Query-based Comparison of Mappings in Ontology-based Data Access
Meghyn Bienvenu, Riccardo Rosati

To cite this version:

HAL Id: lirmm-01367812
https://hal-lirmm.ccsd.cnrs.fr/lirmm-01367812
Submitted on 16 Sep 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Query-based Comparison of Mappings in Ontology-based Data Access

Meghyn Bienvenu
LIRMM - CNRS, INRIA, & Université de Montpellier
Montpellier, France

Riccardo Rosati
DIAG
Sapienza Università di Roma
Rome, Italy

Abstract
An ontology-based data access (OBDA) system is composed of one or more data sources, an ontology that provides a conceptual view of the data, and declarative mappings that relate the data and ontology schemas. In order to debug and optimize such systems, it is important to be able to analyze and compare OBDA specifications. Recent work in this direction compared specifications using classical notions of equivalence and entailment, but an interesting alternative is to consider query-based notions, in which two specifications are deemed equivalent if they give the same answers to the considered query or class of queries for all possible data sources. In this paper, we define such query-based notions of entailment and equivalence of OBDA specifications and investigate the complexity of the resulting analysis tasks when the ontology is formulated in (fragments of) DL-LiteR.

1 Introduction
Ontology-based data access (OBDA) (Poggi et al. 2008) is a recent paradigm that proposes the use of an ontology as a conceptual, reconciled view of the information stored in a set of existing data sources. The connection between the ontology and the data sources is provided by declarative mappings that relate the elements of the ontology with the elements of the data sources. The ontology layer is the virtual interface used to access data, through queries formulated in the vocabulary of the ontology.

Due to the recent availability of techniques and systems for query processing in this setting (Calvanese et al. 2011; Rodriguez-Muro, Kontchakov, and Zakharyaschev 2013), the OBDA approach has begun to be experimented in real applications (see e.g. (Antonioli et al. 2013; Kharlamov et al. 2013; Giese et al. 2015)). In these projects, the construction, debugging and maintenance of the OBDA specification, consisting of the ontology, the schemas of the data sources, and the mapping, is a non-trivial task. In fact, the size and the complexity of the ontology and, especially, the mappings makes the management of such specifications a practical issue in these projects. Providing formal tools for supporting the above activities is therefore crucial for the successful deployment of OBDA solutions.

In addition, the OBDA specification plays a major role in query answering, since its form may affect the system performance in answering queries: different, yet semantically equivalent specifications may give rise to very different execution times for the same query. Thus, the study of notions of equivalence and formal comparison of OBDA specifications is also important for optimizing query processing in OBDA systems. Indeed, some systems already implement forms of optimization based on such transformations (see, e.g., (Rodriguez-Muro, Kontchakov, and Zakharyaschev 2013)).

Most of the work on OBDA thus far has focused on the query answering task, often in a simplified setting without mappings. Very little attention has been devoted to the formal analysis of OBDA specifications. The first approach that explicitly focuses on the formal analysis of OBDA specifications is (Lembo et al. 2014; 2015), whose aim is the discovery of redundancies and inconsistencies in mappings. That work utilizes the classical notions of logical equivalence and entailment to compare OBDA specifications. While it is very natural to resort to such classical notions, an appealing alternative in many applications is to adopt query-based notions of equivalence and entailment, in which two specifications are compared with respect to a given query or a given class of queries, and are deemed equivalent if they give the same answers to the considered queries for all possible data sources. This idea has already been investigated in several works by the DL community for the purpose of comparing TBoxes and knowledge bases (Konev et al. 2009; Lutz and Wolter 2010; Kontchakov, Wolter, and Zakharyaschev 2010; Konev et al. 2011; Botoeva et al. 2014; 2015), as well as in the data exchange and schema mapping literature (Fagin et al. 2008; Gottlob, Pichler, and Savenko 2011). To the best of our knowledge, it has never been explicitly considered for OBDA specifications.

 Conjunctive queries (CQs) are the type of queries most commonly considered in OBDA. Therefore, a first natural choice would be to compare OBDA specifications with respect to the whole class of CQs. We thus define a notion of CQ-entailment between OBDA specifications, which is a natural generalization of previously studied notions of query entailment between ontologies. We also consider the important subclass of instance queries (IQs) and analyze the notion of IQ-entailment between specifications. Moreover, in many application contexts, only a (small) set of predefined queries are available or of interest to the user(s). In such cases, it may be more appropriate to tailor the comparison of specifications to a specific set of queries. For this reason,
we also study the notions of single CQ-entailment and single IQ-entailment, which compare specifications with respect to
a single CQ or IQ.

Each of the preceding notions of entailment induces a corresponding notion of equivalence of OBDA specifications. Being weaker than classical logical equivalence, such query-based notions of equivalence lay the formal foundations for more powerful types of optimizations of OBDA specifications, as it is possible to consider a larger set of semantically well-founded transformations and simplifications of the ontology and mapping. The notions of equivalence based upon particular queries can be used to check that modifications to the specification do not inadvertently impact the answers to queries that should not be concerned by the changes.

We present a first investigation of the computational complexity of deciding the above forms of entailment for a pair of OBDA specifications. We consider three different languages (linear, GAV and GLAV) for expressing mappings in OBDA. As ontology languages, we focus on lightweight description logics (DLs) from the DL-Lite family (Calvanese et al. 2007; Artale et al. 2009), which are a popular choice for OBDA. More precisely, we consider the logic DL-LiteR (also known as DL-LiteR\textsuperscript{\textit{core}}), which is the basis for the OWL 2 QL profile (Motik et al. 2012), as well as its fragments DL-Lite\textsuperscript{\textit{core}} and DL-Lite\textsuperscript{RDFS}. The latter is the DL analogue of the RDFS standard (Guha and Brickley 2014).

The paper is structured as follows. After some preliminaries, we formally define in Section 3 the different notions of entailment between OBDA specifications. In Section 4, we provide characterization results that allow us to restrict the number and form of databases and ABoxes that need to be considered when deciding entailment. These results will be used in Section 5 as central tools in our analysis of the complexity of query entailment between OBDA specifications formulated in the DL and mapping languages mentioned above. To obtain our complexity results, we also exploit connections to query containment in the presence of signature restrictions (Bienvenu, Lutz, and Wolter 2012) and KB query inseparability (Botoeva et al. 2014). The results of our complexity analysis are summarized in Figure 1. As shown in the table, the complexity of query-based entailment between OBDA specifications ranges from NL (non-deterministic logarithmic space) for (single) IQ-entailment with linear mappings (and the three considered DL-Lite dialects) to EXPTIME for CQ-entailment in DL-Lite\textit{R} (for the three types of mapping).

For lack of space, some proofs have been deferred to the appendix of the long version (Bienvenu and Rosati 2016).

### 2 Preliminaries

We start from four pairwise disjoint countably infinite sets of names: the set $\mathbb{N}_C$ of concept names (unary predicates), the set $\mathbb{N}_R$ of role names (binary predicates), the set $\mathbb{N}_{rel}$ of relation names (predicates of arbitrary arity), and the set $\mathbb{N}_I$ of constant names (also called individuals). We will call $\mathbb{N}_C \cup \mathbb{N}_R$ the DL signature and $\mathbb{N}_{rel}$ the database signature. Given a syntactic object $O$, we denote by $\text{sig}(O)$ the set of concept names, role names, and relation names occurring in $O$, and by $\text{const}(O)$ the set of constant names occurring in $O$. If $\vec{a}$ is a tuple of constants and $\Xi \subseteq \mathbb{N}_I$, then the notation $\vec{a} \subseteq \Xi$ means that every constant in $\vec{a}$ belongs to $\Xi$. As usual, we use $|O|$ for the size of $O$, i.e. the total number of occurrences of symbols in $O$.

We recall that a DL knowledge base (KB) is a pair $\langle T, A \rangle$ composed of a TBox $T$, expressing intensional knowledge, and an ABox $A$, specifying factual information about particular individuals. Formally, the TBox is a finite set of axioms built from $\mathbb{N}_C$ and $\mathbb{N}_R$, whose form depends on the DL in question. The ABox $A$ is a finite set of concept and role assertions (ground facts) of the forms $A(a)$ and $R(a, b)$, with $a, b \in \mathbb{N}_I$, $A \in \mathbb{N}_C$, and $R \in \mathbb{N}_R$.

As mentioned in the introduction, we focus on the description logic $\text{DL-Lite}_R$ and on two fragments of this logic. A $\text{DL-Lite}_R$ TBox consists of a finite set of concept inclusions $B \subseteq C$ and role inclusions $S \subseteq Q$, where $B, C, S$, and $Q$ are defined according to the following syntax (where $A$ is a concept name and $R$ is a role name):

$$B \rightarrow A \mid \exists S \ C \rightarrow B \mid \neg B \ \ S \rightarrow R \mid R^- \ Q \rightarrow S \mid \neg S$$

A $\text{DL-Lite}_{\text{core}}$ TBox only allows for concept inclusions of the form $B \subseteq C$, while a $\text{DL-Lite}_{\text{RDFS}}$ TBox allows for concept inclusions $B \subseteq A$ and role inclusions $S \subseteq S'$.

The semantics of DL KBs is based upon interpretations, which take the form $I = (\Delta_I, I)$, where $\Delta_I$ is a non-empty set, and the function $\cdot^I$ maps every $A \in \mathbb{N}_C$ to $A^I \subseteq \Delta_I$, every $R \in \mathbb{N}_R$ to $R^I \subseteq \Delta_I \times \Delta_I$, and every $c \in \mathbb{N}_I$ to $c^I \in \Delta_I$. The function $\cdot^I$ is straightforwardly extended to arbitrary concepts and roles, e.g. $(\exists R)^I = \{(e_1, e_2) \in R^I\}$ and $(R^-)^I = \{(e_2, e_1) \in R^I\}$. An interpretation $I$ satisfies an inclusion $G \subseteq H$ if $G^I \subseteq H^I$, and it satisfies an assertion $A(a)$ (resp. $R(a, b)$) if $a^I \in A^I$ (resp. $(a^I, b^I) \in R^I$). We call $I$ a model of a KB $\langle T, A \rangle$ if it satisfies all inclusions in $T$ and assertions in $A$. If $\langle T, A \rangle$ has no models, it is unsatisfiable, written $\langle T, A \rangle \models \bot$.

Every satisfiable $\text{DL-Lite}_R$ KB $\langle T, A \rangle$ possesses a canonical model $I_{T, A}$ constructed as follows. The domain $\Delta_{T, A}$ of $I_{T, A}$ consists of all words $aR_1 \ldots R_n$ ($n \geq 0$) such that $a \in \text{const}(A)$, $R_i \in \{R, R^-, R \in \mathbb{N}_R\}$, and:

- if $n \geq 1$, then $I_{T, A} \models \exists R_1(a)$;
- for $1 \leq i < n$, $I_{T, A} \models \exists R_i \subseteq \exists R_{i+1}$ and $R^-_i \not= R_{i+1}$.

The interpretation function is defined as follows:

$$a^I_{T, A} = a \text{ for all } a \in \mathbb{N}_I$$

$$A^I_{T, A} = \{a \in \text{const}(A) \mid T, A \models A(a)\}$$

$$\cup \{aR_1 \ldots R_n \mid n \geq 1 \text{ and } T \models \exists R_n \subseteq A\}$$

$$R^I_{T, A} = \{(a, b) \mid R(a, b) \in A\} \cup \{(w_1, w_2) \mid w_2 = w_1S \text{ and } T \models S \subseteq R\} \cup \{(w_2, w_1) \mid w_2 = w_1S \text{ and } T \models S^- \subseteq R\}$$

Observe that the elements of $\Delta_{T, A}$ can be naturally viewed as a set of trees, rooted at the constants.

Now we introduce OBDA specifications, which consist of a DL ontology and a mapping, with the latter serving to specify the semantic relationship holding between elements
of the DL and database vocabularies. In this paper, we focus on the well-known class of GLAV (‘global-and-local-as-view’) mappings (Doan, Halevy, and Ives 2012).

Mappings are formally defined as follows. An atom is an expression $P(\vec{t})$ where $P$ is a predicate and $\vec{t}$ is a tuple of terms (variables and constants). A (GLAV) mapping assertion $m$ is an expression of the form $q_\vec{x}(\vec{m}) \rightarrow q_\vec{y}(\vec{m})$, where $q_\vec{x}(\vec{m})$ (the body of $m$, denoted $body(m)$) is a conjunction of atoms over predicates from $N_m$ and constants from $N_0$, and $q_\vec{y}(\vec{m})$ (called the head of $m$, $head(m)$) is a conjunction of atoms using predicates from $N_C \cup N_R$ and constants from $N_0$. $\vec{x}$ and $\vec{y}$, the frontier variables of $m$, are the variables that appear both in $q_\vec{x}$ and in $q_\vec{y}$. We use $vars(m)$ (resp. $vars(M)$, terms($m$)) to refer to the frontier variables (resp. variables, terms) of $m$. The arity of $m$ is the number of its frontier variables. When $q_\vec{x}(\vec{m})$ has the form $P(\vec{t})$ (i.e., $q_\vec{x}(\vec{m})$ is a single atom whose arguments are $\vec{t}$), we call $m$ a GAV mapping assertion. A linear mapping assertion is a GAV assertion whose body consists of a single atom. A (GLAV) mapping $M$ is a set of mapping assertions. A GAV (resp. linear) mapping is a mapping constituted of GAV (resp. linear) mapping assertions. Without loss of generality, we assume that in every mapping $M$, the every pair of distinct mapping assertions uses pairwise disjoint sets of variables.

An OBDA specification is a pair $\Gamma = \langle T, \mathcal{M} \rangle$, where $T$ is a TBox and $\mathcal{M}$ is a mapping. Given a mapping assertion $m$ of arity $n$ and an $n$-tuple of constants $\vec{a}$, we denote by $m(\vec{a})$ the assertion obtained from $m$ by replacing the frontier variables with the constants in $\vec{a}$.

Given a set of atoms $AT$, the function $gr$ returns a set $gr(AT)$ of ground atoms obtained from $AT$ by replacing every variable symbol $x$ with a fresh constant symbol $c_x$ that does not occur in the considered mapping or database. We assume without loss of generality that if $AT \neq AT'$, then $gr(AT)$ and $gr(AT')$ use distinct fresh constants. Note that there is a natural isomorphism from $AT$ to $gr(AT)$ which maps every constant in $AT$ to itself and every variable $x$ to the corresponding constant symbol $c_x$ in $gr(AT)$.

In this paper, a database (instance) is a finite set of ground atoms using relation names from $N_R$ and constant names from $N_0$. Given a mapping $\mathcal{M}$ and a database $D$, we define the ABox for $D$ and $\mathcal{M}$, denoted as $A_{\mathcal{M}, D}$, as follows:

$$\beta \in gr(head(m(\vec{a}))) \mid m \in \mathcal{M} \text{ and } D \models \exists \vec{y}. body(m(\vec{a}))$$

where we assume that $\vec{y}$ are the variables occurring in $body(m(\vec{a}))$. If $A \subseteq A_{\mathcal{M}, D}$, then we denote by $const(A)\{\vec{x}\}$ the constants in $A$ that occur neither in $D$ nor in $\mathcal{M}$ (i.e., the fresh constants introduced when grounding mapping heads).

