcal{R}$ the set of role inclusions in $\mathcal{T}$ and all $a_1, \ldots, a_k \in \mathsf{Ind}(\mathcal{A})$, we have $\mathcal{I}_\mathcal{K}^r \models q_\mathcal{R}^*[a_1, \ldots, a_k]$ iff $\mathcal{U}_\mathcal{K} \models q[a_1, \ldots, a_k]$, where $q_\mathcal{R}^*$ is the rewriting of q defined in Section 4. In the following, we prove a slightly stronger result.

Let $\mathcal{R}$ be a set of role inclusions and $\mathcal{A}$ an ABox. Let $\mathcal{I}$ be a model in which $\mathsf{Aux}^\mathcal{I} = \Delta^\mathcal{I} \setminus \mathsf{Ind}(\mathcal{A})^\mathcal{I}$. $\mathcal{I}$ is called *$\mathcal{A}$-connected* if every $d \in \Delta^\mathcal{I}$ equals $\mathsf{tail}(p)$ for some $p \in \mathsf{paths}_\mathcal{A}(\mathcal{I})$. It is called *split* if $d \in \mathsf{Aux}^\mathcal{I}$ and $(d,d') \in r^\mathcal{I}$ imply $d' \in \mathsf{Aux}^\mathcal{I}$, for all $r \in \mathsf{N_R}$ and $d, d' \in \Delta^\mathcal{I}$. Clearly $\mathcal{I}_\mathcal{K}^r$ is $\mathcal{A}$-connected and split. Thus, Theorem 5 follows from the following result.

Theorem 11. *Let $\mathcal{I}$ be split and $\mathcal{A}$-connected, and let $\mathcal{I}'$ be the $(\mathcal{A},\mathcal{R})$-unraveling of $\mathcal{I}$. Let q be a k-ary conjunctive query. Then the following holds for all $a_1, \ldots, a_k \in \mathsf{Ind}(\mathcal{A})$:*

$$\mathcal{I} \models q_\mathcal{R}^*[a_1, \ldots, a_k] \text{ iff } \mathcal{I}' \models q[a_1, \ldots, a_k].$$

Proof. Assume $q = \exists \vec{u}.\psi(\vec{u},\vec{v})$. Recall that $q^*_{\mathcal{R}}$ is defined as $\exists \vec{u}.(\psi \wedge \varphi_1 \wedge \varphi_2 \wedge \varphi_3)$, where φ_1, φ_2, and φ_3 are quantifier-free. Recall that $\mathsf{Aux}^{\mathcal{I}'} = \Delta^{\mathcal{I}'} \setminus \mathsf{Ind}(\mathcal{A})^{\mathcal{I}'}$.

($\Leftarrow$) Let π be an $(a_1,\dots,a_k)$-match for $\mathcal{I}'$ and q. Define a mapping $\tau : \mathsf{term}(q) \to \Delta^{\mathcal{I}}$ by setting $\tau(t) := \mathsf{tail}(\pi(t))$ for all $t \in \mathsf{term}(q)$. By definition of τ and $\mathcal{I}'$, $\mathcal{I} \models^{\tau} \psi$ and so τ is an $(a_1,\dots,a_k)$-match for $\mathcal{I}$ and q. It thus remains to show that $\mathcal{I} \models^{\tau} \varphi_1 \wedge \varphi_2 \wedge \varphi_3$. To this end, we first show the following:

Claim 1. Let $s', t' \in \mathsf{term}(q)$ with $s' \sim_q t'$ and $\pi(s') \in \mathsf{Aux}^{\mathcal{I}'}$. Then

(a) $\pi(s') = \pi(t')$;

(b) if $r_1(s,s'), r_2(t,t') \in q$, then $\pi(s) = \pi(t)$.

We start with the proof of Point (a). By definition, $\sim_q$ can be generated by starting with $\mathsf{id}_q = \{(t,t) \mid t \in \mathsf{term}(\mathsf{q})\}$ and then exhaustively applying $(*)$ from Section 4 as a rule and the following rule:

(tc) if $t \sim_q s$ and $s \sim_q t'$, then $t \sim_q t'$.

