Benchmarking Ontology-based Query Rewriting Systems

Martha Imprialou\textsuperscript{1}, Giorgos Stoilos\textsuperscript{2} and Bernardo Cuenca Grau\textsuperscript{1}

\textsuperscript{1}Department of Computer Science, University of Oxford, Oxford, UK
martha.imprional@balliol.ox.ac.uk bernardo.cuenca.grau@cs.ox.ac.uk

\textsuperscript{2}School of Electrical and Computer Engineering National Technical University of Athens, Greece
gstoil@image.ece.ntua.gr

Abstract

Query rewriting is a prominent reasoning technique in ontology-based data access applications. A wide variety of query rewriting algorithms have been proposed in recent years and implemented in highly optimised reasoning systems. Query rewriting systems are complex software programs; even if based on provably correct algorithms, sophisticated optimisations make the systems more complex and slower to develop and test. In this paper, we present techniques that allow us to determine whether a system is unsound and/or incomplete for a given test query and ontology. Our evaluation shows that most publicly available query rewriting systems are unsound and/or incomplete, even on commonly used benchmark ontologies; more importantly, our techniques revealed the precise causes of their correctness issues and the systems were then corrected based on our feedback. Finally, since our evaluation is based on a larger set of test queries than existing benchmarks, which are based on hand-crafted queries, it also provides a better understanding of the scalability behaviour of each system.

Introduction

An important application of ontologies is data access, where an ontology provides the knowledge describing the meaning of application’s data as well as the vocabulary used to formulate user queries. In this setting, query languages are often based on conjunctive queries (CQs), and the answers to queries reflect both the application’s data and the knowledge captured by the ontology (Calvanese et al. 2007; Poggi et al. 2008; Glimm et al. 2007; Lutz, Toman, and Wolter 2009; Ortiz, Calvanese, and Eiter 2006).

The need for efficient query answering in ontology-based data access applications has motivated the development of lightweight ontology languages, such as DL-Lite (Calvanese et al. 2007) (which underpins the OWL 2 QL profile (Motik et al. 2009)), DLP (which underpins the OWL 2 RL profile), as well as (fragments of) datalog \textsuperscript{\pm} (Cali et al. 2010). These languages are tailored such that query answering becomes tractable in data complexity (and even in logarithmic space in the case of the logics underlying OWL 2 QL).

In this setting, query rewriting is a common reasoning technique of choice. Intuitively, a rewriting of a query \( Q \) w.r.t. an ontology \( T \) is another query (typically a union of CQs or a datalog query) that captures the information in \( T \) relevant for answering \( Q \) w.r.t. \( T \) and arbitrary data. Due to the growing interest in ontology-based data access, many rewriting algorithms have been proposed in recent years and implemented in (both commercial and non-commercial) reasoning systems. Examples of such systems include QuOnto (Acciarri et al. 2005), Requiem (Pérez-Urbina, Horrocks, and Motik 2009), Presto (Rosati and Almatelli 2010), Nyaya (Gottlob, Orsi, and Pieris 2011), Rapid (Chortaras, Trivela, and Stamou 2011), and IQAROS.\textsuperscript{1}

Despite being based on provably correct algorithms, sophisticated optimisations needed in practice make these query rewriting systems rather complex and error-prone software programs. Consequently, both system and application developers would greatly benefit from practical and reliable benchmarking approaches to query rewriting.

To the best of our knowledge, little work has been done so far towards the systematic benchmarking of ontology-based query rewriting systems, and developers have relied mostly on “hand-crafted” ontologies and queries for evaluating their systems and comparing them to others. An example of a popular benchmark for query rewriting, which was first used in (Pérez-Urbina, Horrocks, and Motik 2009), consists of nine ontologies and five hand-crafted queries for each of them; systems are evaluated regarding the size of the computed rewritings and the rewriting computation time.

Existing benchmarks, however, have several important limitations. First, they are ad hoc: test queries are manually generated for each specific ontology using the developers’ personal knowledge about the ontologies in the benchmark. Second, given a test query \( Q \) and ontology \( T \), it is currently not possible to automatically check whether the relevant systems are sound (i.e., they compute only actual query answers to \( Q \) w.r.t. \( T \) and arbitrary data) and complete (i.e., they compute all answers to \( Q \) w.r.t. \( T \) and arbitrary data).