**Remark 1.** In the case of GAV mappings, the ABox $A_{\mathcal{M}, D}$ is guaranteed to be of polynomial size, whereas for GLAV mappings, it can be exponentially large due to presence of existential variables in mapping heads. Indeed, take $D = \{T_i(0), T_i(1) \mid 1 \leq i \leq n\}$ and suppose $\mathcal{M}$ contains $T_1(x_1) \land \ldots \land T_n(x_n) \rightarrow \exists z R_1(z, x_1) \land \ldots \land R_n(z, x_n)$

For each $\tau = (v_1, \ldots, v_n) \in \{0, 1\}^n$, we create a fresh constant $c_\tau$ and include $R_1(c_\tau, v_1), \ldots, R_n(c_\tau, v_n)$ in $A_{\mathcal{M}, D}$.

Given an OBDA specification $\Gamma = \langle T, \mathcal{M} \rangle$ and a database $D$, we define the models of $\Gamma$ and $D$, denoted $Mods(\Gamma, D)$, as the set of models of the KB $\langle T, A_{\mathcal{M}, D} \rangle$. When $Mods(\Gamma, D) = \emptyset$, we write $\langle T, \mathcal{M}, D \rangle \models \bot$.

We are interested in the problem of answering instance queries and conjunctive queries over a pair composed of an OBDA specification and a database. A conjunctive query (CQ) takes the form $q(\vec{x}) = \exists \vec{y}(\vec{a}_1) \land \ldots \land \vec{a}_n)$ where every $\vec{a}_i$ is an atomic whose arguments are either constants or variables from $\vec{x} \cup \vec{y}$. The free variables $\vec{x}$ are called answer variables, and a CQ is called Boolean if it has no answer variables. For a non-Boolean CQ $q$ with answer variables $x_1, \ldots, x_k$ a tuple of constants $\vec{a} = (a_1, \ldots, a_k)$ occurring in $A$ is said to be a certain answer for $q$ w.r.t. $A$ just in the case that $A \models q(\vec{a})$, where $q(\vec{a})$ is the Boolean query obtained from $q$ by replacing each $x_i$ by $a_i$. An instance query (IQ) is a CQ consisting of a single atom of the form $A(x)$ or $R(x, y)$, with $A$ concept name, $R$ role name, and $x, y$ distinct answer variables. We use IQ (resp. IQ) to refer the set of all CQs (resp. IQs) over the DL signature $N_C \cup N_R$.

Given an OBDA specification $\Gamma = \langle T, \mathcal{M} \rangle$, a database $D$, and a CQ $q$, we define the certain answers for $q$ w.r.t. $\langle T, \mathcal{M}, D \rangle$ as the tuples of constants from $\text{const}(D) \cup \text{const}(\mathcal{M})$ that are certain answers for $q$ w.r.t. $\langle T, \mathcal{M}, D \rangle$. In particular, for Boolean CQs, we say that $q$ is entailed by $\langle T, \mathcal{M}, D \rangle$, denoted by $\langle T, \mathcal{M}, D \rangle \models q$, if $\forall I \models q$ for every $I \models \text{const}(\Gamma, D)$. Note that for non-Boolean queries, we only consider constants occurring either in $D$ or in $\mathcal{M}$ and thereby exclude the fresh constants in $\text{const}(\mathcal{M})$.

Importantly, canonical models characterize CQ answering over DL-Lite$_R$ KBs. More precisely: for every CQ $q$ and every tuple $\vec{a}$ of constants occurring in $A$, $\langle T, \mathcal{M}, D \rangle \models q(\vec{a})$ iff $I_{T, \mathcal{M}} \models q(\vec{a})$. Given an OBDA specification $\Gamma = \langle T, \mathcal{M} \rangle$ and a data instance $D$, we use $I_{T, \mathcal{M}}$ to denote the canonical model for the KB $\langle T, \mathcal{M}, D \rangle$. The preceding property immediately carries over to the OBDA setting:

**Proposition 1.** Given a DL-Lite$_R$ OBDA specification $\Gamma = \langle T, \mathcal{M} \rangle$, a database $D$, a CQ $q$ and a tuple $\vec{a} \subseteq \text{const}(D) \cup \text{const}(\mathcal{M})$, we have $\langle T, D \rangle \models q(\vec{a})$ iff $I_{T, D} \models q(\vec{a})$.

### 3 Query-based Entailment for OBDA Specifications

We start by recalling the classical notion of entailment between OBDA specifications.

**Definition 1 (Logical entailment).** An OBDA specification $\langle T_1, M_1 \rangle$ logically entails $\langle T_2, M_2 \rangle$, written $\langle T_1, M_1 \rangle \models_{\log} \langle T_2, M_2 \rangle$ if and only the first-order theory $T_1 \cup M_1$ logically entails the first-order theory $T_2 \cup M_2$.

We next formally define the different notions of query-based entailment between OBDA specifications considered in this paper. First, we introduce a notion of entailment that compares specifications based upon the constraints they impose regarding consistency.

**Definition 2 (\bot-entailment).** An OBDA specification $\langle T_1, M_1 \rangle$ \bot-entails $\langle T_2, M_2 \rangle$, written $\langle T_1, M_1 \rangle \models_\bot \langle T_2, M_2 \rangle$, if, for every database $D$,

$$\langle T_2, M_2, D \rangle \models \bot \Rightarrow \langle T_1, M_1, D \rangle \models \bot$$

Next, we define a notion of query entailment between OBDA specifications with respect to a single query.
Definition 3 (Single query entailment). Let $q$ be a query. An OBDA specification $\langle T_1, M_1 \rangle$ $q$-entails $\langle T_2, M_2 \rangle$, written $\langle T_1, M_1 \rangle \models q (\langle T_2, M_2 \rangle)$, if $\langle q, T_1, M_1 \rangle = \top \langle T_2, M_2 \rangle$ and for every database $D$ and tuple $\alpha \in \text{const}(D) \cup \text{const}(M_2)$:

$$\langle T_2, M_2, D \rangle \models q(\alpha) \Rightarrow \langle T_1, M_1, D \rangle \models q(\alpha)$$

When $q$ is an IQ (resp. CQ), we call the preceding entailment relation single IQ-entailment (resp. single CQ-entailment).

We can generalize the previous definition to classes of queries as follows.

Definition 4 (Query entailment). Let $\mathcal{L}$ be a (possibly infinite) set of queries. An OBDA specification $\langle T_1, M_1 \rangle$ $\mathcal{L}$-entails $\langle T_2, M_2 \rangle$, written $\langle T_1, M_1 \rangle \models_{\mathcal{L}} \langle T_2, M_2 \rangle$, if $\langle \mathcal{L}, T_1, M_1 \rangle = \top \langle T_2, M_2 \rangle$ and for every query $q \in \mathcal{L}$, we call the preceding entailment relation IQ-entailment, and for $\mathcal{L} = \text{CQ}$, we use the term CQ-entailment.

Note that each notion of entailment induces a notion of equivalence between OBDA specifications, corresponding to the case when the entailment holds in both directions (we omit the formal definitions due to space limitations).

The following property immediately follows from the preceding definitions.

Proposition 2. Let $\langle T_1, M_1 \rangle, \langle T_2, M_2 \rangle$ be two OBDA specifications, and let $\mathcal{L}_1$ be a set of queries. Then, $\langle T_1, M_1 \rangle \models_{\log} \langle T_2, M_2 \rangle$ implies $\langle T_1, M_1 \rangle \models_{\mathcal{L}_1} \langle T_2, M_2 \rangle$. Moreover, if $\mathcal{L}_2 \subseteq \mathcal{L}_1$, then $\langle T_1, M_1 \rangle \models_{\mathcal{L}_2} \langle T_2, M_2 \rangle$ implies $\langle T_1, M_1 \rangle \models_{\mathcal{L}_1} \langle T_2, M_2 \rangle$.

As a consequence of Proposition 2, we have that logical entailment implies CQ-entailment, and CQ-entailment implies IQ-entailment. The converse implications do not hold, as the following examples demonstrate.

Example 1. We start by illustrating the difference between logical entailment and CQ-entailment. Consider a database containing instances for the relation $\text{EXAM}(\text{studentName, courseName, grade, date})$. Then, let $\Gamma_1 = \langle T_1, M_1 \rangle$, where

$$T_1 = \{ \text{Student} \sqsubseteq \text{Person}, \text{PhDStudent} \sqsubseteq \text{Student} \}$$

$$M_1 = \{ \text{EXAM}(x, y, z, w) \rightarrow \text{Student}(x) \}$$

and let $\Gamma_2 = \langle T_2, M_2 \rangle$, where $T_2 = \{ \text{Student} \sqsubseteq \text{Person} \}$ and $M_2 = M_1$. It is immediate to verify that $\Gamma_2 \not\models_{\log} \Gamma_1$. However, we have that $\Gamma_2 \models_{\text{CQ}} \Gamma_1$. Indeed, $\Gamma_2 \models_{\text{CQ}} \Gamma_1$ can be intuitively explained by the fact that the mapping $M_1$ does not retrieve any instances of the concept $\text{PhDStudent}$ (and there are no subclauses that can indirectly populate it), so the presence of the inclusion $\text{PhDStudent} \sqsubseteq \text{Student}$ in $T_1$ does not have any effect on query answering; in particular, every CQ that mentions the concept $\text{PhDStudent}$ cannot be entailed both under $\Gamma_1$ and under $\Gamma_2$. Notice also that, if we modify the mapping $M_1$ to map $\text{PhDStudent}$ instead of $\text{Student}$ (i.e., if $M_1$ were $\{ \text{EXAM}(x, y, z, w) \rightarrow \text{PhDStudent}(x) \}$), then CQ-entailment between $\Gamma_2$ and $\Gamma_1$ would no longer hold.

Next, consider $\Gamma_3 = \langle T_3, M_3 \rangle$, where $T_3 = \emptyset$ and

$$M_3 = \{ \text{EXAM}(x, y, z, w) \rightarrow \text{Student}(x), \text{EXAM}(x, y, z, w) \rightarrow \text{Person}(x) \}$$

Again, it is immediate to see that $\Gamma_1 \not\models_{\text{CQ}} \Gamma_2$, while we have that $\Gamma_3 \models_{\text{CQ}} \Gamma_2$. Indeed, $\Gamma_3 \models_{\text{CQ}} \Gamma_2$ follows informally from the fact that the mapping $M_3$ is able to "extensionally" simulate the inclusion $\text{Student} \sqsubseteq \text{Person}$ of $T_2$, which is sufficient for $\Gamma_3$ to entail every CQ in the same way as $\Gamma_2$.

Example 2. We slightly modify Example 1 to show the difference between CQ-entailment and IQ-entailment. Consider $\Gamma_1 = \langle T_1, M_1 \rangle$ and $\Gamma_2 = \langle T_2, M_2 \rangle$ where

$$T_1 = \{ \text{Student} \sqsubseteq \text{Person}, \text{Student} \sqsubseteq \exists \text{TakeCourse} \}$$

$$T_2 = \{ \text{Student} \sqsubseteq \text{Person} \}$$

$$M_1 = \{ \text{EXAM}(x, y, z, w) \rightarrow \text{Student}(x) \}$$

It is easily verified that $\Gamma_2 \not\models_{\text{CQ}} \Gamma_1$. Indeed, take the Boolean CQ $\exists x, y \text{TakeCourse}(x, y)$: for every database $D$, this query is not entailed by the pair $(\Gamma_2, D)$, while this is not the case when the specification is $\Gamma_1$. On the other hand, we have that $\Gamma_2 \models_{\text{IQ}} \Gamma_1$: in particular, for every database $D$ and for every pair of individuals $a, b$, neither $(\Gamma_1, D)$ nor $(\Gamma_2, D)$ entails the IQ $\text{TakeCourse}(a, b)$. Finally, let $q$ be the non-Boolean CQ $\exists y \text{TakeCourse}(x, y)$: then, it can be easily verified that the single CQ-entailment $\Gamma_2 \models q \Gamma_1$ holds; while for the CQ $q'$ of the form $\exists y \text{TakeCourse}(x, y)$, the single CQ-entailment $\Gamma_2 \models q' \Gamma_1$ does not hold.

Some of our proofs will exploit results about query entailment for DL knowledge bases, a problem which was first investigated in (Botoeva et al. 2014) and further studied in (Botoeva et al. 2015). The latter work introduces a generalized notion of entailment parameterized both by a query signature and a set of constants, which is defined as follows.

Definition 5. Let $\Sigma$ be a DL signature and $\Upsilon$ a set of constants. A DL KB $\langle T_1, A_1 \rangle (\Sigma, \Upsilon)$-entails a DL KB $\langle T_2, A_2 \rangle$ iff $\langle T_2, A_2 \rangle \models q(\alpha)$ implies $\langle T_1, A_1 \rangle \models q(\alpha)$ for every CQ $q$ with $\text{sig}(q) \subseteq \Sigma$ and every tuple $\alpha \subseteq \Upsilon \cap \text{const}(A_2)$.

The relationship between the two notions of query entailment is witnessed by the following property.

Proposition 3. Let $\langle T_1, M_1 \rangle, \langle T_2, M_2 \rangle$ be two OBDA specifications. Then, $\langle T_1, M_1 \rangle \models_{\text{CQ}} \langle T_2, M_2 \rangle$ iff, for every database $D$, the KB $K_1 = \langle T_1, A_{M_1, D} \rangle (\Sigma, \Upsilon)$-entails the KB $K_2 = \langle T_2, A_{M_2, D} \rangle$, where $\Sigma = \text{sig}(K_2)$ and $\Upsilon = \text{const}(A_{M_2, D}) \setminus \text{const}^2(A_{M_2, D})$.

4 Characterization Results

A key difficulty in developing procedures for checking the various notions of entailment from the previous section is that the definitions quantify over all possible databases $D$. In this section, we show that when ontologies are formulated in DL-Lite$_R$, it is sufficient to consider a finite set of databases. The exact number and shape of these databases will depend on the notion of entailment considered.

To simplify the exposition, we will assume throughout this section that $\Gamma_1 = \langle T_1, M_1 \rangle$ and $\Gamma_2 = \langle T_2, M_2 \rangle$ are OBDA specifications such that $T_1, T_2$ are formulated in DL-Lite$_R$, and $M_1, M_2$ are GLAV mappings.

We first consider $\models_{\text{IQ}}$-entailment and show that it can be characterized in terms of a small number of small databases.

Theorem 1. $\Gamma_1 \models_{\text{IQ}} \Gamma_2$ if and only if $\langle T_1, M_1, D \rangle \models \bot$ for every database $D$ obtained by:
(i) taking two mapping assertions \(m_1, m_2\) from \(M_2\),
(ii) selecting atoms \(a_1 \in \text{head}(m_1)\) and \(a_2 \in \text{head}(m_2)\),
(iii) identifying in \(m_1\) and \(m_2\) some variables from \(a_1\) and \(a_2\) in such a way that \((T, \{\text{gr}(a_1), \text{gr}(a_2)\}) \models \bot\), and
(iv) setting \(D\) equal to \(\text{gr}(\text{body}(m_1) \cup \text{body}(m_2))\).