We prove Point (a) by induction on the number of rule application. The start is trivial. For the step, we distinguish between the two rules:

Rule $(*)$. Let $r_1(s,s'), r_2(t,t') \in q$ and $s' \sim_q t'$. Then $(*)$ adds (s,t) to $\sim_q$. Assume that $\pi(s) \in \mathsf{Aux}^{\mathcal{I}'}$. By construction of $\mathcal{I}'$ and since $(\pi(s),\pi(s')) \in r_1^{\mathcal{I}'}$, we get that $\pi(s') \in \mathsf{Aux}^{\mathcal{I}'}$. By IH, $\pi(s') = \pi(t')$. By construction of $\mathcal{I}'$ and since $(\pi(s),\pi(s')) \in r_1^{\mathcal{I}'}$, $(\pi(t'),\pi(s')) \in r_2^{\mathcal{I}'}$, and $\pi(s') \in \mathsf{Aux}^{\mathcal{I}'}$, we have $\pi(s) = \pi(t)$.

Rule (tc). Let $t \sim_q s$ and $s \sim_q t'$. Then (tc) adds (t,t') to $\sim_q$. Assume that $\pi(t) \in \mathsf{Aux}^{\mathcal{I}'}$. By IH, $\pi(t) = \pi(s)$ and thus $\pi(s) \in \mathsf{Aux}^{\mathcal{I}'}$. Again by IH, $\pi(s) = \pi(t')$ and we obtain $\pi(t) = \pi(t')$.

We come to Point (b). Let $r_1(s,s'), r_2(t,t') \in q$ and $s' \sim_q t'$ and assume that $\pi(s') \in \mathsf{Aux}^{\mathcal{I}'}$. By Point (a), $\pi(s') = \pi(t')$. Hence, by construction of $\mathcal{I}'$ and since $\pi(s') \in \mathsf{Aux}^{\mathcal{I}'}$, $\pi(s) = \pi(t)$. This finishes the proof of Claim 1.

We now show that $\mathcal{I} \models^{\tau} \varphi_1$, i.e., $\tau(v) \in \mathsf{Ind}(\mathcal{A})^{\mathcal{I}}$ for all $v \in \mathsf{avar}(q) \cup \mathsf{Fork}_{\neq} \cup \mathsf{Cyc}$. By definition of $(a_1,\dots,a_k)$-match, we have $\pi(t) \in \mathsf{Ind}(\mathcal{A})^{\mathcal{I}'}$ for all $t \in \mathsf{avar}(q)$. By definition of τ and construction of $\mathcal{I}'$, it follows that $\tau(t) = \pi(t) \in \mathsf{Ind}(\mathcal{A})^{\mathcal{I}'} = \mathsf{Ind}(\mathcal{A})^{\mathcal{I}}$. Now assume that $v \in \mathsf{Fork}_{\neq}$ and that, contrary to what has to be shown, $\tau(v) \in \mathsf{Aux}^{\mathcal{I}}$. Then there is no implicant for $\mathsf{in}([v])$. For each $r \in \mathsf{in}([v])$, there is an atom $r(s_r,t_r) \in q$ with $t_r \sim_q v$. Since $\tau(v) \in \mathsf{Aux}^{\mathcal{I}}$, we have $\pi(v) \in \mathsf{Aux}^{\mathcal{I}'}$. Thus, by Point (a) of Claim 1, $\pi(v) = \pi(t_r)$ for all $r \in \mathsf{in}([v])$. Thus $(\pi(s_r),\pi(v)) \in r^{\mathcal{I}'}$ for all $r \in \mathsf{in}([v])$. By construction of $\mathcal{I}'$ and since $\pi(v) \in \mathsf{Aux}^{\mathcal{I}'}$, we have that $\pi(s_r) = \pi(s_{r'})$ for all $r, r' \in \mathsf{in}([v])$ and there is an implicant for $\mathsf{in}([v])$, which is a contradiction.

Now assume $v \in \mathsf{Cyc}$ and that, contrary to what has to be shown, $\tau(v) \in \mathsf{Aux}^{\mathcal{I}}$. Then there are

$$r_0(t_0,t_0'),\dots,r_m(t_m,t_m'),\dots,r_n(t_n,t_n') \in q,\ n,m \geq 0,$$

with $v \sim_q t_j$ for some $j \leq n$, $t_i' \sim_q t_{i+1}$ for all $i < n$, and $t_n' \sim_q t_m$. Since $\tau(v) \in \mathsf{Aux}^{\mathcal{I}}$ and by Point (a) of Claim 1, $\pi(t_j) \in \mathsf{Aux}^{\mathcal{I}'}$. Since $r_j(t_j,t_j') \in q$, the construction of unravelings yields that $\pi(t_j') = \pi(t_j) \cdot rd$ for some $d \in \Delta^{\mathcal{I}}$. In particular, $\pi(t_j')$ is auxiliary. By Point (a) of Claim 1, $\pi(t_j') = \pi(t_{j+1})$. We can repeat this argument ad infinitum, setting $t_i = t_i \mod n+1$ and $t_i' = t_i' \mod n+1$ for all $i > n$. In each step, the length of the path $\pi(t_{j+\ell})$ increases. This contradicts the fact that $\pi(t_{n+j}) = \pi(t_j)$ (since actually $t_{n+j} = t_j$). We have thus shown that $\mathcal{I} \models^{\tau} \varphi_1$.

