Sound, Complete, and Minimal Query Rewriting for Existential Rules
Mélanie König, Michel Leclère, Marie-Laure Mugnier, Michaël Thomazo

To cite this version:

HAL Id: lirmm-00838791
https://hal-lirmm.ccsd.cnrs.fr/lirmm-00838791
Submitted on 26 Jun 2013

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.
Sound, Complete, and Minimal Query Rewriting for Existential Rules

Mélanie König and Michel Leclère and Marie-Laure Mugnier and Michaël Thomazo
University of Montpellier, France
firstname.lastname@lirmm.fr

Abstract
We address the issue of Ontology-Based Data Access which consists of exploiting the semantics expressed in ontologies while querying data. Ontologies are represented in the framework of existential rules, also known as Datalog+-./. We focus on the backward chaining paradigm, which involves rewriting the query (assumed to be a conjunctive query, CQ) into a set of CQs (seen as a union of CQs). The proposed algorithm accepts any set of existential rules as input and stops for so-called finite unification sets of rules (fus). The rewriting step relies on a graph notion, called a piece, which allows to identify subsets of atoms from the query that must be processed together. We first show that our rewriting method computes a minimal set of CQs when this set is finite, i.e., the set of rules is a fus. We then focus on optimizing the rewriting step. First experiments are reported in the associated technical report.

1 Introduction
In recent years, there has been growing interest in exploiting the semantics expressed in ontologies when querying data, an issue known as ontology-based data access (OBDA). The dominant approach to this issue is based on description logics (DLs), with the most studied DLs in this context being lightweight DLs like DL-Lite and EL families [Baader, 2003; Calvanese et al., 2007] and their Semantic Web counterparts, so-called tractable fragments of OWL2. A newer approach, to which this paper contributes, is based on existential rules. Existential rules have the ability of generating unknown individuals, a feature that has been recognized as crucial in an open-world perspective. These rules are of the form $\text{body} \rightarrow \text{head}$, where $\text{body}$ and $\text{head}$ are conjunctions of atoms (without functions), and variables that occur only in $\text{head}$ are existentially quantified [Baget et al., 2009; 2010; 2011b; Krötzsch and Rudolph, 2011]. They are also known as Datalog± an extension of plain Datalog to database constraints called tuple-generating dependencies [Calì et al., 2008; 2009].

In this paper, we consider knowledge bases (KBs) composed of a set of facts, or data, and of existential rules. We focus on the standard basic queries, namely conjunctive queries (CQs), which can be seen as existentially quantified conjunctions of atoms. The fundamental decision problem associated with query answering can be expressed in several equivalent ways, in particular as a Boolean CQ entailment problem: is a given Boolean CQ logically entailed by a knowledge base?

CQ entailment is undecidable for general existential rules. There has been an intense research effort aimed at finding decidable subsets of rules that provide good tradeoffs between expressivity and complexity of query answering (see [Mugnier, 2011] for a synthesis). Compared to lightweight DLs, these decidable rule fragments are more powerful and flexible. However, the existential rule-based OBDA framework does not come yet with practically usable algorithms, with the exception of very simple classes of rules, which can be seen as slight generalizations of lightweight DLs. In this paper, we undertake a step in this direction.

There are two classical paradigms for processing rules, namely forward chaining and backward chaining, which can both be seen as ways of integrating the rules either into the facts or into the query: forward chaining uses the rules to enrich the facts and the query is entailed by the KB if it maps by homomorphism to the enriched facts, while backward chaining uses the rules to rewrite the query in several ways and the initial query is entailed by the KB if a rewritten query maps to the initial facts. In the context of large data, the obvious advantage of backward chaining is that it does not make the data grow. When the set of rewritten queries is finite, this set can be seen as a single query, which is the union of the queries in the set. An approach initiated with DL-Lite consists of decomposing backward chaining into two steps: (1) rewrite the initial CQ as a union of CQs (2) use a database management system to answer this rewritten query. This approach aims to benefit from the optimizations developed for classical database queries. It is at the core of several systems, such as QuOnto [Calvanese et al., 2007], Requiem [Pérez-Urbina et al., 2009], Nyaya [Gottlob et al., 2011], Rapid [Chortaras et al., 2011], Iqaros [Venetis et al., 2012] and Quest [Rodriguez-Muro and Calvanese, 2012]. While the above work focuses on specific rule sublanguages, in this paper we consider gen-
eral existential rules, i.e., our algorithm accepts as input any set of existential rules, but of course is guaranteed to stop only for a subset of them (called “finite unification sets” in [Baget et al., 2010]), which includes expressive classes of rules.