Proof. The one direction is immediate from the definitions. For the other direction, let us suppose that \(\langle T_1, M_1, D_0 \rangle \models \bot\) for every database \(D\) obtained as in the theorem statement. Let us further suppose that \(\langle T_2, M_2, D_0 \rangle \models \bot\), where \(D_0\) may be any database. We thus have \(\langle T_2, \mathcal{A}_{M_2, D_0} \rangle \models \bot\).

It is well known that every minimal unsatisfiable subset of a DL-Lite KB contains at most two ABox assertions, so there must exist a subset \(A' \subseteq A_{M_2, D_0}\) with \(|A'| \leq 2\) such that \(\langle T_2, A' \rangle \models \bot\). Let \(\gamma\) be the conjunction of atoms obtained by taking for each ABox assertion in \(A'\), a mapping assertion that produced it, identifying the variables (and only those variables) needed to produce the ABox assertion(s), and then taking the conjunction of the atoms in the bodies. We observe that by construction \(D_2 = \text{gr}(\gamma)\) satisfies the conditions of the theorem and is such that \(\langle T_1, M_1, D_0 \rangle \models \bot\). By construction, there is a homomorphism of \(D_2\) into the original database \(D_0\). It follows that \(\langle T_1, M_1, D_0 \rangle \models \bot\).

Next we consider entailment with respect to a specific CQ and show that it suffices to consider a finite (but exponential) number of small databases. We further show that, instead of considering all query answers obtainable from \(A_{M_2, D}\), we can focus on a polynomial subset \(A_2 \subseteq A_{M_2, D}\).

Theorem 2. Let \(q\) be a CQ. Then \(\Gamma_1 \models q \iff \Gamma_2 \models q\) if and only if \(\langle T_2, A_2 \rangle \models q(a)\) for every database \(D\), every tuple \(a\) from \(\text{const}(D) \cup \text{const}(M_2)\), and every ABox \(A_2 \subseteq A_{M_2, D}\) obtained as follows:

(i) take \(k \leq q\) mapping assertions \(m_1, \ldots, m_k \in M_2\),
(ii) identify some of the variables in \(\cup_{i=1}^k \text{fvars}(m_i)\), and
(iii) let \(D = \text{gr}(\cup_{i=1}^k \text{body}(m_i))\) with natural isomorphism \(f\), and \(A_2 = \cup_{i=1}^k \text{gr}(f(\text{head}(m_i)))\).

Proof. Again the one direction is immediate. To show the non-trivial direction, let us suppose that \(\Gamma_1 \models q\) and that \(\langle T_2, M_2, D_0 \rangle \models q\) for every database \(D\) satisfying the conditions of the theorem. Further suppose that \(\langle T_2, M_2, D_0 \rangle \models q(a)\). We focus on the interesting case in which \(\langle T_2, M_2, D_0 \rangle \models \bot\) and \(\langle T_1, M_1, D_0 \rangle \not\models \bot\). Since \(\langle T_2, M_2, D_0 \rangle \models q(a)\), we have \(\langle T_2, A_{M_2, D_0} \rangle \models q(a)\). It is a well-known property of DL-Lite that there exists a subset \(A_2 \subseteq A_{M_2, D_0}\) with \(|A_2| \leq |q|\) such that \(\langle T_2, A_2 \rangle \models q(a)\). Let \(\gamma\) be the least equivalence relation over the assertions in \(A_2\) such that \(\beta \sim \gamma\) for every pair of assertions \(\beta, \gamma \in A_2\) that share a constant from \(\text{const}^*(A_2)\). Next let \(B_1, \ldots, B_k\) be the ABoxes corresponding to the equivalence classes of \(\sim\). By construction, \(A_2\) is the disjoint union of \(B_1, \ldots, B_k\); in particular, we have that \(k \leq |A_2|\) (hence, \(k \leq |q|\)).

Moreover, it follows from the definition of \(\sim\) and the construction of \(A_{M_2, D_0}\) that for every \(B_i\), there exists a mapping assertion \(m_i \in M_2\) and a homomorphism \(h_i : \text{body}(m_i) \to D_0\) such that

\[
B_i \subseteq \text{gr}(h_i(\text{head}(m_i))).
\]

Now let \(FV = \cup_{i=1}^k \text{fvars}(m_i)\), and let \(\equiv\) be the smallest equivalence relation on \(FV\) that satisfies the following property: \(u \equiv v\) if and only if \(u\) and \(v\) appear in \(m_1\) and \(m_j\) respectively, and \(h_1(u) = h_j(v)\). Select one variable from each equivalence class, and let \(m_1', \ldots, m_k'\) be obtained from \(m_1, \ldots, m_k\) by replacing each frontier variable by the representative of its equivalence class (i.e., we merge those frontier variables that are mapped to the same constants). Define a function \(h\) from \(\cup_{i=1}^k \text{terms(body}(m_i))\) to \(\text{const}(D)\) by mapping constants to themselves, every \(v \in \text{vars}(m_i') \setminus \text{fvars}(m_i')\) to \(h_1(v)\), and every \(v \in \text{fvars}(m_i')\) to \(h_j(v)\), where \(v \in \text{fvars}(m_j)\) (note that at least one such \(j\) must exist, and if \(v\) occurs in more than one \(m_j\), the result will be the same no matter which we choose). Observe that for every \(1 \leq i \leq k\), \(h\) is a homomorphism from \(\text{body}(m_i')\) to \(D_0\) that is injective on \(\text{fvars}(m_i')\), and such that \(h(\text{head}(m_i')) = h_i(\text{head}(m_i'))\).

We set \(D' = \text{gr}((\text{body}(m_i') \cup \ldots \cup \text{body}(m_k'))\)). Then let \(f\) be the natural isomorphism from \(\cup_{i=1}^k \text{body}(m_i')\) to \(D'\), and set \(A_2' = \cup_{i=1}^k \text{gr}(f(\text{head}(m_i')))\). Clearly, \(D'\) and \(A_2'\) satisfy the conditions of the theorem statement. We can define an isomorphism \(g\) from \(\cup_{i=1}^k \text{gr}(h(\text{head}(m_i')))\) to \(A_2'\) as follows:

- \(g(c) = c\) for every constant \(c\) in \(\text{head}(m_i')\)
- \(g(d_v) = d'_v\), for every \(v \in \text{vars(\text{head}(m_i'))} \setminus \text{fvars}(m_i')\) which is grounded to the constant \(d_v\) (resp. \(d'_v\) in \(\cup_{i=1}^k \text{gr}(h_i(\text{head}(m_i')))\)) (resp. \(A_2'\))
- \(g(h(v)) = f(v)\), for every \(v \in \text{fvars}(m_i')\)

As \(\langle T_2, A_2 \rangle \models q_0(a)\), \(A_2 \subseteq \cup_{i=1}^k \text{gr}(h_i(\text{head}(m_i')))\), and \(h(\text{head}(m_i')) = h_i(\text{head}(m_i'))\), it follows that \(\langle T_2, A_2' \rangle \models q_0(a)\). Applying our assumption, we obtain \(\langle T_1, M_1, D' \rangle \models q_0(a)\). We next remark that the function which maps \(d \in \text{const}(D')\) to \(h(f^{-1}(d))\) defines a homomorphism from \(D'\) to \(D_0\). Moreover, if we take the tuple \(g(a)\) and replace every constant \(d\) in \(g(a)\) with \(h(f^{-1}(d))\), we get the tuple \(a\). We thus obtain \(\langle T_1, M_1, D_0 \rangle \models q_0(a)\).

If we replace CQs by IQs, the preceding theorem can be refined as follows:

Theorem 3. Let \(q\) be an IQ. Then \(\Gamma_1 \models q \iff \Gamma_2 \models q\) if and only if \(\langle T_2, A_2 \rangle \models q(a)\) for every database \(D\), every tuple \(a\) from \(\text{const}(D) \cup \text{const}(M_2)\), and every ABox \(A_2 \subseteq A_{M_2, D}\) obtained as follows:

(i) taking a mapping assertion \(m\) from \(M_2\) and choosing an atom \(a \in \text{head}(m)\),
(ii) possibly identifying in \(m\) the (at most two) frontier variables appearing in \(a\), and
(iii) letting \(D = \text{gr}((\text{body}(m))\setminus a)\) with natural isomorphism \(f\), and \(A_2 = \text{gr}(f(\text{head}(m)))\).

Proof. We simply note that in the proof of Theorem 2, if we restrict our attention to instance queries, then we have \(k = 1\), and it is only necessary to identify those variables in the head atom of the mapping that leads to introducing the single ABox assertion of interest.
We now consider entailment with respect to query classes. For CQ-entailment, the characterization is actually a bit simpler than for single-CQ entailment, since we only need to consider the database-ABox pairs induced from single mapping assertions (rather than sets of mapping assertions).

**Theorem 4.** $\Gamma_1 \models_{\text{CQ}} \Gamma_2$ if and only if $\langle T_1, M_1, D_1 \rangle \models q(\bar{a})$ implies $\langle T_2, M_2, D_2 \rangle \models q(\bar{a})$ for every CQ $q$, every database $D$, every tuple $\bar{a}$ from $\text{const}(D) \cup \text{const}(M_2)$, and every ABox $A_2 \subseteq A_{M_2,D}$ obtained as follows:

(i) take a mapping assertion $m$ from $M_2$,
(ii) possibly identify some of the variables in $\text{vars}(m)$, and
(iii) let $D' = gr(\text{body}(m))$ with natural isomorphism $f$, and $A_2 = gr(f(\text{head}(m)))$.

**Proof.** Suppose that $\Gamma_1 \models_{\text{CQ}} \Gamma_2$ and $\langle T_2, A_2 \rangle \models q(\bar{a})$ implies $\langle T_1, M_1, D \rangle \models q(\bar{a})$ for every CQ $q$ and every database $D$ and ABox $A_2 \subseteq A_{M_2,D}$ obtained as in the theorem statement. Further suppose that $\langle T_2, M_2, D_2 \rangle \models q_0(\bar{a})$. We focus on the interesting case in which both $\langle T_2, M_2, D_2 \rangle \not\models \perp$ and $\langle T_1, M_1, D \rangle \not\models \perp$. We let $A_{M_1,D}$ and $A_{M_2,D}$ be the corresponding ABoxes, and let $I_1$ and $I_2$ be the canonical models of $\langle T_1, A_{M_1,D} \rangle$ and $\langle T_2, A_{M_2,D} \rangle$ respectively. Since $\langle T_2, M_2, D_2 \rangle \models q(\bar{a})$, we have $I_2 \models q_0(\bar{a})$ by Proposition 1, and thus there must exist a match $\pi$ for $q_0(\bar{a})$ in $I_2$ (i.e., a homomorphism of $q_0$ into $I_2$ that sends the answer variables $\vec{x}$ of $q_0$ to $\vec{a}$). This match gives rise to an equivalence relation $\sim_\pi$ on the atoms in $\pi$, defined as follows: $\alpha \sim_\pi \alpha'$ just in the case that $\alpha$ and $\alpha'$ share a variable $v$ such that $\pi(v) \in (\Delta^2 \setminus \text{const}(A_{M_2,D})) \cup \text{const}^3(A_{M_2,D})$ (that is, $\pi(v)$ is not a constant from $D_0$ nor $M_2$). This equivalence relation in turn induces a partition of $q_0$ into connected subqueries $q_1, \ldots, q_n$, where every $q_i$ corresponds to one of the equivalence classes of $\sim_\pi$. By construction, each of the queries $q_i$ satisfies exactly one of the following conditions:

- $q_i$ consists of a single atom all of whose variables are mapped by $\pi$ to constants from $\text{const}(D_0) \cup \text{const}(M_2)$;
- every atom in $q_i$ contains a variable that is not mapped to a constant occurring in $D_0$ or $M_2$, and at least one term in $q_i$ is mapped to a constant in $A_{M_2,D}$;
- $q_i$ contains no constants and none of its variables are mapped to constants in $A_{M_2,D}$.

We associate a tuple $\vec{x}_i$ of answer variables with each $q_i$. If $q_i$ satisfies the first or second condition, then every variable in $q_i$ that is mapped to a constant from $\text{const}(D_0) \cup \text{const}(M_2)$ is designated as an answer variable. If the third condition holds, then $q_i$ has no answer variables. By construction, we have $x_i \models q_i(\vec{b}_i)$, where $\vec{b}_i = \pi(\vec{x}_i)$.

Our aim is to show that $\langle T_1, M_1, D_0 \rangle \models q_0(\bar{a})$. Because of the way we defined the queries $q_i(\vec{x}_i)$ and tuples $\vec{b}_i$, it is sufficient to show that $\langle T_1, M_1, D_0 \rangle \models q_i(\vec{b}_i)$ for every $1 \leq i \leq n$. Indeed, suppose this is the case, and let $\pi'_i$ be a match of $q_i(\vec{b}_i)$ in the canonical model $I_1$ of $\langle T_1, A_{M_1,D_0} \rangle$. If $v \in \vec{x}_i \cap \vec{x}_j$, then we must have $\pi'_i(v) = \pi'_j(v) \forall i \leq j$. It follows that we can define a match $\pi'$ for $q_0$ in $I_0$ by setting $\pi'(v) = \pi'_i(v)$, where $i$ is chosen such that $v \in \vec{x}_i$.

We briefly sketch the argument that $\langle T_1, M_1, D_0 \rangle \models q_i(\vec{b}_i)$ (details are in the appendix). The key step is to show, using properties of canonical models, that there exists a mapping assertion $m_i \in M_2$ and a homomorphism $h_i$ of $\text{body}(m_i)$ into $D_0$ such that $\langle T_2, gr(h_i(\text{head}(m_i))) \rangle \models q_i(\vec{b}_i)$. We let $m'_i$ be obtained from $m_i$ by identifying frontier variables $y, z$ whenever $h_i(y) = h_i(z)$. We apply the assumption to the pair consisting of the database $D_{m'_i} = gr(\text{body}(m'_i))$ and the ABox $gr(f(\text{head}(m'_i)))$ (with $f$ the natural isomorphism) to obtain $\langle T_1, M_1, D_{m'_i} \rangle \models q_i(\vec{b}_i)$, with $\vec{b}_i$ the tuple of constants from $D_{m'_i}$ corresponding to $\vec{b}_i$.

From this, we can derive $\langle T_1, M_1, D_0 \rangle \models q_i(\vec{b}_i)$.

For IQ entailment, it suffices to consider the same databases and ABoxes as for the single IQ case.

**Theorem 5.** $\Gamma_1 \models_{\text{IQ}} \Gamma_2$ if and only if $\Gamma_1 \models_{\text{IQ}} \Gamma_2$ and $\langle T_2, A_2 \rangle \models q(\bar{a})$ implies $\langle T_1, M_1, D \rangle \models q(\bar{a})$ for every IQ $q$ and every database-ABox pair $(D, A_2)$ and tuple of constants $\bar{a}$ that satisfies the conditions of Theorem 3.