We now show that $\mathcal{I} \models^{\tau} \varphi_2$, i.e., for all $(\{t_1,\dots,t_k\},\zeta) \in \mathsf{Fork}_{=}$, $\tau(t_\zeta) \in \mathsf{Aux}^{\mathcal{I}}$ implies $\tau(t_1) = \cdots = \tau(t_k)$. Thus, let $(\{t_1,\dots,t_k\},\zeta) \in \mathsf{Fork}_{=}$ and assume that $\tau(t_\zeta) \in \mathsf{Aux}^{\mathcal{I}}$. Then $\pi(t_\zeta) \in \mathsf{Aux}^{\mathcal{I}'}$ and there are terms $t_1',\dots,t_k' \in \zeta$ and role names $r_1,\dots,r_k$ such that $r_i(t_i,t_i') \in q$ for $1 \leq i \leq k$. Since $\pi(t_\zeta) \in \mathsf{Aux}^{\mathcal{I}'}$ and by Point (b) of Claim 1, $\pi(t_1) = \cdots = \pi(t_k)$, and thus $\tau(t_1) = \cdots = \tau(t_k)$.

Finally, we show that $\mathcal{I} \models^{\tau} \varphi_3$, i.e., for all $(I,\zeta) \in \mathsf{Fork}_{\mathcal{H}}$, $\tau(t_\zeta) \in \mathsf{Aux}^{\mathcal{I}}$ implies $(\tau(t_\zeta^{\mathsf{pre}}),\tau(t_\zeta)) \in r^{\mathcal{I}}$ for some $r \in I$. Thus, let $(I,\zeta) \in \mathsf{Fork}_{\mathcal{H}}$ and assume that $\tau(t_\zeta) \in \mathsf{Aux}^{\mathcal{I}}$. Then $\mathsf{pre}(\zeta) \neq \emptyset$, i.e., t_ζ^{pre} is defined and the set $\Gamma := \{r \in \mathsf{N_R} \mid (\tau(t_\zeta^{\mathsf{pre}}),\tau(t_\zeta)) \in r^{\mathcal{I}}\}$ is non-empty. By construction of $\mathcal{I}'$ and since $\tau(t_\zeta) \in \mathsf{Aux}^{\mathcal{I}}$, there is an $r \in \Gamma$ that is an implicant for Γ. Since $\tau(t_\zeta) \in \mathsf{Aux}^{\mathcal{I}}$, we have $\pi(t_\zeta) \in \mathsf{Aux}^{\mathcal{I}}$. Thus, Claim 1 and the definition of τ yields

- $\tau(t) = \tau(t_\zeta)$ for all $t \in \zeta$ and
- $\tau(t) = \tau(t_\zeta^{\mathsf{pre}})$ for all $t \in \mathsf{pre}(\zeta)$.

It follows that $\Psi := \{s \in \mathsf{N_R} \mid s(t,t') \in q \text{ for some } t \in \mathsf{pre}(\zeta) \text{ and } t' \in \zeta\} \subseteq \Gamma$ and thus r is an implicant for Ψ. By Assumption (v) from Section 2, there is even a prime implicant $\widehat{r} \in \Gamma$ for Ψ. We have $(\tau(t_\zeta^{\mathsf{pre}}),\tau(t_\zeta)) \in \widehat{r}^{\mathcal{I}}$ and $\widehat{r} \in I$.

($\Rightarrow$) Let π be an $(a_1,\dots,a_k)$-match for $\mathcal{I}$ and $q^*_{\mathcal{R}}$. We start with introducing some notation. The *degree* $d(\zeta)$ of an equivalence class ζ is the length $n \geq 0$ of a longest sequence (if it exists)

$$r_0(t_0,t_0'),\dots,r_n(t_n,t_n') \in q$$

such that $t_0 \in \zeta$ and $t_i' \sim_q t_{i+1}$ for all $i < n$. If no longest sequence exists, we set $d(\zeta) = \infty$.

Claim 2.

(a) If $\pi(t) \in \mathsf{Aux}^{\mathcal{I}}$, then $d([t]) < \infty$.

(b) If $s \sim_q t$ and $\pi(s) \in \mathsf{Aux}^{\mathcal{I}}$, then

  (i) $\pi(s) = \pi(t)$;

  (ii) if $r_1(s',s), r_2(t',t) \in q$, then $\pi(s') = \pi(t')$.