In this paper, we present a novel approach for addressing...
these important limitations. First, we present an algorithm that synthetically generates test queries for an ontology \( T \). Our algorithm is generic, in the sense that it is applicable to a wide range of ontologies and, in particular, to (Horn) ontologies that can be captured by existential rules (Baget et al. 2011; Cali et al. 2010). Furthermore, our algorithm is fully automatic and hence does not require experts’ involvement. Most importantly, each generated query is relevant to \( T \), in the sense that it can be used to verify whether a system correctly performs a certain set of “inferences”, each of which can be traced back to axioms in \( T \). Second, we present techniques that allow us to determine whether a given system is unsound and/or incomplete for given \( Q \) and \( T \); these techniques are easy to implement and in many cases help detect errors with marginal manual effort.

Our evaluation shows that most publicly available query rewriting systems are unsound and/or incomplete, even on commonly used benchmark ontologies; most importantly, our techniques revealed the precise causes of their unsoundness and incompleteness issues, and as a result some of the evaluated systems were actually corrected by their developers based on our feedback. Finally, since our collection of tests queries is more “exhaustive” than those in existing corpora, our evaluation provides further insights into the scalability behaviour of each system.

Preliminaries

We use standard notions of first-order constants, (free and bound) variables, function symbols, terms, substitutions, predicates, atoms, (ground) formulae, and sentences. A fact is a ground atom, and an instance is a finite set of facts. For \( \phi \) a formula, with \( \phi(\vec{x}) \) we denote that \( \vec{x} \) are the free variables of \( \phi \), while for \( \sigma \) a substitution, \( \phi \sigma \) is the result of applying \( \sigma \) to \( \phi \). Satisfiability and entailment are defined as usual.

Ontologies

We use description logics in the wider framework of first order logic, and identify a DL TBox \( T \) with a finite set of first order sentences. Specialised DL syntax will sometimes be used in examples, and we assume that the reader is familiar with the basics of such syntax (Baader et al. 2002).

Existential Rules

An existential rule (Baget et al. 2011; Cali et al. 2010) is a sentence of the form

\[
\forall \vec{x}. \forall \vec{z}. [\phi(\vec{x}, \vec{z}) \rightarrow \exists \vec{y}. \psi(\vec{x}, \vec{y})]
\]

where \( \phi(\vec{x}, \vec{z}) \) and \( \psi(\vec{x}, \vec{y}) \) are conjunctions of function-free atoms and \( \vec{x}, \vec{y} \) and \( \vec{z} \) are pair-wise disjoint. Formula \( \phi \) is the body, formula \( \psi \) is the head, and universal quantifiers are often omitted. Note that, by definition, existential rules are safe—that is, all variables in \( \vec{x} \) occur both in the body and the head. If \( \vec{y} \) is empty, the rule is datalog. Finally, the instantiation of a datalog rule \( r \) w.r.t. a substitution \( \sigma \) mapping all variables in \( r \) to constants is the instance \( I_\sigma \), consisting all facts \( B \sigma \) with \( B \) a body atom in \( r \). If \( \sigma \) is an injective mapping, then \( I_\sigma \) is an injective instantiation of \( r \).

Many popular DLs (e.g., those underlying the OWL 2 profiles (Motik et al. 2009)) can be captured by existential rules.

Queries

A query \( Q \) is a finite set of sentences containing a distinct query predicate \( Q \). A tuple of constants \( \vec{a} \) is an answer to \( Q \) w.r.t. TBox \( T \) and instance \( I \) if the arity of \( \vec{a} \) agrees with the arity of \( Q \) and \( T \cup I \cup Q \vdash Q(\vec{a}) \). We denote with \( \text{cert}(Q, T \cup I) \) the answers to \( Q \) w.r.t. \( T \cup I \). A query \( Q \) is a union of conjunctive queries (UCQ) if it is a set of datalog rules containing \( Q \) in the head but not in the body. In this case, we sometimes abuse notation and use \( Q \) to denote the head atom, rather than the head predicate. A UCQ is a conjunctive query (CQ) if it has exactly one rule.