Finally, we close this section with the following result, which shows that if CQ entailment holds, then this is witnessed by a polynomial-size subset of the ABox $A_{M_1,D}$.

**Theorem 6.** Let $A_2 \subseteq A_{M_2,D}$, and suppose that $\langle T_2, A_2 \rangle \models q(\bar{a})$ implies $\langle T_1, M_1, D \rangle \models q(\bar{a})$ for every Boolean CQ $q(\bar{a})$ with $\bar{a} \subseteq \text{const}(D) \cup \text{const}(M_2)$. Then there exists a subset $A_1 \subseteq A_{M_1,D}$ with $|A_1| \leq 2|T_1| \cdot |\langle 2|T_2 + 1 \rangle \cdot |M_1| \cdot (|A_2| + 2|T_2| \cdot (|D| + |M_1|))^2$, such that $\langle T_2, A_2 \rangle \models q(\bar{a})$ implies $\langle T_1, M_1, D \rangle \models q(\bar{a})$ for every Boolean CQ $q(\bar{a})$ with $\bar{a} \subseteq \text{const}(D) \cup \text{const}(M_2)$. It follows from results in (Botoeva et al. 2015) that the latter property is true iff there is a homomorphism from $I_2$ to $I_1$ that is the identity on the constants in $D$ and $M_2$, or more precisely, a function $h : \Delta^2 \rightarrow \Delta^1$ such that:

- $h(e) = e$ for every $e \in \text{const}(D) \cup \text{const}(M_2)$
- $e \in A^2$ implies $h(e) \in A^1$
- $(e, e') \in R^2$ implies $h(e), h(e') \in R^1$

Note that the constants in $\text{const}^3(A_{M_2,D})$ need not be mapped to themselves.

Our objective is to show how to use $h$ to construct a homomorphism $g$ from $I_2$ to $I_1$ with the property that the set $\Theta = \{a | a = \text{range}(g)\}$ has size polynomially bounded in the size of $D, A_2, T_2,$ and $M_1$. This will allow us to identify a small subset of $A_{M_1,D}$ that is sufficient to infer all of the queries involving constants from $D$ or $M_2$ that are entailed by $\langle T_2, A_2 \rangle$. To construct the desired homomorphism $g$, we will define a sequence $g_0, g_1, g_2, \ldots$ of
partial homomorphisms from $I_2$ to $I_1$ such that $g_{i+1}$ extends $g_i$ for every $i \geq 0$ (that is, $\text{dom}(g_i) \subseteq \text{dom}(g_{i+1})$ and $g_{i+1}(e) = g_i(e)$ for all $e \in \text{dom}(g_i)$). We will ensure that for every $e \in \Delta^{2i}$, there is some $i$ such that $e \in \text{dom}(g_i)$, which allows us to take $g$ to be the limit of this sequence.

For the initial partial homomorphism $g_0$, we include in the domain $\text{dom}(g_0)$: (i) all constants in $\text{const}(A_2)$, and (ii) all elements $awR$ such that $h(e) \not\in \text{const}(D) \cup \text{const}(M_1)$ for every $e$ that is a prefix of $awR$, and set $g_0(e) = h(e)$ for every $e \in \text{dom}(g_0)$. Intuitively, $g_0$ copies the value of $h$ for all constants in $A_2$ as well as for all elements $e$ such that every ‘ancestor’ of $e$ is mapped to an element of $\Delta^{2i} \setminus (\text{const}(D) \cup \text{const}(M_1))$.

At each stage $i \geq 1$, we will extend $g_{i-1}$ to some additional elements from $\Delta^{2i} \setminus \text{const}(A_2)$ in such a way that the cardinality of the set $\{a \mid aw \in \text{range}(g_0)\}$ does not exceed the required polynomial bound. The construction is lengthy and rather involved (see the appendix for details), but the basic intuition is as follows: we exploit the regularity of canonical models in order to identify subtrees of $I_2$ that are similar in structure, and we modify the homomorphism $h$ so similar subtrees are mapped in the same way into $I_1$.

Figure 1: Complexity of query-based entailment between OBDA specifications. All results are completeness results, unless otherwise indicated.

5 Complexity Results

In this section, we investigate the computational properties of the different notions of entailment between OBDA specifications defined in the previous section.

The results of our complexity analysis are displayed in Figure 1. In what follows, we formally state the different complexity results and provide some ideas about the proofs.

We begin by considering the complexity of deciding classical entailment between OBDA specifications.

Theorem 7. For DL-Lite$_{RDFS}$, DL-Lite$_{core}$ and DL-Lite$_R$, logical entailment between OBDA specifications is NP-complete for GAV and GLAV mappings, and NL-complete for linear mappings.

Proof. Let $\Gamma_1 = \langle T_1, M_1 \rangle$, $\Gamma_2 = \langle T_2, M_2 \rangle$. First, it is easy to see that $\Gamma_1 \models_{\text{log}} \Gamma_2$ iff (i) $T_1 \models T_2$; and (ii) $\Gamma_1 \models_{\text{log}} M_2$. In the case of a DL-Lite$_R$ TBox, property (i) can be decided in NL (Artale et al. 2009). Property (ii) can be decided by an algorithm that, for every assertion $m \in M_2$, first builds a database $D$ corresponding to $\text{gr(body}(m))$, and then checks whether $(\Gamma_1, D)$ entails the CQ corresponding to the head of $m$ whose frontier variables have been replaced by the corresponding constants. This algorithm runs in NP in the case of GAV and GLAV mappings, and it can be made to run in NL in the case of linear mappings (see the appendix for details), which implies the overall upper bounds in the theorem statement. For both DL-Lite$_{core}$ and DL-Lite$_{RDFS}$ TBoxes, the lower bound for GAV mappings can be obtained through an easy reduction of CQ containment to logical entailment, while the lower bounds for linear mappings follow from a reduction of the entailment of a concept inclusion axiom in a TBox that is both in DL-Lite$_{core}$ and in DL-Lite$_{RDFS}$.

Using Theorem 1, we can pinpoint the complexity of $\perp$-entailment. We remark that $\perp$-entailment is trivial in DL-Lite$_{RDFS}$, since every KB in this logic is satisfiable.

Theorem 8. For both DL-Lite$_R$ and DL-Lite$_{core}$ TBoxes, the $\perp$-entailment problem between OBDA specifications is NP-complete in the case of GAV / GLAV mappings, and NL-complete in the case of linear mappings.

Proof. We know from Theorem 1 that $\Gamma_1 \not\models \Gamma_2$ iff $\langle T_1, M_1, D \rangle \models \perp$ for every database $D$ satisfying the conditions of the theorem statement. For the GAV / GLAV case, we compute these databases in polynomial time and for every such database $D$, we guess a polynomial-size proof that $\langle T_1, M_1, D \rangle \models \perp$. For the linear case, we observe that the databases satisfying the conditions of Theorem 1 correspond to taking the conjunction of two single-atom mapping bodies and possibly performing a bounded number of variable unifications, and (representations of) such databases can be enumerated in logarithmic space. For every such database $D$, we can check using an NL oracle whether $\langle T_1, M_1, D \rangle \models \perp$. Since L$^N_L = NL$, we obtain an NL procedure.

We next analyze the complexity of single CQ-entailment, showing it to be $\Pi^2_N$-complete for all of the DLs and mapping languages considered in this paper.

Theorem 9. The single CQ-entailment problem is $\Pi^2_N$-complete for OBDA specifications based upon DL-Lite$_R$ TBoxes and GLAV mappings. The lower bound holds even for linear mappings and when both TBoxes are empty.

Proof. For the upper bound, consider two OBDA specifications $\Gamma_1 = \langle T_1, M_1 \rangle$ and $\Gamma_2 = \langle T_2, M_2 \rangle$. From Theorems 1 and 2, we know that $\Gamma_1 \not\models_{\text{CQ}} \Gamma_2$ if and only if one of the following holds:

- there is a database $D$ satisfying the conditions of Theorem 1 such that $\langle T_1, M_1, D \rangle \not\models \perp$;
- there is a pair $(D, A_2)$ satisfying the conditions of Theorem 2 such that $\langle T_2, A_2 \rangle \models q(\bar{a})$, $\langle T_2, A_2 \rangle \not\models \perp$, and $\langle T_1, M_1, D \rangle \not\models \perp$.

The first item can be checked using an NP oracle (by Theorem 8). To check the second item, we remark that for every pair $(D, A_2)$ satisfying the conditions of Theorem 2, the number of facts in $D$ (resp. $A_2$) cannot exceed $|q| \cdot \text{maxbody}$ (resp. $|q| \cdot \text{maxhead}$) where $\text{maxbody}$ (resp. $\text{maxhead}$) is the maximum number of atoms appearing in the body (resp.
head) of a mapping assertion in \( M_2 \). It follows that to show that the second item above is violated, we can guess a pair \( (D, A_2) \) respecting these size bounds, together with a tuple of constants \( \vec{a} \) and a polynomial-size proof that \( \langle T_2, A_2 \rangle \models q(\vec{a}) \), and then we can verify using an NP oracle that \( \langle T_1, M_1, D \rangle \not\models q(\vec{a}) \). We therefore obtain a \( \Sigma^p_2 \) procedure for deciding the complement of our problem.

For the lower bound, we utilize a result from (Bienvenu, Lutz, and Wolter 2012) on query containment over signature-restricted ABoxes. In that paper, it is shown how, given a 2QBF \( \forall \vec{x} \exists \vec{y} \varphi(\vec{x}, \vec{y}) \), one can construct a TBox \( T \), Boolean CQs \( q_1 \) and \( q_2 \), and a signature \( \Sigma \) such that \( \forall \vec{x} \exists \vec{y} \varphi(\vec{x}, \vec{y}) \) is valid iff \( T, A \models q_1 \Rightarrow T, A \models q_2 \) for all \( \Sigma \)-ABoxes \( A \). We will not detail the construction but simply remark that the same TBox \( T = \{ T \subseteq V, F \subseteq V \} \) is used for all QBFs, the signature \( \Sigma \) is given by \( \{ \text{sig}(T) \cup \text{sig}(q_1) \cup \text{sig}(q_2) \} \setminus \{ V \} \), and the query \( q_2 \) is such that \( V \not\models \text{sig}(q_2) \).

In what follows, we will show how given \( T, q_1, q_2, \) and \( \Sigma \) as above, we can construct the problem of testing whether \( T, A \models q_1 \) implies \( T, A \models q_2 \) for all \( \Sigma \)-ABoxes to the problem of single CQ entailment. We will use \( \Sigma \) for our database instances, and we create two copies \( \Sigma_1 = \{ P^1 \mid P \in \Sigma \} \) and \( \Sigma_2 = \{ P^2 \mid P \in \Sigma \} \) of the signature \( \Sigma \) to be used in the head of mapping assertions. Next, we define sets of mapping assertions \( \text{copy}^1(\Sigma) \) and \( \text{copy}^2(\Sigma) \) that simply copy all of the predicates in \( \Sigma \) into the corresponding symbol in \( \Sigma_1 \) (resp. \( \Sigma_2 \)). Formally, for \( j \in \{ 1, 2 \} \),

\[
\text{copy}^j(\Sigma) = \{ A(x) \rightarrow A^j(x) \mid A \in \Sigma \cap N_C \} \cup \{ R(x, y) \rightarrow R^j(x, y) \mid R \in \Sigma \cap N_R \}
\]

We further define, given a data signature \( \Lambda_1 \) and DL signature \( \Lambda_2 \), a set \( \text{pop}(\Lambda_1, \Lambda_2) \) of mapping assertions that populates the relations in \( \Lambda_2 \) using all possible combinations of the constants appearing in tuples over \( \Lambda_1 \):

\[
\text{pop}(\Lambda_1, \Lambda_2) = \{ P(\vec{x}) \rightarrow A(x) \mid P \in \Lambda_1, A \in \Lambda_2 \cap N_C, x \in \vec{x} \} \cup \{ P(\vec{x}) \rightarrow R(x, y) \mid R \in \Lambda_1, y \in \Lambda_2 \cap N_R, x \in \vec{x} \}
\]

Using \( \text{copy}^1(\Sigma) \), \( \text{copy}^2(\Sigma) \), \( \text{pop}(\Sigma, \Sigma^1) \), and \( \text{pop}(\Sigma, \Sigma^2) \), we construct the following mappings:

\[
M_1 = \text{pop}(\Sigma, \Sigma \cup \{ V \}) \cup \text{copy}^1(\Sigma) \\
M_2 = \text{copy}^1(\Sigma) \cup \text{pop}(\Sigma, \Sigma^2) \cup \{ T(x) \rightarrow V(x), F(x) \rightarrow V(x) \}
\]

Observe that both mappings are linear. For the query, we let \( q_1' \) (resp. \( q_2' \)) be obtained from \( q_1 \) (resp. \( q_2 \)) by replacing every predicate \( P \) by \( P^1 \) (resp. \( P^2 \)). We also rename variables so that \( q_1' \) and \( q_2' \) do not share any variables. We then let \( q \) be the CQ obtained by taking the conjunction of \( q_1' \) and \( q_2' \) and existentially quantifying all variables. In the appendix, we show that \( \langle 0, M_1 \rangle \models q \langle 0, M_2 \rangle \) iff \( T, A \models q_1 \Rightarrow T, A \models q_2 \) for all \( \Sigma \)-ABoxes. By combining this with the reduction from (Bienvenu, Lutz, and Wolter 2012), we obtain a reduction from universal 2QBF to the single CQ-entailment problem, establishing \( \Pi^p_2 \)-hardness of the latter.

If we consider IQs instead, the complexity drops to either NP- or NL-complete.

**Theorem 10.** For DL-Lite_RDFS, DL-Lite_core, and DL-Lite_R, the single IQ-entailment problem between OBDA specifications is NP-complete for both GAV and GLAV mappings, and is NL-complete for linear mappings.

**Proof.** We give the arguments for GAV and GLAV mappings (for the linear case, see the appendix). For the NP upper bound, consider specifications \( \Gamma_1 = \langle T_1, M_1 \rangle \) and \( \Gamma_2 = \langle T_2, M_2 \rangle \), and let \( q \) be an IQ. By Theorem 3, \( \Gamma_1 \models q \) iff \( \Gamma_1 \models_1 \Gamma_2 \) and \( \langle T_2, A_2 \rangle \models q(\vec{a}) \) implies \( \langle T_1, M_1, D \rangle \models q(\vec{a}) \) for every database-ABox pair \( (D, A_2) \) obtained in the manner described by the theorem statement.

We already know that it is in NP to test whether \( \Gamma_1 \models_1 \Gamma_2 \). For the second property, observe that there are only polynomially many pairs \( (D, A_2) \) satisfying the required conditions, since each corresponds to choosing a mapping assertion \( m \) in \( M_2 \), an atom \( \alpha \in \text{head}(m) \), and deciding whether or not to identify the (at most two) variables in \( \alpha \). For every such pair \( (D, A_2) \), we compute in polynomial time the set of Boolean IQs \( \beta \) obtained by instanciating the IQ \( q \) with constants \( \text{const}(D) \cup \text{const}(M_2) \) for which \( \langle T_2, A_2 \rangle \models \beta \). For every such \( \beta \), we guess a polynomial-size proof that \( \langle T_1, M_1, D \rangle \models \beta \). If all of our polynomially many guesses succeed, the procedure returns yes; else it returns no. By grouping the guesses together, we obtain an NP decision procedure.