We start with (a). Assume to contrary of what has to be shown that there is a t_0 with $\pi(t_0) \in \mathsf{Aux}^{\mathcal{I}}$ and an infinite sequence

$$r_0(t_0,t_0'), r_1(t_1,t_1'),\dots$$

with $t'_i \sim_q t_{i+1}$ for all $i \geq 0$. By definition of $(a_1, \ldots, a_k)$-match, $\pi(t_0) \in \mathsf{Aux}^{\mathcal{I}}$ implies that $t_0 \in \mathsf{qvar}(q)$. As q is finite, there exist m, n with $0 \leq m \leq n$ such that $t'_n = t'_m$. It follows that $t_0 \in \mathsf{Cyc}$. Hence φ_1 contains the conjunct $\neg\mathsf{Aux}(t_0)$ and we have derived a contradiction to $\pi(t_0) \in \mathsf{Aux}^{\mathcal{I}}$.

Now for (b). Because of (a), Point (i) of (b) can be proved by induction on $n := d([s]) = d([t])$. For the induction start, let $s \sim_q t$ with $\pi(s) \in \mathsf{Aux}^{\mathcal{I}}$ and $d([s']) = 0$. By definition of $\sim_q$, we have $[s] = \{s\}$ and thus $s = t$. Therefore, $\pi(s) = \pi(t)$ trivially holds. For the induction step, define

$$\begin{aligned}
\sim_q^{(0)} &:= \{(t,t) \mid t \in \mathsf{term}(q)\} \\
\sim_q^{(i+1)} &:= \sim_q^{(i)} \cup \\
&\quad \{(s,t) \mid \text{there is } s' \text{ with } s \sim_q^{(i)} s' \text{ and } s' \sim_q^{(i)} t\} \cup \\
&\quad \{(s,t) \mid \text{there are } r_1(s,s'), r_2(t,t') \in q \text{ with } s' \sim_q^{(i-1)} t'\}
\end{aligned}$$

for all $i \geq 0$. It is not hard to see that $\sim_q = \bigcup_{i \geq 0} \sim_q^{(i)}$. We show by induction on i that if $s \sim_q^{(i)} t$, $d([s]) = n$, and $\pi(s) \in \mathsf{Aux}^{\mathcal{I}}$, then $\pi(s) = \pi(t)$. The induction start is trivial since $s \sim_q^{(0)} t$ implies $s = t$. For the induction step, we distinguish two cases:

- There is s' with $s \sim_q^{(i)} s'$ and $s' \sim_q^{(i)} t$.
  By (inner) IH, $\pi(s) = \pi(s')$ and thus $\pi(s') \in \mathsf{Aux}^{\mathcal{I}}$. Since $s \sim_q^{(i)} s'$, we have $[s] = [s']$, and thus $d([s']) = n$. We can thus apply (inner IH) once more to derive $\pi(s') = \pi(t)$, thus $\pi(s) = \pi(t)$.
- There are $r_1(s,s'), r_2(t,t') \in q$ such that $s' \sim_q^{(i-1)} t'$.
  As $\mathcal{I}$ is split, $r(s,s') \in q$ and $\pi(s) \in \mathsf{Aux}^{\mathcal{I}}$ entails $\pi(s') \in \mathsf{Aux}^{\mathcal{I}}$. By definition of depth, $d([s']) < d([s])$. We can thus apply (outer) IH to obtain $\pi(s') = \pi(t_{[s']})$. Hence, $\pi(t_{[s']}) \in \mathsf{Aux}^{\mathcal{I}}$. Thus, from the conjunct φ_2 of $q^*_{\mathcal{R}}$, we obtain $\pi(s) = \pi(t)$.

Now for Point (ii). Assume $\pi(s) \in \mathsf{Aux}^{\mathcal{I}}$, $r_1(s',s), r_2(t',t) \in q$, and $s \sim_q t$. By Point (i), $\pi(s) = \pi(t_{[s]})$. Hence, by the conjunct φ_2 of $q^*_{\mathcal{R}}$, $\pi(s') = \pi(t')$. This finishes the proof of Claim 2.