The originality of our work lies in the rewriting step based on a notion stemming from a graph view of a set of atoms, that of a piece. Briefly, a piece is a subset of atoms from the query that must be rewritten together. Classically, in logic programming, rules and queries are processed atom by atom: at each step, an atom of a query \( Q \) is unified with the head of a rule \( R \) (a single atom) and a new query is generated by replacing \( a \) in \( Q \) by the body of \( R \) (precisely: let \( u \) be the unifier, the new query is \( u(\text{body}(R)) \cup u(Q \setminus \{ a \}) \)). Here, existential variables in rule heads have to be taken into account, which prevents the use of atomic unification. Instead, subsets of atoms (the pieces) have to be considered at once. We present below a very simple example.

Example 1 Let the rule \( R = \forall x \ (q(x) \rightarrow \exists y \ p(x,y)) \), and the Boolean CQ \( Q = \exists u \exists v \forall w \ (p(u,v) \land p(w,v) \land r(u,w)) \). Assume we want to unify the atom \( p(u,v) \) from \( Q \) with \( p(x,y) \) by a substitution \( \{ (u,x), (v,y) \} \). Since \( v \) is unified with the existential variable \( y \), all other atoms containing \( v \) must also be considered: indeed, simply rewriting \( Q \) into \( q(x) \land p(u,y) \land r(x,w) \) would be incorrect: intuitively, the fact that the atoms \( p(u,v) \) and \( p(w,v) \) in \( Q \) share a variable would be lost in atoms \( q(x) \) and \( p(w,y) \). Thus, \( p(u,v) \) and \( p(u,w) \) have to be both unified with the head of \( R \) by means of the following substitution: \( \{ (u,x), (v,y), (w,x) \} \). \{\( p(u,v), p(w,v) \}\) is called a piece. The corresponding rewriting of \( Q \) is \( q(x) \land r(x,x) \).

An alternative method would be to consider the Skolem form of rules, i.e., to replace existential variables in the head by Skolem functions of variables occurring in the body, however we think it is simpler and more intuitive to keep the original rule language. This framework established, we then posed ourselves the following questions:

1. Can we ensure that we produce a minimal set of rewritten conjunctive queries, in the sense that no sound and complete algorithm can produce a smaller set?

2. How to optimize the rewriting step? The problem of deciding whether there is a piece-unifier between a query and a rule head is NP-complete and the number of piece-unifiers can be exponential in the size of the query.

With respect to the first question, we first point out that any sound and complete set of CQs remains sound and complete when it is restricted to its most general elements (w.r.t. the generalization relation induced by homomorphism). We then show that all sound and complete sets of CQs restricted to their most general CQs have the same cardinality, which is minimal w.r.t. the completeness property.

With respect to the second question, we consider rules with an atomic head. This is not a restriction in terms of expressivity, since any rule can be decomposed into an equivalent set of atomic-head rules by simply introducing a new predicate for each rule [Cali et al., 2008; Baget et al., 2009]. Besides, many rules found in the literature have an atomic head. We exploit the fact that each atom in a CQ \( Q \) belongs to at most one piece with respect to a rule \( R \) (which is false for existential rules with non-atomic head) to efficiently compute a rewriting step, i.e., generate all CQs obtainable from \( R \) and \( Q \). An algorithm producing a sound and complete minimal set of rewritten CQs, and benefiting from the above optimizations, has been implemented.