The NP lower bound is by reduction from the NP-complete CQ containment problem: given two CQs \( q_1, q_2 \) both having a single answer variable \( x \), we have that \( q_1 \) is contained in \( q_2 \) just in the case that \( \langle 0, \{ q_2 \rightarrow A(x) \} \rangle \models_{A(x)} \langle 0, \{ q_1 \rightarrow A(x) \} \rangle \), where \( A \not\models \text{sig}(q_1) \cup \text{sig}(q_2) \).

Finally, we consider entailment with respect to entire classes of queries. We show that testing CQ-entailment is significantly more difficult than for single CQs. Both the upper and lower bounds use recent results on KB query inexpressibility (Botoeva et al. 2014; 2015).

**Theorem 11.** For DL-Lite_R, the CQ-entailment problem between OBDA specifications is EXPTIME-complete for linear, GAV and GLAV mappings.

**Proof.** We start with the proof of the membership results. Consider OBDA specifications \( \Gamma_1 = \langle T_1, M_1 \rangle \) and \( \Gamma_2 = \langle T_2, M_2 \rangle \). By Theorem 4, \( \Gamma_1 \models_\Sigma \Gamma_2 \) if and only if \( \Gamma_1 \models_1 \Gamma_2 \) and \( \langle T_2, A_2 \rangle \models q(\vec{a}) \) for every query \( q(\vec{a}) \) and database-ABox pair \( (D, A_2) \) satisfying the conditions of the theorem statement. We know that testing \( \Gamma_1 \models_1 \Gamma_2 \) can be done in NP (Theorem 8). To decide whether the second property holds, we consider each of the single exponentially many database-ABox pairs \( (D, A_2) \), and for every such pair, we set \( N_{D, A_2} = 2|T_2| \cdot |T_2| \cdot (|M_1|+|A_2|+2|T_2| \cdot (|D|+|M_1|))^2 \).

By Theorem 6 and Proposition 3, we know that \( \langle T_2, A_2 \rangle \models q(\vec{a}) \) implies \( \langle T_1, M_1, D \rangle \models q(\vec{a}) \) with \( \vec{a} \subseteq \text{const}(D) \cup \text{const}(M_2) \) iff there is a subset \( A_1 \subseteq A_{M_1, D} \) with \( |A_1| \leq N_{D, A_2} \) such that the KB \( K_1 = \langle T_1, A_1 \rangle \) (\( \Sigma, T \))-entails the KB \( K_2 = \langle T_2, A_2 \rangle \), where \( \Sigma = \text{sig}(K_2) \) and \( Y = \text{const}(A_2) \setminus \text{const}(M_2) \). The latter
check can be performed in EXPTIME, as proven in (Botoeva et al. 2015). We thus obtain an EXPTIME procedure by iterating over all exponentially many pairs \((D, A_2)\) and ABoxes \(A_1 \subseteq A_{M_1, D}\) with \(|A_1| \leq N_{D, A_2}\), performing the KB query entailment check for every such combination, and outputting yes if for every pair \((D, A_2)\), there is some ABox \(A_1\) for which the entailment check succeeds.

Our lower bound also exploits recent results on query inseparability of DL-Lite\(R\) KBs. In (Botoeva et al. 2014), the following problem is shown to be EXPTIME-complete: given DL-Lite\(R\) TBoxes \(T_1\) and \(T_2\) that are consistent with the ABox \(\{A(c)\}\), decide whether the certain answers for every CQ \(q\) w.r.t. \((T_2, \{A(c)\})\) are contained in those for \((T_1, \{A(c)\})\) for every CQ \(q\) with \(\text{sig}(q) \subseteq \text{sig}(T_2)\). To reduce this problem to the CQ-entailment problem for OBDA specifications, we take the following linear mapping which populates the concept \(A\) with all constants appearing in the unary database relation \(A'\) (refer to the proof of Theorem 9 for the definition of \(\text{pop}\): \(M_1 = M_2 = \text{pop}(\{A'\}, \{A\})\)). To complete the proof, we prove in the appendix that \((T_1, M_1) \models_{CQ} (T_2, \{A(c)\})\) iff \((T_2, \{A(c)\}) \models q(a) \iff (T_1, \{A(c)\}) \models q(a)\) for every CQ \(q\) with \(\text{sig}(q) \subseteq \text{sig}(T_2)\).

The complexity of CQ-entailment is significantly lower for DL-Lite\(core\) and DL-Lite\(RDFS\):

**Theorem 12.** For DL-Lite\(RDFS\), the CQ-entailment problem between OBDA specifications is \(\Pi_2^p\)-complete for linear mappings and NP-complete for GAV and GLAV mappings; for DL-Lite\(core\), the problem is in PTIME for linear mappings, NP-complete for GAV mappings, and in \(\Pi_2^p\) for GLAV mappings.

**Proof.** We give the arguments for DL-Lite\(core\) (the rest of the proof is in the appendix). For linear mappings, membership in PTIME easily follows from the fact that, by Theorem 4, to check \((T_1, M_1) \models_{CQ} (T_2, M_2)\) it is sufficient to construct in polynomial time each of the databases \(D\) from the theorem statement, together with the corresponding ABoxes \(A_{M_1, D}\) and \(A_2\), and then check whether the KB \((T_1, A_{M_1, D})\) (\(\text{sig}(K_2), \text{const}(A_2)\))-entails the KB \((T_2, A_2)\) (note that since we consider linear mappings, \(\text{const}(A_2) = \emptyset\)). We return yes if all of these checks succeed. Since KB query entailment without constant restrictions was proven to be in PTIME for DL-Lite\(core\) KBs (Botoeva et al. 2014), membership in PTIME follows.

For GAV mappings, NP-hardness can be proven by an easy reduction of the CQ containment problem. As for the upper bound, we know from Theorem 4 that to check \((T_1, M_1) \models_{CQ} (T_2, M_2)\) it is sufficient to check a polynomial number of database-ABox pairs \((D, A_2)\). For each such pair \((D, A_2)\), we guess a subset \(A' \subseteq A_{M_1, D}\) (together with a polynomial proof that every \(a \in A'\) belongs to \(A_{M_1, D}\)) and then check in PTIME that \((T_1, A')\) (\(\text{sig}(K_2), \text{const}(A_2)\))-entails the KB \((T_2, A_2)\), returning yes if all checks succeed. By grouping all of the above guesses into a single one, we obtain an NP decision procedure.

Finally, in the case of GLAV mappings, we can show membership in \(\Pi_2^p\) as follows. By Theorem 4, \((T_1, M_1) \models_{CQ} (T_2, M_2)\) iff \((T_1, M_1) \models_{\bot} (T_2, M_2)\) (which can be checked in NP by Theorem 8) and for every pair \((D, A_2)\) constructed according to the theorem statement, we have \((T_2, M_2) \models q(\bar{a}) \iff (T_1, M_1, D) \models q(\bar{a})\) for every Boolean CQ \(q(\bar{a})\) with \(\bar{a} \subseteq \text{const}(D) \cup \text{const}(M_2)\). Because of Theorem 6, we know that the latter holds iff there exists a subset \(A_1 \subseteq A_{M_1, D}\) whose size is bounded polynomially in \(|D|, |A_2|, |T_2|, |M_1|\), and \(|M_1|\) such that the KB \(K_1 = (T_1, A_1)\) (\(\Sigma, \Upsilon\))-entails the KB \(K_2 = (T_2, A_2)\), where \(\Sigma = \text{sig}(K_2)\) and \(\Upsilon = \text{const}(A_2) \setminus \text{const}(T_2)\). Since KB entailment with both signature and constant restrictions is NP-complete for DL-Lite\(core\) KBs (Botoeva et al. 2015), it is in NP to decide whether such a subset \(A_1\) exists. We can thus define a \(\Pi_2^p\) procedure which ranges over all possible pairs \((D, A_2)\), and for each such pair, uses an NP oracle to check both \((T_1, M_1) \models_{CQ} (T_2, M_2)\) and that there exists a subset \(A_1 \subseteq A_{M_1, D}\) with the required properties.

Our final result shows that IQ-entailment has the same complexity as single IQ-entailment. The proof exploits Theorem 5 and proceeds similarly to the proof of Theorem 10.

**Theorem 13.** For DL-Lite\(RDFS\), DL-Lite\(core\), and DL-Lite\(R\), the IQ-entailment problem between OBDA specifications is NP-complete for both GAV and GLAV mappings, and is \(\Pi_2^p\)-complete for linear mappings.

### 6 Conclusion and Future Work

In this paper, we have introduced notions of query-based entailment of OBDA specifications and have analyzed the complexity of checking query-based entailment for different classes of queries and mappings and for TBoxes formulated in DL-Lite\(R\) and its sublogics DL-Lite\(core\) and DL-Lite\(RDFS\).

The present work constitutes only a first step towards a full analysis of query-based forms of comparing OBDA specifications and can be extended in several directions:

- First, we could introduce a query signature and only test entailment for queries formulated in the given signature, as has been done for TBox and KB query inseparability (Botoeva et al. 2014). In fact, for GAV mappings (which are the most commonly used in practice), our complexity upper bounds continue to hold in the presence of a query signature. We leave open whether this is the case for CQ-entailment with GLAV mappings.

- Second, it would be interesting to extend the computational analysis of query entailment to other DLs beyond DL-Lite\(R\). We expect that the techniques from this paper can be extended to handle Horn versions of DL-Lite, but that fundamentally new ideas will be needed for DLs allowing for role functionality or recursion.

- Third, other forms of mapping beyond GAV and GLAV could be analyzed. In particular, we would like to see whether decidability of query entailment is preserved if we add some restricted form of inequality or negation to the mapping bodies. This is especially relevant in the context of R2RML (Das, Cyganiak, and Sundara 2012), the W3C language for mapping relational databases to RDF. The query-based comparison of specifications based upon R2RML mappings is in general undecidable, since
R2RML can express arbitrary SQL queries over the database. It would thus be interesting to identify fragments for which query-based entailment is decidable.

- Fourth, to explore the impact of restricting the set of possible databases, we could extend the computational analysis to database schemas with integrity constraints.
- Finally, it would be interesting to develop and experiment practical algorithms for comparing OBDA specifications using query-based notions of entailment.

7 Acknowledgements

This work was partially supported by the EU (FP7 project Optique, grant n. FP7-318338) and the French National Research Agency (ANR JCJC grant n. ANR-12-JS02-007-01).

References


A Proof details for Section 4

We complete the proof of Theorem 4.

Proof of Theorem 4 (continued). Suppose that $\Gamma_1 \models \Gamma_2$ and $\langle T_2, A_2 \rangle \models q(\vec{a})$ implies $\langle T_1, M_1, D \rangle \models q(\vec{a})$ for every CQ $q$ and every database $D$ and ABox $A_2 \subseteq A_{M_2, D}$ obtained as in the theorem statement. Let us further suppose that $\langle T_2, M_2, D_0 \rangle \models q_0(\vec{a})$, where $D_0$ is another database. The first possibility is that $\langle T_2, M_2, D_0 \rangle \models \perp$, in which case $\Gamma_1 \models \Gamma_2$ yields $\langle T_1, M_1, D_0 \rangle \models \perp$ and thus, $\langle T_1, M_1, D_0 \rangle \models q_0(\vec{a})$. The other possibility is that $\langle T_2, M_2, D_0 \rangle \models q_0(\vec{a})$ and $\langle T_2, M_2, D_0 \rangle \not\models \perp$. If $\langle T_1, M_1, D_0 \rangle \models \perp$, we are done.

Otherwise, we are in the case in which both $\langle T_2, M_2, D_0 \rangle \not\models \perp$ and $\langle T_1, M_1, D_0 \rangle \not\models \perp$. The beginning of the argument for this case was given in the body of the text. We complete the proof by showing that $\langle T_1, M_1, D_0 \rangle \models q_i(\vec{b}_i)$, for every $1 \leq i \leq n$.

We begin by establishing the following claim, which shows that only a small number of assertions in $A_{M_2, D_0}$ are needed to derive $q_i(\vec{b}_i)$.

Claim. For every $1 \leq i \leq n$, there exists a mapping assertion $m_i \in M_2$ and a homomorphism $h_i$ of $\text{body}(m_i)$ into $D_0$ such that $\langle T_2, m_i \rangle \models \langle T_2, h_i(\text{head}(m_i)) \rangle \models q_i(\vec{b}_i)$.

Proof of claim. First suppose that $q_i$ is of the first type. Then $q_i = \alpha$ for some atom $\alpha \in q(i)$ such that all variables in $\alpha$ are mapped by $\pi$ to constants from $D_0$ or $M_2$. It is a well-known property of DL-Lite$\mathbb{K}$ (see e.g., (Calvanesi et al. 2007)) that there must exist an assertion $\beta \in A_{M_2, D_0}$ such that $\langle T_2, \{\beta\} \rangle \models \pi(\alpha)$. Since $\beta \in A_{M_2, D_0}$, there is some mapping $m_i \in M_2$ and a homomorphism $h_i$ of $\text{body}(m_i)$ into $D_0$ such that $\beta \in \text{gr}(h_i(\text{head}(m_i)))$ and hence such that $\langle T_2, m_i \rangle \models \langle T_2, h_i(\text{head}(m_i)) \rangle \models q_i(\vec{b}_i)$.

Next suppose that $q_i$ is of the second type. We know that for every pair of atoms $\alpha, \alpha' \in q(i)$, there exists a sequence of atoms $\alpha = \alpha_0, \alpha_1, \ldots, \alpha_{\ell-1}, \alpha_{\ell} = \alpha'$ such that for every $0 \leq j < \ell$, there exists $v_j \in \text{vars}(\alpha_j) \cap \text{vars}(\alpha_{j+1})$ such that $h_{\ell}(v_j) \in (\Delta^{\text{2}} \setminus \text{const}(A_{M_2, D_0})) \cup \text{const}^3(A_{M_2, D_0})$.

We then note that an element $aw \in (\Delta^{\text{2}} \setminus \text{const}(A_{M_2, D_0}))$ is only connected via roles to elements of the form $aw'$ (with $w'$ possibly empty), and constants from $\text{const}^3(A_{M_2, D_0})$ can only be connected via roles to constants that occur in the same mapping head. It follows that there must exist a mapping assertion $m_i \in M_2$ and a homomorphism $h_i$ of $\text{body}(m_i)$ into $D_0$ such that all constants in $\{a \mid \exists v \in \text{vars}(q(i)), h(v) = aw \}$ appear in $\text{gr}(h_i(\text{head}(m_i)))$.

Finally, consider the third case, in which $q_i$ contains no constants and none of its variables are mapped to constants in $A_{M_2, D}$. In this case, we know that $\{h(v) \mid v \in q_i\}$ forms a connected subset of $\Delta^{\text{2}} \setminus \text{const}(A_{M_2, D_0})$, and thus there must exist a constant $a \in \text{const}(A_{M_2, D_0})$ and a role $R$ such that every element in $\{h(v) \mid v \in q_i\}$ takes the form $aRw$ for some (possibly empty) word $w$.