Let $\sim_\pi$ be the transitive closure of

$$\begin{aligned}
&\{(t,t) \mid t \in \mathsf{term}(q) \cup \\
&\{(s,t) \in \mathsf{term}(q)^2 \mid s \sim_q t, \pi(s), \pi(t) \in \mathsf{Aux}^{\mathcal{I}}\} \cup \\
&\{(s,t) \in \mathsf{term}(q)^2 \mid \exists\, r_1(s,s'), r_2(t,t') \in q : \\
&\qquad\qquad \pi(s') \in \mathsf{Aux}^{\mathcal{I}} \wedge s' \sim_q t'\}.
\end{aligned}$$

By Claim 2, we have

$(*)$ $\pi(s) = \pi(t)$ whenever $s \sim_\pi t$.

Note that $\sim_\pi$ is an equivalence relation because it is, by Claim 2, the transitive closure of a symmetric relation.

Now let the query q' be obtained from q by identifying all terms $t, t' \in \mathsf{term}(q)$ such that $t \sim_\pi t'$. More precisely, choose from each $\sim_\pi$-equivalence class ξ a fixed term $t_\xi \in \xi$ and replace each occurrence of an element of ξ in q by t_ξ. By $(*)$, π is a match for $\mathcal{I}$ and the resulting query q'. Next, we show the following:

(I) if $v \in \mathsf{qvar}(q')$ with $\pi(v) \in \mathsf{Aux}^{\mathcal{I}}$, then there is at most one $t \in \mathsf{term}(q')$ such that $r(t,v) \in q'$, for some $r \in \mathsf{N_R}$;

(II) if $v \in \mathsf{qvar}(q')$ with $\pi(v) \in \mathsf{Aux}^{\mathcal{I}}$ and $t \in \mathsf{term}(q')$ such that $\Gamma = \{r \mid r(t,v) \in q'\} \neq \emptyset$, then there is an implicant s for Γ with $(\pi(t), \pi(v)) \in s^{\mathcal{I}}$;

(III) if $q' \supseteq \{r_0(t_0,t_1), \ldots, r_{n-1}(t_{n-1}, t_n)\}$ with $t_0 = t_n$, then $\pi(t_i) \notin \mathsf{Aux}^{\mathcal{I}}$, for all $i \leq n$.

First for (I). Let $\pi(v) \in \mathsf{Aux}^{\mathcal{I}}$, and let $r_1(t_1,v), r_2(t_2,v) \in q'$. Then there are $r_1(s_1,s'_1), r_2(s_2,s'_2) \in q$ such that $s_1 \sim_\pi t$, $s_2 \sim_\pi t'$, and $s'_1 \sim_\pi v \sim_\pi s'_2$. By $(*)$, $\pi(s'_1) = \pi(v)$, and thus $\pi(s'_1) \in \mathsf{Aux}^{\mathcal{I}}$. By definition of $\sim_\pi$, $s'_1 \sim_\pi s'_2$ implies $s'_1 \sim_q s'_2$. Summing up, we thus have $t_1 \sim_\pi t_2$. Since both t_1 and t_2 occur in q', we have $t_1 = t_2$.

Now for (II). Let $\pi(v) \in \mathsf{Aux}^{\mathcal{I}}$ and $\Gamma \neq \emptyset$. Due to the use of $\mathsf{Fork}_{\neq}$ in φ_1 and since $\pi(v) \in \mathsf{Aux}^{\mathcal{I}}$, there is an implicant for $\mathsf{in}([v])$. By φ_3, there thus is an implicant s for $\mathsf{in}([v])$ with $(\pi(t^{\mathsf{pre}}_{[v]}), \pi(t_{[v]})) \in s^{\mathcal{I}}$. Since $\pi(v) \in \mathsf{Aux}^{\mathcal{I}}$ we have $t^{\mathsf{pre}}_{[v]} \sim_\pi t$ and $t_{[v]} \sim_\pi v$. By $(*)$, $\pi(t^{\mathsf{pre}}_{[v]}) = \pi(t)$ and $\pi(t_{[v]}) = \pi(v)$, thus $(\pi(t), \pi(v)) \in s^{\mathcal{I}}$. Since $\Gamma \subseteq \mathsf{in}([v])$, s is the required implicant for Γ.