The paper is organized as follows. Section 2 introduces our framework. Sections 3 and 4 are respectively devoted to the first and to the second question. Finally, Section 5 reports first experiments and outlines further work. See [König et al., 2012a] for the associated technical report with all proofs.

2 Framework

An atom is of the form \( p(t_1, \ldots, t_k) \) where \( p \) is a predicate with arity \( k \), and the \( t_i \) are terms, i.e., variables or constants. Given an atom or a set of atoms \( A \), \( \text{vars}(A) \), \( \text{consts}(A) \) and \( \text{terms}(A) \) denote its set of variables, of constants and of terms, respectively. In the following examples, all the terms are variables (denoted by \( x, y, z \), etc.). \( \models \) denotes the classical logical consequence. A fact is an existentially closed conjunction of atoms. A conjunctive query (CQ) is an existentially quantified conjunction of atoms. When it is a closed formula, it is called a Boolean CQ (BCQ). Hence facts and BCQs have the same logical form. In the following, we will see them as sets of atoms. Given sets of atoms \( A \) and \( B \), a homomorphism \( h \) from \( A \) to \( B \) is a substitution of \( \text{vars}(A) \) by terms \( (B) \) s.t. \( h(A) \subseteq B \). We say that \( A \) maps to \( B \) by \( h \). If there is a homomorphism from \( A \) to \( B \), we say that \( A \) is more general than \( B \) (or \( B \) is more specific than \( A \)), which is denoted \( A \geq B \) (or \( B \leq A \)). Given a fact \( F \) and a BCQ \( Q \), the answer to \( F \) in \( Q \) is positive if \( F \models Q \). It is well-known that \( F \models Q \) iff there is a homomorphism from \( Q \) to \( F \).

Definition 1 (Existential rule) An existential rule (or simply rule) is a formula \( R = \forall x y \ (B[x,y] \rightarrow \exists z H[y,z]) \) where \( B = \text{body}(R) \) and \( H = \text{head}(R) \) are conjunctions of atoms, resp. called the body and the head of \( R \). The frontier of \( R \), noted \( \text{fr}(R) \), is the set \( \text{vars}(B) \cap \text{vars}(H) = \emptyset \).

A knowledge base (KB) \( K = (F,R) \) is composed of a finite set of facts (seen as a single fact) \( F \) and a finite set of existential rules \( R \). The BCQ entailment problem takes as input a KB \( K = (F,R) \) and a BCQ \( Q \), and asks if \( F \models Q \). Other notations: Throughout the paper we note respectively \( R \) and \( Q \) the considered rule and query. We always assume that \( R \) and \( Q \) have no variables in common. Given \( Q' \subseteq Q \), we note \( Q' \) the set \( Q \setminus Q' \). The variables in \( \text{vars}(Q') \cap \text{vars}(Q) \) are called separating variables and noted \( \text{sep}(Q',Q) \). In the examples, we will omit quantifiers in facts and rules since there is no ambiguity.

As explained in the introduction, the rewriting step relies on a specific unification operation based on “pieces”.

A piece-unifier is a pair \( (Q',u) \), where \( Q' \subseteq Q \) and \( u \) is a substitution that “unifies” \( Q' \) with some \( H' \subseteq \text{head}(R) \).

1Pieces come from earlier work on conceptual graph rules [Salvat and Mugnier, 1996] recast in the framework of existential rules in [Baget et al., 2009; 2011a].

2We generalize the classical notion of a fact in order to take existential variables into account.
in the sense that \( u(Q') = u(H') \). The substitution \( u \) can be decomposed as follows: (1) it specializes \( fr(R) \), thus head \( R \geq u(\text{head}(R)) \), with existential variables left unchanged; (2) it maps \( Q' \) to \( u(\text{head}(R)) \), while satisfying the following constraint: the separating variables in \( Q' \) are not mapped to existential variables.