Let $m_i$ and $h_i$ be as in the preceding claim, and let $m_i'$ be obtained from $m_i$ by identifying all pairs of frontier variables $y, z$ such that $h_i(y) = h_i(z)$. By construction, $h_i$ is an injective homomorphism of $\text{body}(m_i')$ into $D_0$, and $\langle T_2, \text{gr}(h_i(\text{head}(m_i'))) \rangle \models q_i(\vec{b}_i)$.

We set $D_{m_i'} = \text{gr}(\text{body}(m_i'))$, and we let $f_i$ be the natural isomorphism from $\text{body}(m_i')$ into $D_{m_i'}$ that maps each variable in $\text{body}(m_i')$ to the corresponding constant in $D_{m_i'}$. Observe that $D_{m_i'} = f_i(\text{body}(m_i'))$. We define two further functions relating constants in $D_0$ and $D_{m_i'}$:

- a function $g_i$, mapping the constants in $D_{m_i'}$ to constants in $D_0$ defined by setting $g_i(c) = h_i(f_i^{-1}(c))$;
- a function $g_i'$, mapping the constants in $D_0$ that occur in the image of $h_i$ to constants in $D_{m_i'}$ defined by setting $g_i'(c) = f_i(h_i^{-1}(c))$.

Note that this definition relies upon the injectivity of $h_i$ and $f_i$, which ensures that $h_i^{-1}$ and $f_i^{-1}$ are well defined.

We can extend $g_i'$ to all of the constants occurring in $\text{gr}(h_i(\text{head}(m_i'))) \setminus \text{body}(m_i')$ by mapping the unique constant in $\text{gr}(h_i(\text{head}(m_i'))) \cap \text{const}(D_{m_i'})$ to the constant in $\text{gr}(f_i(\text{head}(m_i'))) \setminus \text{body}(m_i')$. Let $g_i''$ be the natural isomorphism $h_i(\text{head}(m_i'))$ into $\text{gr}(f_i(\text{head}(m_i')))$. From $\langle T_2, \text{gr}(h_i(\text{head}(m_i'))) \rangle \models q_i(\vec{b}_i)$ and the fact that $g_i''$ is a homomorphism of $\text{gr}(h_i(\text{head}(m_i'))) \setminus \text{body}(m_i')$ into $\text{gr}(f_i(\text{head}(m_i')))$. If $\langle T_2, \text{gr}(f_i(\text{head}(m_i'))) \rangle \models q_i(\vec{b}_i)$, and hence $\langle T_2, M_2, D_{m_i'} \rangle \models q_i(g_i''(\vec{b}_i))$. By construction, $D_{m_i'}$ satisfies the conditions of the theorem statement, so we can apply our initial assumption to obtain $\langle T_1, M_1, D_{m_i'} \rangle \models q_i(g_i'(\vec{b}_i))$.

As $f_i^{-1}(D_{m_i'}) = \text{body}(m_i')$ and $h_i(\text{body}(m_i')) \subseteq D_0$, it follows that $g_i(D_{m_i'}) \subseteq D_0$. In other words, $g_i$ defines a homomorphism of $D_{m_i'}$ into $D_0$. It follows that $\langle T_1, M_1, D_0 \rangle \models q_i(g_i''(\vec{b}_i))$. Since $g_i(g_i''(\vec{b}_i)) = \vec{b}_i$, this yields $\langle T_1, M_1, D_0 \rangle \models q_i(\vec{b}_i)$.

Proof of Theorem 5. We can use a similar argument to that used in the proof of Theorem 4. The main differences can be summarized as follows:

- We do not need to break down the instance query $q_0$ into subqueries, as it consists of a single atom, all of whose variables are mapped to constants from $D_0$ or $M_2$.
- We prove the following stronger claim: there exists a mapping assertion $m \in M_2$, an atom $\alpha \in m$, and a homomorphism $h$ of $\text{body}(m)$ into $D_0$ such that $\langle T_2, \text{gr}(h(\alpha)) \rangle \models q_0(\vec{a})$.
- Since we are only concerned with the head atom $\alpha$, we let $m'$ be the mapping obtained by unifying the at most two variables occurring in $\alpha$, if they are mapped to the same constant by $h$.

The latter point ensures that the database $D_{m'}$ is of the required form.
In preparation for the proof of Theorem 6, we introduce some additional notation and recall some facts about the canonical model construction.

The notation \( \leq \) will be used to indicate the prefix relation between words (or domain elements): \( u < u' \) means that there exists a (possibly empty) word \( u'' \) such that \( u' = uu'' \); for the strict prefix relation (when \( u'' \) must be non-empty), we use \( < \) instead. Note that if \( e, e' \in \Delta^{T_{A}} \) and \( e \leq e' \), then \( e' \) is a proper subset of \( e \). Thus, the elements of \( \Delta^{T_{A}} \) can be naturally viewed as a set of trees. We call \( w \in \Delta^{T_{A}} \) an ancestor of \( w' \in \Delta^{T_{A}} \) if \( w < w' \); \( w \) is the parent of \( w' \) if \( w' = wR \).

With every element \( e \in \Delta^{T_{A}} \) is naturally associated a subinterpretation \( \mathcal{I}^{T_{A}} | e \) obtained by restricting \( \mathcal{I}^{T_{A}} \) to those elements \( e' \) of \( \Delta^{T_{A}} \) such that \( e \leq e' \). Observe that if \( e = awRw' \) and \( e = bwRw'' \) are elements of \( \Delta^{T_{A}} \) that end with the same role \( R \), then the function \( f \) mapping \( awRw' \) to \( bwRw'' \) defines an isomorphism from \( \mathcal{I}^{T_{A}} | e \) to \( \mathcal{I}^{T_{A}} | e' \).

In the following proof, given a (partial) function \( f \), we will use \( \text{dom}(f) \) to refer to the domain of \( f \) (i.e., the elements for which \( f \) is defined), and we will use \( \text{range}(f) \) to refer to the range of \( f \).

For the sake of readability, we provide the full proof of Theorem 6 (including the parts that were included in the proof sketch in the main text).

**Proof of Theorem 6.** Suppose that \( A_{2} \subseteq A_{M_{2}, D} \) and that \( \langle T_{2}, A_{2} \rangle \models q(a) \). Then \( \langle T_{1}, A_{1}, D \rangle \models q(a) \) for every Boolean CQ \( q(a) \) such that \( a \in \text{const}(D) \cup \text{const}(M_{2}) \).

First consider the case in which \( \langle T_{1}, A_{1}, D \rangle \models \perp \), and hence \( \langle T_{2}, A_{2} \rangle \models \perp \). It is well known that if an ABox is inconsistent with a DL-Lite \( R \) TBox, then there exists a subset of the ABox of cardinality at most two that is inconsistent with the TBox. We can thus extract a subset \( A_{1} \subseteq A_{M_{1}, D} \) such that \( |A_{1}| \leq 2 \) and \( \langle T_{1}, A_{1} \rangle \models \perp \).

Let us now consider the more interesting case in which \( \langle T_{1}, A_{1}, D \rangle \models \perp \). By definition, \( \langle T_{2}, A_{2} \rangle \models \perp \), so \( \langle T_{2}, A_{2} \rangle \models \perp \). We let \( I_{2} \) and \( I_{3} \) be the canonical models of \( \langle T_{2}, A_{2}, D \rangle \) and \( \langle T_{2}, A_{2}, D \rangle \) respectively. By assumption, \( \langle T_{2}, A_{2} \rangle \models q(a) \) implies \( \langle T_{1}, A_{1}, D \rangle \models q(a) \) for every Boolean CQ \( q(a) \) with \( a \in \text{const}(D) \cup \text{const}(M_{2}) \).

It follows from results in (Botoeva et al. 2015) that the latter property is true just in the case that there is a homomorphism from \( I_{2} \) to \( I_{3} \) that is the identity on the constants from \( D \) and \( M_{2} \). More precisely, there is a function \( h : \Delta^{T_{2}} \rightarrow \Delta^{T_{1}} \) such that:

- \( h(e) = e \) for every \( e \in \text{const}(D) \cup \text{const}(M_{2}) \)
- \( e \in A^{T_{3}} \) implies \( h(e) \in A^{T_{2}} \)
- \( (e, e') \in P^{T_{3}} \) implies \( (h(e), h(e')) \in P^{T_{2}} \)

Note that the constants in \( \text{const}(A_{M_{2}, D}) \) need not be mapped to themselves.

Our objective is to show how to use \( h \) to construct a homomorphism \( g \) from \( I_{2} \) to \( I_{1} \) with the property that the set \( \Theta = \{ a | aw \in \text{range}(g) \} \) has size polynomially bounded in the size of \( D, A_{3}, T_{2}, \) and \( M_{1} \) (more precisely: of size at most \( |A_{2}| \cdot |M_{2}| + 2 |D| \cdot |T_{2}| \cdot |M_{1}| \)). This will allow us to identify a small subset of \( A_{M_{1}, D} \) that is sufficient to infer all of the queries involving constants from \( D \) or \( M_{2} \) that are entailed by \( \langle T_{2}, A_{2} \rangle \). To construct the desired homomorphism \( g \), we will define a sequence \( g_{0}, g_{1}, g_{2} \ldots \) of partial homomorphisms from \( I_{2} \) to \( I_{1} \) such that \( g_{i+1} \) extends \( g_{i} \) for every \( i \geq 0 \) (that is, \( \text{dom}(g_{i}) \subseteq \text{dom}(g_{i+1}) \) and \( g_{i+1}(e) = g_{i}(e) \) for all \( e \in \text{dom}(g_{i}) \)). We will ensure that for every \( e \in \Delta^{T_{2}} \), there is some \( i \) such that \( e \in \text{dom}(g_{i}) \), which allows us to take \( g \) to be the limit of this sequence. In what follows, we will use \( \Theta_{1} \) to refer to the set \( \{ a | aw \in \text{range}(g_{i}) \} \).

For the initial partial homomorphism \( g_{0} \), we set \( \text{dom}(g_{0}) \) equal to

\[
\text{const}(A_{2}) \cup \{ awR \mid \text{const}(D) \cup \text{const}(M_{1}) \} \quad \text{for all } e < awR
\]

and set \( g_{0}(e) = h(e) \) for every \( e \in \text{dom}(g_{0}) \). Intuitively, \( g_{0} \) copies the value of \( h \) for all constants in \( A_{2} \) as well as for all elements \( e \) such that every ancestor of \( e \) is mapped to an element of \( \Delta^{T_{1}} \in \text{const}(D) \cup \text{const}(M_{1}) \). Because of the way we defined \( \text{dom}(g_{0}) \) and the fact that roles can only link constants from \( \text{const}(A_{M_{2}, D}) \) to constants issuing from the same grounded mapping head, we know that \( g_{0}(b) = h(b) = aw, bw \in \text{dom}(g_{0}) \), and \( g_{0}(bw) = h(bw) = a'u' \), then there exists a mapping assertion \( m \in M_{1} \) such that \( D \models \exists g_{0}. \text{body}(m(a)) \) and \( a', e' \in \text{const}(g_{0}(\text{head}(m(a)))) \). Thus, \( |\Theta_{0}| \leq |A_{2}| \cdot |M_{1}| \).

At each stage \( i \geq 1 \), we will extend \( g_{i-1} \) to some additional elements from \( \Delta^{T_{2}} \setminus \text{const}(A_{2}) \). The idea is to exploit the regularity of canonical models in order to identify subtrees of \( I_{2} \) with similar structure and to modify the homomorphism \( h \) so that similar subtrees are mapped in the same way into \( I_{1} \). We will use a table MapTrees to keep track of how we have mapped some representative subtrees and to reuse this information when extending our partial homomorphisms. Formally, MapTrees will contain tuples of the form \( (d, R, W, \sigma, e) \) where:

- \( d \) is a constant from \( \text{const}(D) \cup \text{const}(M_{1}) \)
- \( R \in N_{R}^{k} \cap \text{sig}(T_{2}) \)
- \( W \) is a (possibly infinite) prefix-closed set of words over the alphabet \( N_{R}^{k} \cap \text{sig}(T_{2}) \)
- \( \sigma \) is a function mapping words in \( W \) to elements in \( \Delta^{T_{1}} \)
- \( e \) is an element from \( \Delta^{T_{2}} \)

where we use \( N_{R}^{k} \) as an abbreviation for \( \{ P, P^{-} | P \in N_{R} \} \). Intuitively, the tuple \( (d, R, W, \sigma, e) \) means that when the parent of an element \( e' \in \Delta^{T_{2}} \) is mapped to \( d \) and the final letter in \( e' \) is \( R \), then we need to extend the homomorphism to all elements in \( \{ e'w \mid w \in W \} \) by mapping \( e'w \) to \( \sigma(w) \). The final argument \( e \) indicates that \( \sigma \) copies information from the subtree rooted at \( e \) (this will be formalized below). Like \( g_{0} \), the table MapTrees will be constructed in stages; we will use MapTrees, to denote the state of MapTrees at the end of stage \( i \). We start with an empty table (MapTrees\(_{0} = \emptyset \)), and at each stage, we will add zero or one new tuples (MapTrees\(_{i} \subseteq \text{MapTrees}, \) and |MapTrees\(_{i} \setminus \text{MapTrees}_{i-1} | \leq 1 \).
For correctness and to achieve the required bound on the cardinality of the sets $\Theta_i$, we will ensure that the following conditions are satisfied by $g_i$ and $\text{MapTrees}_i$ for every $i \geq 0$:

**(P1)** If $\text{MapTrees}_i \setminus \text{MapTrees}_{i-1} = \{(d, R, W, \sigma, e)\}$, then:
- $\text{MapTrees}_{i-1}$ does not contain any tuple $(d, R, \ldots)$
- $e \in \text{dom}(g_{i-1})$ and $g_{i-1}(e) = d$
- for every $w \in W$, $eRw \in \Delta^{T_2} \setminus \text{dom}(g_{i-1})$, $eRw \in \text{dom}(g_i)$, and $g_i(eRw) = h(eRw) = \sigma(e)$
- $\text{dom}(g_i) \setminus \text{dom}(g_{i-1}) = \{eRw \mid w \in W\}$

**(P2)** If $\text{MapTrees}_i = \text{MapTrees}_{i-1}$, then there exists $(d, R, W, \sigma, e) \in \text{MapTrees}_{i-1}$ and $e'R \in \Delta^{T_2} \setminus \text{dom}(g_{i-1})$ such that:
- $e' \neq e$, $e' \in \text{dom}(g_{i-1})$, and $g_{i-1}(e') = d$
- for every $w \in W$, $e'Rw \in \Delta^{T_2}$, $e'Rw \in \text{dom}(g_i)$, $g_i(e'Rw) = \sigma(w)$
- $\text{dom}(g_i) \setminus \text{dom}(g_{i-1}) = \{e'Rw \mid w \in W\}$

**(P3)** For every tuple $(d, R, W, \sigma, e) \in \text{MapTrees}_i$ and every leaf $w \in W$ (that is, there is no $w' \in W$ with $w < w'$), we have $\sigma(w) \in \text{const}(D) \cup \text{const}(M_2)$.