Query Rewriting

Intuitively, a rewriting of \( Q \) w.r.t. \( T \) is another query that captures all the information from \( T \) relevant for answering \( Q \) over an arbitrary instance \( I \) (Calvanese et al. 2007; Pérez-Urbina, Motik, and Horrocks 2010; Gottlob, Orsi, and Piers 2011). UCQs and datalog are common target languages for query rewriting; furthermore, virtually all TBoxes to which existing rewriting techniques are applicable can be captured by existential rules.

Definition 1. A datalog rewriting of a CQ \( Q \) w.r.t. TBox \( T \) is a tuple \( \langle \mathcal{R}_D, \mathcal{R}_Q \rangle \) with \( \mathcal{R}_D \) a set of datalog rules not mentioning \( Q \) and \( \mathcal{R}_Q \) a UCQ with query predicate \( Q \) whose body atoms mention only predicates from \( T \), and where for each \( I \) using only predicates from \( T \) we have

\[
\text{cert}(Q, T \cup I) = \text{cert}(\mathcal{R}_Q, \mathcal{R}_D \cup I).
\]

The rewriting \( \langle \mathcal{R}_D, \mathcal{R}_Q \rangle \) is a UCQ rewriting if \( \mathcal{R}_D = \emptyset \).

The Chase

Given a set of existential rules \( \mathcal{R} \) and an instance \( I \), CQ answering over \( \mathcal{R} \cup I \) can be characterised using the chase: a technique that computes a (possibly infinite) set \( \text{chase}(\mathcal{R} \cup I) \) of facts implied by \( \mathcal{R} \) and \( I \) in a forward-chaining manner; the result is a universal model over which the CQ can then be evaluated. Although many chase procedures have been proposed in the literature (Fagin et al. 2005; Deutsch, Nash, and Remmel 2008; Marnette 2009), all variants of the chase satisfy the following basic properties:

- \( \text{cert}(Q, \mathcal{R} \cup I) = \text{cert}(Q, \text{chase}(\mathcal{R} \cup I)) \) for all \( Q, \mathcal{R}, I \).
- For each ground fact \( \alpha \) mentioning only constants and functions from \( \mathcal{R} \cup I \), \( \mathcal{R} \cup I \vdash \alpha \) if and only if \( \alpha \in \text{chase}(\mathcal{R} \cup I) \).

Query Rewriting Systems

Many optimised query rewriting systems for various ontology languages have been developed in recent years (Calvanese et al. 2007; Rosati and Almatselli 2010; Chortaras, Trivela, and Stamou 2011; Gottlob, Orsi, and Piers 2011). To abstract from the implementation details, and focus only on the properties of the system that we want to test for, we introduce an abstract notion of a query rewriting system.

Definition 2. A query rewriting system \( \text{rew} \) for a DL \( \mathcal{L} \) is a computable function that, for each TBox \( T \in \mathcal{L} \) and each CQ \( Q \) with head predicate \( Q \) computes in a finite number of steps a tuple \( \text{rew}(Q, T) = \langle \mathcal{R}_D, \mathcal{R}_Q \rangle \) where \( \mathcal{R}_D \) is a set of datalog rules not containing \( Q \) and \( \mathcal{R}_Q \) is a UCQ with head predicate \( Q \) whose body atoms mention only predicates from \( T \). The system \( \text{rew} \) is a (\( Q, T \))-sound if \( \text{cert}(\mathcal{R}_Q, \mathcal{R}_D \cup I) \subseteq \text{cert}(Q, T \cup I) \) for every instance \( I \) containing only predicates from \( T \). It is a (\( Q, T \))-complete if \( \text{cert}(Q, T \cup I) \subseteq \text{cert}(\mathcal{R}_Q, \mathcal{R}_D \cup I) \) for every instance \( I \) containing only predicates from \( T \).
Most query rewriting systems are based on algorithms whose behaviour on input $T$ and $Q$ can be characterised by the application of the following two steps:

1. **Pre-processing**, where the input TBox $T$ is transformed into a set $R_T$ of existential rules.

2. **Query rewriting**, where a calculus (often a variant of resolution) is used to derive $R_D$ and $R_Q$ from $R_T$ and $Q$.

Consequently, two main sources of implementation errors for rewriting systems can be identified, namely those caused by either an incorrect pre-processing, or an incorrect implementation of the calculus.