**Definition 2 (Piece-unifier)** A piece-unifier of \( Q \) with \( R \) is a pair \( \mu = (Q', u) \) with \( Q' \subseteq Q \), \( Q' \neq \emptyset \), \( u \) a substitution of \( fr(R) \cup \text{vars}(Q') \) by terms head \( \text{(head}(R)) \cup \text{consts}(Q' \cup \text{head}(R)) \) s.t.:

1. for all \( x \in fr(R) \), \( u(x) \in fr(R) \cup \text{consts}(Q' \cup \text{head}(R)) \) (for technical convenience, we allow \( u(x) = x \));
2. for all \( x \in \text{sep}(Q') \), \( u(x) \in fr(R) \cup \text{consts}(Q' \cup \text{head}(R)) \);
3. \( u(Q') \subseteq u(\text{head}(R)) \).

**Example 2** Let \( R = q(x) \rightarrow p(x, y) \) and \( Q = p(u, v) \land p(z, v) \land p(w, t) \land r(u, w) \). There are three “most general”, cf. Sect. 4) piece-unifiers of \( Q \) w.r.t. \( R \):

- \( \mu_1 = (Q_1, u_1) \) with \( Q_1 = \{ p(u, v), p(z, v) \} \) and \( u_1 = \{(u, x), (v, y), (z, x)\} \): we omit identity pairs in all examples, \( f.i. \), \( u_1 \) contains \( (x, x) \)
- \( \mu_2 = (Q_2, u_2) \) with \( Q_2 = \{ p(w, t) \} \) and \( u_2 = \{(w, x), (t, y)\} \)
- \( \mu_3 = (Q_3, u_3) \) with \( Q_3 = \{ p(v, p), p(z, v), p(w, t) \} \) and \( u_3 = \{(u, x), (v, y), (z, x), (w, x), (t, y)\} \)

We are now able to formally define pieces. Generally speaking, a set of atoms can be partitioned into subsets (called pieces) w.r.t. a set \( T \) of variables acting as “cutpoints”: two atoms are in the same piece if they are connected by a path of variables that are not in \( T \). Here, \( T \) is the set of variables from \( Q' \) that are not mapped to existential variables by \( u \).

**Definition 3 (Piece)** [Baget et al., 2011a] Let \( A \) be a set of atoms and \( T \subseteq \text{vars}(A) \). A piece of \( A \) according to \( T \) is a minimal non-empty subset \( P \) of \( A \) s.t. for all \( a, a' \in A \), if \( a \in P \) and \( \text{vars}(a) \cap \text{vars}(a') \not\subseteq T \), then \( a' \notin P \).

**Definition 4 (Cutpoint, Piece of Q)** Given a piece-unifier \( \mu = (Q', u) \) of \( Q \) with \( R \), a variable \( x \in Q' \) is a cutpoint if \( u(x) \in fr(R) \cup \text{consts}(Q' \cup \text{head}(R)) \). The set of cutpoints associated with \( \mu \) is denoted by \( T_Q(\mu) \). We call piece of \( Q \) (for \( \mu \)) a piece of \( Q \) according to \( T_Q(\mu) \).

**Example 2** (contd) \( Q_1 \) and \( Q_2 \) are pieces for \( \mu_1 \) and \( \mu_2 \) respectively. Both are pieces for \( \mu_3 \).

In fact, for any piece-unifier \( \mu = (Q', u) \), \( Q' \) is a set of pieces of \( Q \), which justifies the name “piece-unifier”. To sum up, a piece of \( Q \) is a minimal subset of atoms that must be considered together once cutpoints in \( Q \) have been defined (indeed, an atom of \( Q \) may belong to different pieces according to different piece-unifiers). Finally, note that in rules without existential variables, such as in plain Datalog, each piece is restricted to a single atom.