**(P4)** $g_i$ is a partial homomorphism from $I_2$ to $I_1$ that is the identity on constants from $D$ and $M_2$.

**(P5)** For every function $\sigma$ appearing in the fourth column of $\text{MapTrees}_i$, there exists a mapping assertion $m \in M_1$ such that $D = \{x \mid \exists \bar{y}. \text{body}(m)\}$ and $\{a \mid a \in \text{range}(\sigma)\} \subseteq \text{const}(\text{gr}(\text{head}(m)))$.

Observe that the preceding properties are trivially satisfied by the initial function $g_0$ and table $\text{MapTrees}_0$.

We now explain how to construct $g_i$ and $\text{MapTrees}_i$, starting from $g_{i-1}$ and $\text{MapTrees}_{i-1}$. We first select an element $awR \in \Delta^{T_2}$ such that $aw \in \text{dom}(g_{i-1})$, $awR \not\in \text{dom}(g_{i-1})$, and there is no $aw'R \in \Delta^{T_2}$ with $aw'R \in \text{dom}(g_{i-1})$ and $aw'R \not\in \text{dom}(g_{i-1})$ such that either:
- $|\rho(aw')| < |aw|$, or
- $\rho(aw') = |aw|$ and there exists $j$ such that $aw' \in \text{dom}(g_j)$ but $aw \not\in \text{dom}(g_j)$.

It follows from properties (P1)-(P3) and the definition of $g_0$ that $g_{i-1}(aw) \in \text{const}(D) \cup \text{const}(M_1)$. There are two cases to consider:

**Case 1:** $\text{MapTrees}_{i-1} \not\subset \text{MapTrees}_{i}$ does not contain any tuple of the form $(g_{i-1}(aw), R, \sigma, e)$

We define a set of $W^R_{aw}$ of words as follows:

$W^R_{aw} = \{w' \mid \text{for every } w'' < w', h(awRw') \not\in \text{const}(D) \cup \text{const}(M_2)\}$

and we set $\sigma^R_{aw}(aw') = h(awRw')$ for every $w' \in W^R_{aw}$. We let $\text{MapTrees}_i$ be the result of adding the tuple $(g_{i-1}(aw), R, W^R_{aw}, \sigma^R_{aw})$ to $\text{MapTrees}_{i-1}$, and we extend $g_i$ to the set of elements $\{awRw' \mid w' \in W^R_{aw}\}$ by setting $g_i(awRw') = h(awRw')$ (equivalently, $g_i(awRw') = \sigma_{aw}(aw')$). By construction, $g_i$ and $\text{MapTrees}_i$ satisfy properties (P1) and (P3), and (P2) trivially holds. It remains to show that (P5) and (P4) hold.

For property (P4), we begin by showing that $g_{i-1}(aw) = h(aw)$. Suppose for a contradiction that this is not the case. Then we know that $aw \not\in \text{dom}(g_0)$ (since $g_0(e) = h(e)$ for all $e \in \text{dom}(g_0)$). Let $j$ be such that $aw \in \text{dom}(g_j)$ \setminus \text{dom}(g_{j-1}).$ We must have $\text{MapTrees}_i = \text{MapTrees}_{j-1}$, since otherwise, by property (P1), we would have $g_j(aw) = h(aw)$, hence $g_{j-1}(aw) = h(aw)$. By properties (P1) and (P2), there exist $(c, S, W, \sigma, e) \in \text{MapTrees}_{j-1} \setminus \Delta^{T_2} \setminus \text{dom}(g_{j-1})$ such that:
- $aw \in \text{dom}(g_{j-1})$ and $g_{j-1}(aw) = c$
- there exists $w' \in W$ such that $aw = auSu'$
- $g_j(aw) = g_j(auSu') = \sigma(w')$

Since $(c, S, W, \sigma, e) \in \text{MapTrees}_{j-1}$ and $w' \in W$, the element $uSu'$ belongs to $\text{dom}(g_{j-1})$ by (P1), whereas $aw = auSu' \not\in \text{dom}(g_{j-1})$. Moreover, because of the way we select elements for treatment, we know that $|e| \leq |au|$, hence $|eSu'| \leq |auSu'| = |aw|$. It follows that at stage $i$, our strategy for selecting elements would lead us to choose $eSu'$ rather than $aw$, a contradiction. We have thus established that $g_{j-1}(aw) = h(aw)$. Since $aw$ is the only element in $\text{dom}(g_{j-1})$ that is connected via a role to an element in $\text{dom}(g_j)$ \setminus $\text{dom}(g_{j-1})$, $g_{j'}(eSu') = h(eSu')$ for every $eSu' \in \text{dom}(g_j) \setminus \text{dom}(g_{j-1})$, and $h$ is a homomorphism from $I_2$ to $I_1$, it follows that $g_i$ is a partial homomorphism from $I_2$ to $I_1$. Thus, $g_i$ satisfies property (P4).

Finally, we consider property (P5). Recall that we have defined $\sigma^R_{aw}(awRw') = \text{body}(m(\bar{a}))$ and $\sigma^R_{aw}(awRw') = h(awRw')$ for every $w' \in W^R_{aw}$. Thus, $\{b \mid b \in \text{range}(\sigma^R_{aw})\} = \{h(awRw') \mid w' \in W^R_{aw}\}$. Since $g_i$ is a partial homomorphism (property (P4)), we know that if two elements $e_1, e_2 \in \{awRw' \mid w' \in W^R_{aw}\}$ are related by a role in $I_2$, then $h(e_1)$ and $h(e_2)$ must also be related by a role in $I_1$.

By the structure of the ABox $A_{M_1}$ and the definition of canonical models, if there is a role between $h(e_1)$ and $h(e_2)$, then one of the following must hold:
- $h(e_2) = h(e_1)T$ for some role $T$
- $h(e_1) = h(e_2)T$ for some role $T$
- $h(e_1), h(e_2) \in \text{const}(A_{M_1}, D)$ and there exists a mapping assertion $m \in M_1$ such that $D = \{\exists \bar{y}. \text{body}(m)\}$ and $h(e_1), h(e_2)$ occur in a role assertion $\text{gr}(\text{head}(m))$.

Note that in the first two cases, $h(e_1)$ and $h(e_2)$ share the same initial constant. Recall also that by construction, $g_i(awRw') \not\subset \text{const}(D) \cup \text{const}(M_1)$ for every $w' \in W^R_{aw}$ that is not a leaf of $W^R_{aw}$ and by property (P3), $g_i(awRw') \not\subset \text{const}(D) \cup \text{const}(M_1)$ for every leaf $w' \in W^R_{aw}$.

It follows that every path from $h(awRw')$ to $h(awRw')$ with $w' \in W^R_{aw}$ a leaf in $W^R_{aw}$ contains exactly one element from $\text{const}(D) \cup \text{const}(M_1)$, namely, $h(awRw')$.

Earlier in the path, we can move between adjacent elements in the tree rooted at a constant, between constants from $\text{const}^2(A_{M_1}, D)$ that occur in the same grounded mapping head, or at the end of the path, from a constant from $\text{const}^2(A_{M_1}, D)$ to a constant from $\text{const}(D) \cup \text{const}(M_1)$. Thus, must exist a single mapping assertion $m \in M_1$ such that $D = \{\exists \bar{y}. \text{body}(m)\}$ and $\{b \mid h(awRw') = bv \mid w' \in W^R_{aw}\} \subseteq \text{const}(\text{gr}(\text{head}(m)))$. 


Case 2: \( \text{MapTrees}_{-1} \) contains \((g_{i-1}(aw), R, W, \sigma, e)\)

We extend \(g_i\) to set of elements \(\{awRw' \mid w' \in W\}\) by setting \(g_i(awRw') = \sigma(w').\) We do not modify the table: \(\text{MapTrees}_i \subseteq \text{MapTrees}_{i-1}\). Properties \((P1), (P3), \) and \((P5)\) are trivially satisfied for \(g_i\) and \(\text{MapTrees}_i\). For \((P2)\), we note that since \((g_{i-1}(aw), R, W, \sigma, e) \in \text{MapTrees}_{i-1}\) but \(awR \not\in \text{dom}(g_{i-1})\), we must have \(aw \neq e\). The remainder of the first bullet, as well as the other two bullets, are satisfied by construction. To show the remaining property \((P4)\), first note that by property \((P1)\), we have \(g_i(aw) = g_{i-1}(e)\) and \(g_i(awRw') = g_{i-1}(eRw')\) for every \(w' \in W\) since \(g_{i-1}\) satisfies \((P4)\), we know that \(g_{i-1}\) is a partial homomorphism.

To conclude the argument, we recall that the subinterpretation \(I_2|_{awR} \approx I_2|_{eRw'}\) is isomorphic to \(I_2|_{eRw'}\) with isomorphism \(f\) given by \(f(awRw') = ewRw'\).

We have defined the sequence of partial homomorphisms \(g_0, g_1, g_2, \ldots \) from \(I_2\) to \(I_1\). Our selection strategy ensures that for every element \(e \in \Delta^{R_2}\) there exists \(i\) such that \(e \in \text{dom}(g_i)\). Moreover, if \(e \in \text{dom}(g_i)\), then \(g_i(e) = g_{i+1}(e)\) for every \(j \geq i\). We can thus define \(g\) as follows: \(g(e) = g_i(e)\) for some (equivalently, every) \(i\) with \(e \in \text{dom}(g_i)\). Since \(g_i\) is a partial homomorphism, it follows that \(g\) is a (total) homomorphism from \(I_2\) to \(I_1\) that is the identity on \(\text{const}(D) \cup \text{const}(M_2)\).

Let us now show that \(\Theta = \{a \mid aw \in \text{range}(g)\}\) is of size at most \(|\mathcal{A}_2| + 2|\mathcal{T}_2| \cdot |\mathcal{M}_1| \cdot (|D| + |\mathcal{M}_1|)\). As noted earlier, \(|\Theta_0| \leq |\mathcal{A}_2| \cdot |\mathcal{M}_1|\). Let us now consider the size of \(\Theta \setminus \Theta_0 = \{a \mid aw \in \text{range}(g) \setminus \text{range}(g_0)\}\).

We know from properties \((P1)\) and \((P2)\) that for every \(e \in \text{dom}(g_i)\), there exists a tuple \((d, R, W, \sigma, e')\) in \(\text{MapTrees}_i\) such that \(g_i(e) \in \text{range}(\sigma)\). Thus, we need to determine the cardinality of \(\{a \mid aw \in \text{range}(\sigma)\}\) for some tuple \((d, R, W, \sigma, e')\) in \(\text{MapTrees}_i\).

Since there can be at most one tuple of the form \((d, R, W, \sigma, e')\) per \((d, R, W, \sigma, e')\in R\) the total number of different functions \(\sigma\) cannot exceed \(2|\mathcal{T}_2| \cdot (|\text{const}(D)| + |\text{const}(M_1)|)\) (here \(2|\mathcal{T}_2|\) is used to bound the possible values of \(R\)). By property \((P5)\), for every such function \(F\) there exists a mapping assertion \(m \in M_1\) such that \(D \models \exists \bar{y}.\text{body}(m(\bar{a}))\) and \(\{a \mid aw \in \text{range}(\sigma)\} \subseteq \text{const}(gr(\text{head}(m(\bar{a}))))\). The cardinality of the latter set cannot exceed the maximal number of terms appearing in the head of a mapping assertion in \(M_1\), and hence is bounded above by \(|\mathcal{M}_1|\). Putting all of this together yields the desired bound of \(|\mathcal{A}_2| \cdot |\mathcal{M}_1| + 2|\mathcal{T}_2| \cdot |\mathcal{M}_1| \cdot (|D| + |\mathcal{M}_1|)\).

Now we let \(A_{M_2, \alpha}\) be the restriction of \(A_{M, \alpha}\) to the constants in \(\Theta\). To define \(A_1\), for every constant \(c \in \mathcal{A}_0\) such that there exists \(d \in \{c, d\} \in R^{Z_i}\), we pick one such constant \(d\), which we will refer to by \(w(c, R)\). Then the \(\text{ABox}_1\) is defined as follows:

\[ A_1 = A_0 \cup \{R(c, \text{wit}(c, R)) \mid c \in \text{const}(A_0), \exists d R(c, d) \in A_0\} \]

Note that \(R\) may be an inverse role \(r^-\), in which case we use \(R(a, b)\) to refer to the assertion \(r^-(b, a)\). Take \(I_2\) to be the restriction of \(I_2\) to \(\{aw \mid a \in \Theta\}\), and let \(\mathcal{J}\) be the canonical model of the KB \((T_1, A_1)\). It is readily verified, using the definition of canonical models and of the \(\text{ABox}_1\), that the identify function is a homomorphism of \(I_2\) into \(\mathcal{J}\). Since \(g\) is a homomorphism of \(I_1\) into \(I_2\) whose range is contained in \(\{aw \mid a \in \Theta\}\), \(g\) must also be a homomorphism of \(I_1\) into \(I_2\), and hence also into \(\mathcal{J}\). Moreover, \(g\) maps all constants in \(\text{const}(D) \cup \text{const}(M_2)\) to themselves. We can thus conclude that \(\langle I_2, A_2 \rangle \models q(\bar{a})\) if and only if \(\langle I_1, A_1 \rangle \models q(\bar{a})\) for every Boolean CQ \(q(\bar{a})\) with \(\bar{a} \subseteq \text{const}(D) \cup \text{const}(M_2)\).

**B Proof details for Section 5**

We provide details of the argument for the NL upper bound in the proof of Theorem 7.

**Proof of Theorem 7 (continued).**

We consider the case in which \(\Gamma_1 = \langle T_1, M_1 \rangle\), \(\Gamma_2 = \langle T_2, M_2 \rangle\) both have linear mappings, and we explain how to decide in NL whether \(\Gamma_1 \models_{\text{log}} \Gamma_2\). The procedure will consider in turn each of the mapping assertions in \(M_2\), which can be done in logarithmic space by keeping only the id of mapping assertions in memory, rather than the assertions themselves. For each mapping assertion \(m \in M_2\), we do not build the database \(D = gr(\text{body}(m))\) explicitly, but instead exploit the fact that \(D\) has precisely the same structure as \(\text{body}(m)\), which is available in the input. In other words, instead of replacing every variable by a constant, we simply treat the existing variables as constants.

We store in memory (a binary representation of) the atom \(\alpha = \text{head}(m)\); this requires only logarithmic space since the arity of \(\alpha\) is at most two. Our aim is to determine whether \(\Gamma_1, \text{body}(m) \models \alpha\), where the variables in \(\text{body}(m)\) and \(\alpha\) are treated as constants. Because \(T_1\) is a DL-Lite\_TBox, we know that this entailment holds iff there exists a mapping assertion \(m' \in M_1\) and a homomorphism \(h\) from \(\text{body}(m)\) to \(\text{body}(m')\) such that \(\langle T_1, h(\text{head}(m')) \rangle \models \alpha\). Moreover, since we are working with linear mappings, both \(\text{body}(m)\) and \(\text{body}(m')\) consist of a single atom, which means that \(\text{body}(m)\) and \(\text{body}(m')\) must use the same relation name and \(h\) is the unique function such that \(h(\text{body}(m')) = \text{body}(m)\).