For (III), let $q' \supseteq \{r_0(t_0,t_1), \ldots, r_{n-1}(t_{n-1}, t_n)\}$ with $t_0 = t_n$. Then there are $r_0(s_0,s'_0), \ldots, r_{n-1}(s_{n-1}, s'_{n-1}) \in q$ with $s_i \sim_\pi t_i$ and $s'_i \sim_\pi t_{i+1} \bmod n$ for all $i < n$. It follows that $s'_i \sim_\pi s_{i+1} \bmod n$ for all $i < n$. Assume now, contrary to what has to be shown, that $\pi(t_i) \in \mathsf{Aux}^{\mathcal{I}}$ for some $i < n$. Since $s_i \sim_\pi t_i$, $(*)$ yields $\pi(s_i) = \pi(t_i)$. Thus $\pi(s_i) \in \mathsf{Aux}^{\mathcal{I}}$, which implies $s_i \in \mathsf{qvar}(q)$ by definition of $(a_1, \ldots, a_k)$-matches. Together with $\sim_\pi \subseteq \sim_q$, $s_i \in \mathsf{qvar}(q)$ implies $s_i \in \mathsf{Cyc}$. Thus, $\neg\mathsf{Aux}(s_i)$ is a conjunct of φ_1 and $\pi(s_i) \notin \mathsf{Aux}^{\mathcal{I}}$, which is a contradiction. This finishes the proof of (I)-(III).

We inductively define a mapping $\tau : \mathsf{term}(q') \to \Delta^{\mathcal{I}'}$ such that $\mathsf{tail}(\tau(t)) = \pi(t)$ for all $t \in \mathsf{term}(q')$. For the induction start, we distinguish two cases:

- for all $t \in \mathsf{term}(q')$ with $\pi(t) \notin \mathsf{Aux}^{\mathcal{I}}$, set $\tau(t) := \pi(t)$. Observe that this defines $\tau(t)$ for all $t \in \mathsf{avar}(q') \cup (\mathsf{term}(q') \cap \mathsf{N_I})$.
- for all $v \in \mathsf{qvar}(q')$ with $\pi(v) \in \mathsf{Aux}^{\mathcal{I}}$ and such that there is no atom $r(t,v) \in q$, do the following. By definition of $\mathcal{I}'$ and because each $d \in \Delta^{\mathcal{I}}$ is reachable from an element of $\mathsf{Ind}(\mathcal{A})^{\mathcal{I}}$, there is a sequence $d_0, \ldots, d_n \in \Delta^{\mathcal{I}}$ and a sequence $r_0, \ldots, r_{n-1}$ of role names such that $d_0 \in \mathsf{Ind}(\mathcal{A})^{\mathcal{I}}$, $d_n = \pi(v)$, and $(d_i, d_{i+1}) \in r^{\mathcal{I}}_{i+1}$ for all $i < n$. Set $\tau(v) := d_0 r_0 d_1 \cdots r_{n-1} d_n \in \Delta^{\mathcal{I}'}$.

For the induction step, proceed as follows. If $\tau(v)$ is undefined and there exists $r(t,v) \in q'$ with $\tau(t)$ defined, then (II) yields an implicant s for $\Gamma = \{r \mid r(t,v) \in q'\} \neq \emptyset$ with $(\pi(t), \pi(v)) \in s^{\mathcal{I}}$. Set $\tau(v) := \tau(t) \cdot s\pi(v)$. Since $\mathsf{tail}(\tau(t)) = \pi(t)$ and $(\pi(t), \pi(v)) \in s^{\mathcal{I}}$, we have $\tau(v) \in \Delta^{\mathcal{I}'}$.

By (I), the mapping τ is well-defined, i.e., the term t in the induction step is unique. By (III), τ is total, i.e., $\tau(t)$ is defined for all $t \in \mathsf{term}(q')$. To see this, suppose that $\tau(t)$ is undefined. Since $\tau(t)$ is not defined in the induction start, we have $\pi(v) \in \mathsf{Aux}^{\mathcal{I}}$ and there is an atom $r(s,t) \in q$. Since $\tau(t)$ is not defined in the induction step, $\tau(s)$ is undefined. We can repeat this argument ad infinitum. Since q' is finite, this means that there is a sequence $q' \supseteq \{r_0(s_0,s_1),\dots,r_{n-1}(s_{n-1},s_n)\}$ with $s_0 = s_n$ and $\pi(s_i) \in \mathsf{Aux}^{\mathcal{I}}$ for all $i \leq n$, in contradiction to (III).

The constructed τ is a match for $\mathcal{I}'$ and q'. It is immediate that $\mathcal{I}' \models^{\tau} A(t)$ for all $A(t) \in q'$ since $\mathsf{tail}(\tau(t)) = \pi(t)$ and $p \in A^{\mathcal{I}'}$ iff $\mathsf{tail}(p) \in A^{\mathcal{I}}$ for all $p \in \Delta^{\mathcal{I}'}$. Now let $r(t,t') \in q'$. If $\pi(t), \pi(t') \notin \mathsf{Aux}^{\mathcal{I}}$, then $\tau(t) = \pi(t), \tau(t') = \pi(t')$, and $(\pi(t),\pi(t')) \in r^{\mathcal{I}'}$. If $\pi(t') \in \mathsf{Aux}^{\mathcal{I}}$, then the construction of τ implies that $\tau(t') = \tau(t) \cdot s\pi(t')$ with $\mathcal{T} \models s \sqsubseteq r$. By definition of $\mathcal{I}'$, it follows that $(\tau(t),\tau(t')) \in r^{\mathcal{I}}$. The case that $\pi(t) \in \mathsf{Aux}^{\mathcal{I}}$ and $\pi(t') \notin \mathsf{Aux}^{\mathcal{I}}$ cannot occur since $\mathcal{I}$ is a split model.