**Example 3.** Let $T$ consist of the following DL axiom:

$$\text{Student} \equiv \text{Person} \sqcap \exists\text{enrolled}.\text{Course} \quad (2)$$

The TBox $T$ can be translated into the following set $R$ of equivalent existential rules:

1. $\text{Student}(x) \rightarrow \text{Person}(x)$ \quad (3)
2. $\text{Person}(x) \land \text{enrolled}(x,y) \land \text{Course}(y) \rightarrow \text{Student}(x)$ \quad (4)
3. $\text{Student}(x) \rightarrow \exists y. (\text{enrolled}(x,y) \land \text{Course}(y))$ \quad (5)

Some rewriting systems that would accept $T$ as a valid input do not handle qualified existential restrictions (i.e., concepts such as $\exists\text{enrolled}.\text{Course}$). If pre-processing $T$ naively, one such system $\text{rew}_1$ might “ignore” the conjunction involving sub-concept $\exists\text{enrolled}.\text{Course}$ and translate $T$ into $R_T^0$ consisting only of rules (3) and rule $\text{Person}(x) \rightarrow \text{Student}(x)$. Such system would be unsound for $T$: given query $Q_1 = \text{Student}(x) \rightarrow Q(x)$, it would compute the UCQ rewriting $\langle\emptyset,\{Q_1, Q'_1\}\rangle$, with $Q'_1 = \text{Person}(x) \rightarrow Q(x)$, which for instance $I_1 = \{\text{Person}(c)\}$ leads to the incorrect answer $c$.

Consider now a system $\text{rew}_2$ that can handle qualified existential restrictions, but which, due to an implementation error, translates $T$ into $R_T^0$ consisting of rules (3), (4), and rule $\text{Person}(x) \rightarrow \exists y. (\text{enrolled}(x,y))$, thus leaving out conjunct $\text{Course}(y)$ in rule (5). Such system is incomplete for $T$, as witnessed by $Q_2 = \text{enrolled}(x,y) \land \text{Course}(y) \rightarrow Q(x)$ and $I_2 = \{\text{Student}(d)\}$.

As described later on in our evaluation section, the aforementioned normalisation issues are related to actual errors we detected in the Rapid system (Chortaras, Trivela, and Stamou 2011) using our approach, and which were not revealed by existing benchmarks.

Finally, as already mentioned, state of the art implementations of rewriting calculi are highly optimised. As described later on, our approach also allowed us to detect several correctness issues in implemented optimisations; these include errors in ordering criteria for query atoms in IQAROS, and issues with subsumption optimisations in Requiem (Pérez-Urbina, Horrocks, and Motik 2009), among others. \hfill \Diamond

**Test Query Generation**

Virtually all TBoxes to which state of the art rewriting systems are applicable can be captured by existential rules containing only unary and binary predicates. Thus, we only consider such TBoxes $T$ in this section. Translation of $T$ into existential rules $R$ can be performed using well-known structural transformations, which might introduce “fresh” symbols $\Sigma$ (e.g., see (Motik, Shearer, and Horrocks 2009)).

Algorithm 1 can then be used to compute test queries from $R$. The algorithm directly exploits the chase technique, and works in three main stages:

1. **Chase initialisation**, via instantiation of body atoms in $R$.

2. **Chase expansion**, up to a pre-defined bound.

3. **Query generation**, by “navigating” the expanded chase while replacing ground terms by fresh variables.

**Chase initialisation** For each rule $r \in R$, Algorithm 1 instantiates all body atoms using fresh constants, which are marked as *root constants*, thus constructing an instance $I_0$ that “triggers” the application of each rule in $R$ (lines 2–6). For the rules in Example 3 we obtain the following initial instance, where root $= \{a, (b,c)\}$:

$$I_0 = \{\text{Student}(a), \text{Person}(b), \text{enrolled}(b,c), \text{Course}(c)\}$$

**Chase expansion** A chase procedure is then used to materialise new implied facts from existing ones by forward chaining application of rules in $R$ (lines 8–10). Intuitively, when initialised with $I_0$, the properties of the chase ensure that each rule application represents an inference that is relevant to derive some answer to some query $Q$ w.r.t. some input instance (and for the given, fixed, TBox $T$).