**Definition 5 (Immediate Rewriting)** Given a piece-unifier \( \mu = (Q', u) \) of \( Q \) with \( R \), the immediate rewriting of \( Q \) according to \( \mu \), denoted \( \beta(Q, R, \mu) \), is \( u(\text{body}(R)) \cup u(Q') \).

**Definition 6 (R-rewriting of Q)** An \( R \)-rewriting of \( Q \) is a CQ \( Q_{k+1} \) obtained by a finite sequence \( Q_0 = Q, Q_1, \ldots, Q_k \) s.t. for all \( 0 \leq i < k \), there is \( R_i \in R \) and a piece-unifier \( \mu_i \) of \( Q_i \) with \( R_i \) s.t. \( Q_{i+1} = \beta(Q_i, R_i, \mu_i) \).

To evaluate the correctness of different rewriting mechanisms, we introduce the notions of soundness and completeness of a set of CQs with respect to \( Q \) and \( R \) (such a set is called a rewriting set hereafter):

**Definition 7 (Sound and Complete (rewriting) set of CQs)** Let \( R \) be a set of existential rules and \( Q \) be a (B)CQ. Let \( Q \) be a set of CQs. \( Q \) is said to be sound w.r.t. \( R \) and \( Q \) if for all facts \( F \), for all \( Q_i \in Q \), if \( Q_i \) maps to \( F \) then \( R, F \models Q \). Reciprocally, \( Q \) is said to be complete w.r.t. \( R \) and \( Q \) if for all \( F, R, F \models Q \) then there is \( Q_i \in Q \) s.t. \( Q_i \geq F \).

**3 Minimal Rewriting Sets**

We first point out that only the most general elements of a rewriting set need to be considered. Indeed, let \( Q_1 \) and \( Q_2 \) be two elements of a rewriting set s.t. \( Q_2 \subseteq Q_1 \) and let \( F \) be any fact: if \( Q_1 \) maps to \( F \), then \( Q_2 \) is useless; if \( Q_2 \) does not map to \( F \), neither does \( Q_1 \); thus removing \( Q_2 \) will not undermine completeness (nor soundness). The output of a rewriting algorithm should thus be a minimal set of incomparable queries that “covers” all rewritings of the initial query:

**Definition 8 (Cover)** Let \( Q \) be a set of BCQs. A cover of \( Q \) is a set of BCQs \( Q^c \subseteq Q \) s.t.:

1. for any \( Q \in Q \), there is \( Q' \in Q^c \) s.t. \( Q \leq Q' \)
2. elements of \( Q^c \) are pairwise incomparable w.r.t. \( \leq \).

It can be easily checked that all covers of \( Q \) have the same cardinality. Note that the set of rewritings of \( Q \) can have a finite cover even when it is infinite (Example 3).

**Example 3** Let \( Q = t(u), R = t(x) \land p(x, y) \rightarrow t(y) \). The set of \( R \)-rewritings of \( Q \) with \( \{R\} \) is infinite. The first generated queries are the following (note that rule variables are renamed when needed):

- \( Q_0 = t(u) \)
- \( Q_1 = t(x) \land p(x, y) \land \lnot Q_0 \) and \( R \) with \( \{u, y\} \)
- \( Q_2 = t(x_0) \land p(x_0, y_0) \land p(y_0, y) \land \lnot Q_1 \) and \( R \)
- \( Q_3 = t(x_1) \land p(x_1, y_1) \land p(y_1, y_0) \land p(y_0, y) \) and so on . . .

However, the set of the most general \( R \)-rewritings is \( \{Q_0\} \) since any other obtainable query is more specific than \( Q_0 \).