Finally, we extend τ to a mapping from $\mathsf{term}(q)$ to $\Delta^{\mathcal{I}'}$ by setting $\tau(t) := \tau(t')$ if $t \in \mathsf{term}(q) \setminus \mathsf{term}(q')$ and $t \sim_\pi t'$. It is straightforward to verify that τ is a match for $\mathcal{I}'$ and q. Since $\tau(t) = \pi(t)$ if $\pi(t) \in \mathsf{Ind}(\mathcal{A})^{\mathcal{I}}$ for all $t \in \mathsf{term}(q')$, it is also clear that τ is an $(a_1,\dots,a_k)$-match. ❑

C Proof of Theorem 7

Theorem 7. Let $\mathcal{K}$ be a consistent literal $\mathcal{ELH}^{dr}_{\bot}$-KB. Then $\mathcal{K}'$ can be computed in polynomial time and $\mathcal{K} \models_e q[a_1,\dots,a_k]$ iff $\mathcal{K}' \models q'[a_1,\dots,a_k]$ for all literal CQs q and all $a_1,\dots,a_k \in \mathsf{Ind}(\mathcal{A})$.

Proof. We first show that $\mathcal{K}'$ can be computed in polynomial time. For $\mathcal{T}'$, this is obvious. For $\mathcal{A}'$, it suffices to show that given a literal KB $\mathcal{K} = (\mathcal{T},\mathcal{A})$ and a ground literal query $\neg A(a)$, we can decide in polynomial time whether $\mathcal{K} \models \neg A(a)$. To see that this is the case, set

$$L := \{\neg B(b) \mid \neg B(b) \in \mathcal{A}\}.$$

It is rather easy to see that

$$\mathcal{K} \models \neg A(a) \text{ iff } \mathcal{K}_0 \models \bigsqcup_{\neg B(b) \in L} B(b),$$

where $\mathcal{K}_0 = (\mathcal{T}, (\mathcal{A} \setminus L) \cup \{A(a)\})$. Since instance checking in $\mathcal{ELH}^{dr}_{\bot}$ is in PTIME, it now suffices to show that the latter consequence holds iff

$$\mathcal{K}_0 \models B(b) \text{ for some } \neg B(b) \in L.$$

To see this equivalence, assume that $\mathcal{K}_0 \not\models B(b)$ for all $\neg B(b) \in L$. The canonical model $\mathcal{I}_{\mathcal{K}_0}$ is a model of $\mathcal{K}_0$ and $\mathcal{I}_{\mathcal{K}_0} \models B(b)$ iff $\mathcal{K}_0 \models B(b)$, for all $\neg B(b)$ in L. Thus,

$$\mathcal{I}_{\mathcal{K}_0} \not\models \bigsqcup_{\neg B(b) \in L} B(b),$$

and, therefore, $\mathcal{K}_0 \not\models \bigsqcup_{\neg B(b) \in L} B(b)$.

We now prove that $\mathcal{K} \models_e q[a_1,\dots,a_k]$ iff $\mathcal{K}' \models q'[a_1,\dots,a_k]$ for all literal CQs q and all $a_1,\dots,a_k \in \mathsf{Ind}(\mathcal{A})$.