\begin{algorithm}
\begin{algorithmic}
\State \textbf{input:} Existential rules $R$; integer bound $\mu$. \hfill \textbf{Algorithm 1 QueryGeneration}($R$, $\mu$, $\Sigma$)
\State $I_0 := \emptyset$ and root := $\emptyset$
\ForAll{$r \in R$} \hfill 1: \State $I_0 := I_0 \cup I'_0$
\EndFor \hfill 6: \State $I_e := I_e \cup \text{applyChaseRule}(R, I_e)$
\Repeat \hfill 7: \State $I_e := I_e \cup \text{applyChaseRule}(R, I_e)$
\Until terminate($I_e$, bound) \hfill 8: \State \textbf{repeat}
\State $CQ := CQ \cup \{\bigwedge_{P(\tilde{b}) \in \Gamma} P(\sigma(\tilde{b})) \rightarrow Q(\sigma(\tilde{a}))\}$ \hfill 9: \State $I := I_e \setminus I_0$ and CQ := $\emptyset$
\ForAll{$\tilde{a} \in \text{root}$} \hfill 10: \State $I_e := I_e \cup \text{applyChaseRule}(R, I_e)$
\EndFor \hfill 11: \State $I := I_e \setminus I_0$ and CQ := $\emptyset$
\ForAll{$I' \in \text{paths}(\tilde{a}, I, \Sigma)$} \hfill 12: \State \textbf{for all} $I' \in \text{paths}(\tilde{a}, I, \Sigma)$ do
\EndFor \hfill 13: \State $\textbf{for all} I' \in \text{paths}(\tilde{a}, I, \Sigma)$ do
\EndFor \hfill 14: \State $\textbf{for all} I' \in \text{paths}(\tilde{a}, I, \Sigma)$ do
\EndFor \hfill 15: \State Let $\sigma$ map each constant in $I'$ to fresh variable \hfill 16: \State $CQ := CQ \cup \{\bigwedge_{P(\tilde{b}) \in \Gamma} P(\sigma(\tilde{b})) \rightarrow Q(\sigma(\tilde{a}))\}$ \hfill 17: \State \textbf{for all} $I' \in \text{paths}(\tilde{a}, I, \Sigma)$ do
\EndFor \hfill 18: \State \textbf{return} CQ \hfill 19: \EndFor \end{algorithmic}
\end{algorithm}

\footnote{Note that rules (3) and (5) share the same body, so it suffices to instantiate one of them.}
Since the chase might not terminate, the algorithm accepts an integer bound (e.g., maximum number of generated facts), which can be used as a termination condition. Alternatively, one can use any of the available acyclicity conditions for existential rules (e.g., (Fagin et al. 2005; Marnette 2009)), which analyse the information flow between the rules to ensure that no cyclic generation of fresh terms occurs. If \( R \) satisfies any such condition, the chase terminates, and one can dispense with the input bound. In our example, chase expansion would lead to the following instance \( I_c \), where \( a_1 \) and \( b_1 \) are fresh constants generated by the application of rule (5).

\[
I_c = I_0 \cup \{ \text{Person}(a_1), \text{enrolled}(a_1, a_1), \text{Course}(a_1), \text{Student}(b), \text{enrolled}(b, b_1), \text{Course}(b) \}
\]

**Query generation** Instance \( I = I_c \setminus I_0 \) contains all facts that have been derived from \( I_0 \) during chase expansion. The properties of the chase ensure that \( I \) is forest-shaped. More precisely, since \( I \) contains only unary and binary predicates, we can define \( G(I) \) as the graph whose nodes are the constants in \( I \) and which contains an (undirected) edge between \( c \) and \( d \) iff \( c \) and \( d \) occur together in a fact from \( I \); then \( G(I) \) consists of a set of disconnected trees each of which is rooted at an individual from root.

Algorithm 1 generates queries by traversing \( G(I) \) (lines 12–17) along “paths”. More precisely, paths \((\bar{a}, I, \Sigma)\) (see line 13) is the set of all instances \( I' \subseteq I \) not containing predicates from \( \Sigma \) and s.t. \( G(I') \) is a path in \( G(I) \) involving a constant from \( \bar{a} \). Each such \( I' \) is then transformed into a CQ by mapping each constant to a fresh variable s.t. root constants are mapped to answer variables in the head of the query. Clearly, all queries mentioned in Example 3 would be generated by Algorithm 1 for input rules (3)–(5).