A set of rules \( R \) for which it is ensured that the set of \( R \)-rewritings of any query has a finite cover is called a finite unification set (fus). The fus property is not recognizable [Baget et al., 2011a], but several recognizable fus classes have been exhibited in the literature [Baget et al., 2009; Call et al., 2009; 2010]. Following Algorithm 1 is a breadth-first algorithm that, given a fus \( R \) and a query \( Q \), generates a cover of the set of \( R \)-rewritings of \( Q \). “Exploring” a query consists of computing the set of immediate rewritings of this query with all rules. Initially, \( Q \) is the only query to explore; at each step (a while loop iteration), all queries generated at the preceding step and kept in the current cover are explored. The following lemma justifies the fact that only the most general rewritings are kept at each step of the algorithm.
Theorem 1 Let \( u \) be a sound and complete rewriting algorithm can be "optimized" so that if \( Q \) is an \( \{ R \} \)-rewriting of \( Q \) obtained by considering exclusively most general single-piece unifiers s.t. \( Q^* \geq \beta(Q, R, \mu) \).

Property 2 For any piece-unifier \( \mu \) of \( Q \) with \( R \), there is \( Q^* \) an \( \{ R \} \)-rewriting of \( Q \) obtained by considering exclusively most general single-piece unifiers s.t. \( Q^* \geq \beta(Q, R, \mu) \).

From Lemma 1 and Property 2, we obtain:

Theorem 2 Given a set of rules \( \mathcal{R} \), the set of \( \{ R \} \)-rewritings of \( Q \) obtained by considering exclusively most general single-piece unifiers is sound and complete.

However, single-piece unifiers cannot be used as such in Algorithm 1. The next example shows that, despite the completeness result of Theorem 2, the restriction to single-piece unifiers is not compatible with selecting most general rewritings at each step, as done in Algorithm 1.

Example 5 Let \( Q = \{ p(y, z) \land p(z, y) \} \) and \( R = \{ r(x, x) \} \). There are two single-piece unifiers of \( Q \) with \( R \), \( \mu_1 = \{ p(y, z), u \} \) and \( \mu_2 = \{ p(z, y), u \} \), which yield the same rewriting \( Q_1 = r(x, x) \land p(x, x) \). There is also a two-piece unifier \( \mu = \{ Q, u \} \), which yields \( Q' = r(x, x) \). A query equivalent to \( Q' \) can be obtained from \( Q_1 \) by a further single-piece unification.

Now, assume that we restrict unifiers to single-piece unifiers and keep most general rewritings at each step. Since \( Q \geq Q_1 \), \( Q_1 \) is not kept, so \( Q' \) will never be generated, whereas it is incomparable with \( Q \).

To keep the correctness of Algorithm 1, we have to combine single-piece unifiers when they are compatible: two piece-unifiers of \( Q \) with \( R \), \( \mu_1 = (Q_1', u_1) \) and \( \mu_2 = (Q_2', u_2) \), are said to be compatible if (1) \( Q_1' \cap Q_2' = \emptyset \) and (2) for all \( x \in \text{vars}(Q_1') \cap \text{vars}(Q_2') \), when \( u_1(x) \) and \( u_2(x) \) are both constants, it holds that \( u_1(x) = u_2(x) \). To sum up, we keep the schema of Algorithm 1 but, instead of computing all the piece-unifiers at a given step, we compute the single-piece unifiers, then apply together the compatible ones.

5 Perspectives

First experiments were led with the same benchmark as [Gottlob et al., 2011], then extended using the query generator from [Imprrialou et al., 2012]. We compared our algorithm to Nyaya/NY* rewriting engine [Gottlob et al., 2011]. Both running times were comparable, however we found that NY* did not output minimal rewriting sets (although its output was already shown smaller than the output of other existing systems). Since the benchmark we used consists of very simple ontologies we need to consider larger and more complex rule bases. We also have to compare to other recent rewriting systems, though these systems deal with more restricted classes of rules. Finally, our rewriting mechanism is yet far from being optimized. For instance, the number of explored queries is still very large w.r.t. the size of the final cover. The question of whether it is worthwhile, when rules do not have atomic heads, to deal directly with them, still needs to be addressed.
References