We can thus proceed as follows:

1. We guess the (id of a) mapping assertion \(m' \in M_1\).
2. We check in the input that \(\text{body}(m)\) and \(\text{body}(m')\) use the same relation name.
3. We iterate over the positions \(1 \leq i \leq \text{arity}(P)\) and over positions \(i < j \leq \text{arity}(P)\) (by representing the positions in binary, we need only logarithmic space to store the pair \((i, j)\)). If \(\text{body}(m')\) contains the same term in positions \(i\) and \(j\), then we check that the term in position \(i\) of \(\text{body}(m)\) is the same as that in position \(j\).
4. We compute and store binary representations of the (at most two) frontier variables in \(m'\). Then, for each frontier variable \(v\), we scan \(\text{body}(m')\) for the first position \(i\) in which \(v\) occurs, and we memorize the term occurring in position \(i\) of \(\text{body}(m)\).
5. We let \(\beta\) be obtained by taking \(\text{head}(m')\) and replacing each frontier variable \(v\) by the term in \(m\) that we memorized in the previous step. A binary representation of the atom \(\beta\) can be stored in logarithmic space.
6. Finally, since IQ answering in $DL-Lite_R$ is in NL, we can make a call to an NL oracle to verify that $(\langle T_1, \beta \rangle \models \alpha)$. If all of the preceding checks succeeded, we return yes.

Otherwise, we return no.

Note that if the checks in Steps 2 and 3 succeed, then the atom $\beta$ constructed in Step 5 is precisely $h(\text{head}(m'))$. Thus, the procedure correctly checks whether $(\langle \Gamma_1, \text{body}(m) \rangle \models \alpha)$. Moreover, since the procedure runs in logarithmic space with access to an NL oracle and $L^{\text{NL}} = NL$, we obtain an NL process.

The next lemma establishes the claim from the proof of Theorem 9.

**Lemma 1.** $(\emptyset, \mathcal{M}_1) \models_q (\emptyset, \mathcal{M}_2)$ iff $(\langle T, \mathcal{A} \rangle \models q_1 \Rightarrow (\langle T, \mathcal{A} \rangle \models q_2)$ for all $\Sigma$-ABoxes.

**Proof.** First suppose that $(\emptyset, \mathcal{M}_1) \models_q (\emptyset, \mathcal{M}_2)$, and let $\mathcal{A}$ be an ABox such that $\text{sig}(\mathcal{A}) \subseteq \Sigma$ and $(\langle T, \mathcal{A} \rangle \models q_1 \Rightarrow (\langle T, \mathcal{A} \rangle \models q_2$ for all $\Sigma$-ABoxes. Let $D$ be the database instance consisting of the facts in $\mathcal{A}$. Since $\mathcal{M}_2$ contains copy of $\Sigma$ as well as $\{T(x) \rightarrow V(x), F(x) \rightarrow V(x)\}$, it follows that $(\emptyset, \mathcal{M}_2, D) \models q'_2$. Moreover, because $\mathcal{M}_2$ contains $\text{pop}(\Sigma, \Sigma^2)$, the corresponding ABox $\mathcal{A}_{M_2, D}$ contains for every constant $c$ in $D$ (equiv. $A$), every $\Sigma^2$-fact whose only constant is $c$. More precisely, for every such constant $c$, the ABox $\mathcal{A}_{M_2, D}$ contains $A^2(c)$ for every concept name $A \in \Sigma$ and $R^2(c, c)$ for every role name $R \in \Sigma$. Using the fact that $q'_2$ is Boolean and only uses predicates from $\Sigma^2$, we can infer that $(\emptyset, \mathcal{M}_2, D) \models q'_2$ and hence that $(\emptyset, \mathcal{M}_2, D) \models q$. By our assumption that $(\emptyset, \mathcal{M}_1) \models_q (\emptyset, \mathcal{M}_2)$, we obtain $(\emptyset, \mathcal{M}_1, D) \models q$, which implies that $(\emptyset, \mathcal{M}_1, D) \models q'_2$. We then observe that $\mathcal{A}_{M_1, D}$ contains $P^2(c)$ iff $A$ contains the corresponding assertion $P(c)$. It follows that the homomorphism witnessing that $(\emptyset, \mathcal{M}_1, D) \models q'_2$ can be reproduced in $\mathcal{A}$ using the original predicates, so $\mathcal{A} \models q_2$.

For the other direction, suppose that for all $\Sigma$-ABoxes, $(\langle T, \mathcal{A} \rangle \models q_1$ implies $(\langle T, \mathcal{A} \rangle \models q_2$. Let $D$ be a database instance such that $(\emptyset, \mathcal{M}_2, D) \models q$. It follows that $(\emptyset, M_2, D) \models q'_1$. Note that since the left-hand sides of mapping assertions in $\mathcal{M}_2$ only use predicates from $\Sigma$, we may assume w.l.o.g. that $\text{sig}(D) \subseteq \Sigma$. Since $\mathcal{M}_2$ contains the facts in $\mathcal{D}$, from $(\emptyset, \mathcal{M}_2, D) \models q'_1$, and the fact that the two inclusions in $T$ simulate the effect of the mapping assertions $T(x) \rightarrow V(x)$ and $F(x) \rightarrow V(x)$, we can infer that $(\langle T, \mathcal{A}_D \rangle \models q'_1$. Applying our assumption, we obtain $(\langle T, \mathcal{A}_D \rangle \models q'_2$, which yields $(\emptyset, \mathcal{M}_1, D) \models q'_2$ because of the mapping assertions in $\text{copy}^2(\Sigma)$. We can also show that $(\emptyset, \mathcal{M}_1, D) \models q'_2$ using the mapping assertions in $\text{pop}(\Sigma, \Sigma^2 \cup \{V\})$, from which we obtain $(\emptyset, \mathcal{M}_1, D) \models q$. as desired.

We give the missing proof of the linear case for Theorem 10.

**Proof of Theorem 10 (continued).** In the case of linear mappings, we can again use Theorem 3 to restrict the number and form of the databases that need to be considered. By Theorem 8, we can check in NL whether $\Gamma_1 \models \bot \Gamma_2$. We can next iterate over all of pairs $(D, A_2)$ constructed in the manner specified in Theorems 3. Observe that because $M_2$ is a linear mapping, every pair $(D, A_2)$ corresponds to taking a mapping assertion $m \in M_2$, possibly unifying the (at most two) frontier variables in $m$, and setting $D = gr(body(m))$ and $A_2 = gr(head(m))$. As in the proof of Theorems 7 and 8, we can work directly with $\text{body}(m)$ and $\text{head}(m)$, rather than introducing new constants, and we do not need to explicitly construct $\text{body}(m)$ but can instead store the id of $m$ and the decision of whether or not to merge the frontier variables. We next iterate (again in logarithmic space) over all (unary or binary) tuples $\alpha \subseteq \text{terms(head}(m))$ of the same arity as $q$. For every such tuple $\alpha$, we make a call to an NL oracle to check whether the IQ $q(\alpha)$ is entailed from $(\langle T_2, \text{body}(m) \rangle$. If the response is yes, then we make a second NL oracle call to check whether $(\langle \Gamma_1, \text{body}(m) \rangle \models q(\alpha)$ (refer to the proof of Theorem 7 for a more detailed description of how to perform this entailment check in NL). The procedure returns no if there is some $m$ obtained from a mapping assertion in $M_2$ by possibly unifying the frontier variables and some tuple $\alpha \subseteq \text{terms(head}(m))$ for which $(\langle T_2, \text{body}(m) \rangle \models q(\alpha)$ and $(\langle \Gamma_1, \text{body}(m) \rangle \models \neg q(\alpha)$; otherwise, it returns yes. The correctness of the procedure is easily verified, and since $L^{\text{NL}} = NL$, we obtain an NL decision procedure.

For the NL lower bound, we reduce concept subsumption to single IQ entailment with linear mappings as follows: $T \models A \subseteq B$ iff $(\langle T, \{A'(x) \rightarrow A(x)\} \models_B (\langle T, \{A'(x) \rightarrow B(x)\})$. The following lemma, which will be used in the proof of Theorem 12, shows that if we consider OBDA specifications based up on $DL-Lite_R$ TBoxes and GAV mappings, then CQ-entailment and IQ-entailment coincide.

**Lemma 2.** Consider OBDA specifications $\Gamma_1 = (\langle T_1, \mathcal{M}_1 \rangle$ and $\Gamma_2 = (\langle T_2, \mathcal{M}_2 \rangle$ where $T_1, T_2$ are $DL-Lite_R$ TBoxes and $\mathcal{M}_1, \mathcal{M}_2$ are GAV mappings. Then $\Gamma_1 \models_{\text{CQ}} \Gamma_2$ iff $\Gamma_1 \models_{\text{IQ}} \Gamma_2$.

**Proof.** By Proposition 2, $\Gamma_1 \models_{\text{CQ}} \Gamma_2$ implies $\Gamma_1 \models_{\text{IQ}} \Gamma_2$. For the other direction, suppose that $\Gamma_1 \models_{\text{IQ}} \Gamma_2$ and $(\langle T_2, \mathcal{M}_2, D) \models q(\alpha)$. Since $T_2$ is formulated in $DL-Lite_R$, we know that it does not contain any existentials on the right-hand-side of inclusions. It must thus be the case that there is a function $h : \text{terms}(q) \rightarrow \text{constr}(\mathcal{A}_{M_2, D})$ that maps every constant to itself and the tuple of answer variables of $q$ to the tuple of constants $\alpha$ and which satisfies $\{h(\alpha) \mid \alpha \in q\} \subseteq \mathcal{A}_{M_2, D}$. From $\Gamma_1 \models_{\text{IQ}} \Gamma_2$, we can infer that $\{h(\alpha) \mid \alpha \in q\} \subseteq \mathcal{A}_{M_1, D}$, and hence that $(\langle \Gamma_1, \mathcal{M}_1 \rangle \models q(\alpha)$. This establishes that $\Gamma_1 \models_{\text{CQ}} \Gamma_2$.

We now complete the proof of Theorem 12 by proving the complexity results for $DL-Lite_{RDFS}$.

**Proof of Theorem 12 (continued).** For linear mappings, NL-hardness is easily obtained by a reduction of concept subsumption in $DL-Lite_{RDFS}$ similar to the previous ones. More precisely, it is immediate to verify that $T \models A \subseteq B$ iff $(\langle T, (T(x) \rightarrow A(x)) \models_{\text{CQ}} (\langle T, (T(x) \rightarrow A(x), T(x) \rightarrow B(x))\). (Note that this same
sat in polynomial time the saturated mappings $\text{DL-Lite}_{\text{RDFS}}$ following NP decision procedure for CQ-entailment for $\text{sat}$ and $\text{GLAV}$ mappings, we let $\text{sat}(M,T)$ denote the saturation of $M$ by $T$, which is obtained by replacing each mapping assertion $m \in M$ by the mapping assertion $m^*$ which has the same body as $m$, but its head contains all atoms in $\text{head}(m)$, plus all atoms that can be inferred from $\text{head}(m)$ using $T$. For example, if $T = \{ A \subseteq B, \exists S \in D, R \subseteq S \}$, we would replace the mapping assertion $P(x,y,z) \rightarrow \exists w.A(x) \land R(x,w) \land S(x,w) \land B(x) \land D(w)$ by $P(x,y,z) \rightarrow \exists w.A(x) \land R(x,w) \land S(x,w) \land B(x) \land D(w)$.

It is not hard to see that for every database $D$, the ABoxes $A_{M,D}$ and $A_{\emptyset,\text{sat}(M,T)}$ are identical (modulo renaming of the constants issuing from the grounding of existential variables). It follows that $(T,M) \equiv_{\text{CQ}} (\emptyset,\text{sat}(M,T))$. Thus, we have $(T_1,M_1) \equiv_{\text{CQ}} (T_2,M_2)$ iff $(\emptyset,\text{sat}(M_1,T_1)) \equiv_{\text{CQ}} (\emptyset,\text{sat}(M_2,T_2))$. We next argue that the latter holds just in the case that $\text{sat}(M_1,T_1) \models_{\text{log}} \text{sat}(M_2,T_2)$. The direction from $\text{sat}(M_1,T_1) \models_{\text{log}} \text{sat}(M_2,T_2)$ to $(\emptyset,\text{sat}(M_1,T_1)) \equiv_{\text{CQ}} (\emptyset,\text{sat}(M_2,T_2))$ is a simple consequence of Proposition 2. For the other direction, suppose that $(\emptyset,\text{sat}(M_1,T_1)) \equiv_{\text{CQ}} (\emptyset,\text{sat}(M_2,T_2))$, and let $m \in \text{sat}(M_2,T_2)$. Let $D_m = g\gamma(\text{body}(m))$ be the database corresponding to the body of $m$, let $q = \text{head}(m)$, and let $b$ be the constants in $D_m$ corresponding to the grounding of the frontier variables of $m$. By construction, we have $(\emptyset,\text{sat}(M_2,T_2),D_m) \models q(b)$, and so applying our assumption, we obtain $(\emptyset,\text{sat}(M_1,T_1),D_m) \models q(b)$. It follows that $\text{sat}(M_1,T_1) \models m$, which yields $\text{sat}(M_1,T_1) \models_{\text{log}} \text{sat}(M_2,T_2)$. We thus obtain the following NP decision procedure for CQ-entailment for $\text{DL-Lite}_{\text{RDFS}}$ TBoxes and GLAV mappings: (i) construct in polynomial time the saturated mappings $\text{sat}(M_1,T_1)$ and $\text{sat}(M_2,T_2)$, (ii) test whether $\text{sat}(M_1,T_1) \models_{\text{log}} \text{sat}(M_2,T_2)$ using the NP procedure described in the proof of Theorem 7, and (iii) return yes iff the procedure in (ii) returned yes.

Proof of Theorem 13. The proof of all the upper bounds follows from Theorem 10 and from the fact that $(T_1,M_1) \models_{\text{IQ}} (T_2,M_2)$ iff $(T_1,M_1) \models_{\alpha} (T_2,M_2)$ for every instance query $\alpha$ over $\text{sig}((T_2,A_{M_2,D}))$. Since the number of such queries $\alpha$ is linearly bounded (and the queries can be enumerated in logarithmic space), it follows that all the upper bounds of Theorem 10 extend to the IQ-entailment problem.

All the NL lower bounds for all the DLs and linear mappings follow from an easy reduction of the problem of concept subsumption in a TBox constituted of inclusions between atomic concepts (such a TBox belongs to all the DLs considered), which is an NL-hard problem.

Finally, it is easy to prove the NP lower bound for the case of empty TBoxes and GAV mappings through a reduction of containment between conjunctive queries: given two Boolean CQs $q_1, q_2, q_3$ is contained in $q_2$ if and only if $(\emptyset,\{q_2 \rightarrow A(a)\}) \models_{\text{IQ}} (\emptyset,\{q_1 \rightarrow A(a)\})$. □