Note that, in the worst-case, Algorithm 1 might generate exponentially many queries w.r.t. the size of the input. As shown in our evaluation, however, the number of generated queries for commonly-used benchmark ontologies is relatively modest, and they can be computed in just a few seconds. Finally, note that relevant test queries could also be generated from \( I \) by considering instances \( I' \subseteq I \) whose graph is a sub-tree of \( G(I) \), rather than simply a path; however, on the one hand, this results in a blowup in the number of queries in practice and, on the other hand, our evaluation suggests that path queries suffice for uncovering many errors in systems and for analysing and comparing their behaviour.

**Testing Correctness**

Algorithm 1 provides a flexible way for generating relevant test queries. In this section, we present practical techniques for testing soundness and completeness of a rewriting system for a given test query and a given TBox.

**Testing Soundness**

Assume that \( \text{rew}(Q, T) = (R_D, R_Q) \). Clearly, if we have \( T \cup Q \models R_D \cup R_Q \), the properties of first-order logic entailment ensure that each answer to \( Q \) w.r.t. \( R_D \cup R_Q \) and an instance \( I \) is also an answer to \( Q \) w.r.t. \( T \cup I \), and hence the query rewriting system \( \text{rew} \) is \((Q, T)\)-sound.

Proposition 4 shows that \( T \cup Q \models R_D \cup R_Q \) can be checked using any reasoner that is sound and complete for checking entailment of a fact by \( T \) and an instance. For standard DLs, example such reasoners include HermiT (Motik, Siderovski, and Horrocks 2009) and Pellet (Sirin et al. 2007).

**Proposition 4.** Let \( \text{rew}(Q, T) = (R_D, R_Q) \). Then, we have that \( T \cup Q \models R_D \cup R_Q \) iff the following conditions hold:

1. \( T \cup Q \models R_D \cup R_Q \) for each \( r \in R_D \), where \( H \) is the head of \( r \) and \( I_r^\sigma \) an injective instantiation.
2. \( T \cup Q \models \sigma \) for each \( r \in R_Q \), where \( Q \) is the head of \( r \) and \( I_r^\sigma \) an injective instantiation.

**Proof.** Since \( Q \) does not occur in \( R_D \), we have that \( T \cup Q \models R_D \cup R_Q \) iff \( T \models R_D \) and \( T \cup Q \models R_Q \). Furthermore, it is easy to check that a set \( \mathcal{F} \) of sentences entails a datalog rule \( r \) with head \( H \) iff \( \mathcal{F} \cup I_r^\sigma \models H \sigma \) for \( I_r^\sigma \) an injective instantiation; thus, since \( R_D \) and \( R_Q \) consist of datalog rules, we have \( T \models R_D \) iff Condition 1 holds and \( T \cup Q \models R_Q \) iff Condition 2 holds.

In practice, however, checking soundness of \( \text{rew} \) using a fully-fledged DL reasoner, as suggested by Proposition 4, has the drawback that the process is subject to implementation errors of the DL reasoner. Although this problem can be alleviated by performing the required entailment tests using several DL reasoners to detect possible discrepancies, an alternative is to use a chase procedure, such as the one we used for query generation (see Algorithm 1). Such procedure does not need to be optimised, and consequently its implementation can be rather simple and transparent.

The following proposition shows how to exploit the chase for checking \((Q, T)\)-soundness of \( \text{rew} \).

**Proposition 5.** Let \( \text{rew}(Q, T) = (R_D, R_Q) \) and let \( R \) be a set of existential rules such that \( T \) and \( R \) are logically equivalent. The following conditions imply that \( \text{rew} \) is \((Q, T)\)-sound:

1. \( H \sigma \in \text{chase}(R \cup I_r^\sigma) \) for each \( r \in R_D \), where \( H \) is the head of \( r \) and \( I_r^\sigma \) an injective instantiation.
2. \( \bar{a} \in \text{cert}(Q, \text{chase}(R \cup I_r^\sigma)) \) for each \( r \in R_Q \), where \( Q \) is the head of \( r \) and \( I_r^\sigma \) an injective instantiation mapping the variables in \( Q \) to \( \bar{a} \).

Furthermore, if Condition 2 does not hold, then \( \text{rew} \) is not \((Q, T)\)-sound.