"$\Rightarrow$". Assume that $\mathcal{K}' \not\models q'[a_1,\dots,a_k]$. Then $\mathcal{U}_{\mathcal{K}'} \not\models q'[a_1,\dots,a_k]$. We first show that $\mathcal{U}_{\mathcal{K}'}$ is a model of $\mathcal{K}$: first, we can assume standard names since every $d \notin \mathsf{Ind}(\mathcal{A})^{\mathcal{U}_{\mathcal{K}'}}$ can be replaced with an $a_d \in \mathsf{N_I} \setminus \mathsf{Ind}(\mathcal{A})$. Since $\mathcal{U}_{\mathcal{K}'}$ is a model of $\mathcal{K}'$, $\mathcal{U}_{\mathcal{K}'}$ is a model of $\mathcal{T}$ and satisfies all positive atoms of $\mathcal{A}$. It thus remains to show that $\mathcal{U}_{\mathcal{K}'} \models \neg A(a)$ for every $\neg A(a) \in \mathcal{A}$. Since $\mathcal{K}$ is consistent and $\neg A(a) \in \mathcal{A}$, there is a model $\mathcal{I}$ of $\mathcal{K}$ with $a^{\mathcal{I}} \notin A^{\mathcal{I}}$. We can extend $\mathcal{I}$ to a model of $\mathcal{K}'$ by setting $\overline{B}^{\mathcal{I}} = \{b \mid \mathcal{K} \models \neg B(b)\}$ for all concept names B. Note that all $\top \sqsubseteq \overline{B} \in \mathcal{T}'$ are satisfied by construction of $\mathcal{K}'$ and since $\mathcal{K}$ is consistent. Thus $\mathcal{K}' \not\models A(a)$ and thus $\mathcal{U}_{\mathcal{K}'} \models \neg A(a)$ by definition of $\mathcal{U}_{\mathcal{K}'}$.

It thus remains to show that $\mathcal{U}_{\mathcal{K}'} \not\models_e q[a_1,\dots,a_k]$. Suppose to the contrary that $\mathcal{U}_{\mathcal{K}'} \models^{\pi}_e q[a_1,\dots,a_k]$ for some match π. Then $\mathcal{U}_{\mathcal{K}'} \models^{\pi} q'[a_1,\dots,a_k]$. Clearly, π satisfies all atoms in q' that are also in q. Thus let $\overline{A}(t)$ be an atom that is in q', but not in q. Then $\neg A(t)$ is in q and thus $\mathcal{K} \models \neg A(\pi(t))$. This implies that (i) $\mathcal{T} \models A \sqsubseteq \bot$ or (ii) $\pi(t) \in \mathsf{Ind}(\mathcal{A})$. To see this, suppose that neither (i) nor (ii) holds. By failure of (i), we can find a model $\mathcal{I}$ of $\mathcal{T}$ such that there is a $d \in A^{\mathcal{I}}$; since $\mathcal{K}$ is consistent, there is a model $\mathcal{J}$ of $\mathcal{K}$; let $\mathcal{I} \uplus \mathcal{J}$ be the union of two disjoint copies of $\mathcal{I}$ and $\mathcal{J}$ with elements renamed such that the standard names condition is satisfied, all individual names are interpreted as in $\mathcal{J}$, and the new name for d is $\pi(t)$ (this naming scheme is possible by failure of (ii)); then $\mathcal{I} \uplus \mathcal{J}$ is a model of $\mathcal{K}$ and $\pi(t) \in A^{\mathcal{I} \uplus \mathcal{J}}$, in contradiction to $\mathcal{K} \models \neg A(\pi(t))$. From (i) and (ii), it follows that $\top \sqsubseteq \overline{A} \in \mathcal{T}'$ or $\overline{A}(\pi(t)) \in \mathcal{A}'$. Since $\mathcal{U}_{\mathcal{K}'}$ is a model of $\mathcal{K}'$, we have $\pi(t) \in \overline{A}^{\mathcal{I}'}$ as required. We have thus established a contradiction to the fact that $\mathcal{U}_{\mathcal{K}'} \not\models q'[a_1,\dots,a_k]$.

"$\Leftarrow$". Assume that $\mathcal{I}$ is a model of $\mathcal{K}$ with $\mathcal{I} \not\models_e q[a_1,\dots,a_k]$. Convert $\mathcal{I}$ into a new interpretation $\mathcal{I}'$ by setting $\overline{A}^{\mathcal{I}} = \{a \in \mathsf{N_I} \mid \mathcal{K} \models \neg A(a)\}$ for all concept names A. It is easy to see that $\mathcal{I}'$ is a model of $\mathcal{K}'$. It thus remains to show that $\mathcal{I}' \not\models q'[a_1,\dots,a_k]$. Suppose to the contrary that $\mathcal{I}' \models^{\pi} q'[a_1,\dots,a_k]$ for some match π. Then $\mathcal{I} \models^{\pi}_e q[a_1,\dots,a_k]$: let $\neg A(t)$ be an atom that is in q, but not in q'. Then $\overline{A}(t)$ is in q'. Since π satisfies this atom, we have $\pi(t) \in \overline{A}^{\mathcal{I}'}$. By definition of $\mathcal{I}'$, $\mathcal{K} \models \neg A(\pi(t))$ as required. We have thus established a contradiction to the fact that $\mathcal{I} \not\models_e q[a_1,\dots,a_k]$. ❑