**Proof.** To show the first claim in the proposition, it suffices to prove that Condition 1 (resp. Condition 2) in the proposition implies Condition 1 (resp. Condition 2) in Proposition 4. Let \( r \in R_D \), and assume that \( H \sigma \in \text{chase}(R \cup I_r^\sigma) \) with \( I_r^\sigma \) an injective instantiation; since \( H \sigma \) mentions only individuals from \( I_r^\sigma \) (recall that rules are assumed to be safe), the properties of the chase (see preliminaries) ensure that \( R \cup I_r^\sigma \models H \sigma \); but then, since \( T \models R \), we have \( T \cup I_r^\sigma \models H \sigma \), as required. Finally, let \( r \in R_Q \) and let \( \bar{a} \in \text{cert}(Q, \text{chase}(R \cup I_r^\sigma)) \), where \( I_r^\sigma \) is an injective instantiation mapping the variables in \( Q \) to \( \bar{a} \). The properties of chase ensure that \( \bar{a} \in \text{cert}(Q, R \cup I_r^\sigma) \), which implies \( R \cup Q \cup I_r^\sigma \models Q \sigma \). Since \( T \models R \), we then have \( T \cup Q \cup I_r^\sigma \models Q \sigma \), as required.
To show proposition’s last claim, consider $r \in R_Q$ and an injective instantiation $I^r_\sigma$, where $\vec{x}$ are the variables in $Q$ and $\vec{a} = \sigma(\vec{x})$. Clearly, $\vec{a} \in \text{cert}(R_Q, R_D \cup I_\sigma)$; furthermore, $I^r_\sigma$ contains only predicates from $T$. Next, assume $\vec{a} \not\in \text{cert}(Q, \text{chase}(R \cup I^r_\sigma))$; then, the properties of the chase imply that $\vec{a} \not\in \text{cert}(Q, R \cup I^r_\sigma)$; finally, since $R \models T$, we have $\vec{a} \not\in \text{cert}(Q, T \cup I^r_\sigma)$. Thus, rew is not $(Q, T)$-sound.

Note that Proposition 5 provides only a sufficient condition for $(Q, T)$-soundness of rew; indeed, it may be the case that Condition 1 fails (i.e., $H \sigma \not\in \text{chase}(R \cup I^r_\sigma)$ for some $r \in R_D$), but rew is $(Q, T)$-sound. In contrast, as shown, failure of Condition 2 for some $r \in R_Q$ provides a counter-example for $(Q, T)$-soundness. Thus, if rew $(Q, T)$ is a UCQ (i.e., $R_D = \emptyset$), Proposition 5 provides both a necessary and sufficient condition for $(Q, T)$-soundness.

**Example 6.** Consider TBox $T$, rules $R$, queries $Q_1$, $Q_2$ and system rew_1 from Example 3. For $Q_1$ and $T$, rew_1 computes rew_1$(Q_1, T) = \langle \emptyset, \{Q_1, Q'_1\} \rangle$. Let $I$ = $\{ \text{Person}(a) \}$ be an injective instantiation of $Q'_1$ for $\sigma = \{ x \rightarrow a \}$. Clearly, $T \models R$ and $\vec{a} \not\in \text{cert}(Q_1, \text{chase}(R \cup I^r_\sigma))$. Thus, Proposition 5 implies that rew_1 is indeed unsound for $Q_1$ and $T$.

In practice, to ensure $(Q, T)$-soundness of rew it suffices to check that Conditions 1 and 2 in Proposition 5 hold already for a subset of the corresponding chase, in which case chase termination does not need to be ensured. On the one hand, these conditions involve checking whether a particular fact can be found in the chase (or whether a particular answer can be derived from the chase); on the other hand, chase construction is monotonic (i.e., derived facts are never deleted), so if a fact can be found in (or an answer can be derived from) a finite subset of the chase, then the fact (or answer) is guaranteed to hold in the final chase.

To prove unsoundness, however, one needs to check failure of Condition 2, which requires the full expansion of chase $(R \cup I^r_\sigma)$ for each relevant $r \in R_Q$; thus, chase termination becomes an important issue. Our evaluation will show that these issues can be dealt with, and Proposition 5 can be used effectively to determine soundness and unsoundness in practice for typical benchmark ontologies.

### Testing Completeness

Our strategy for detecting sources of incompleteness is to look for and (semi)automatically explicate "disagreements" between query rewriting systems for given $Q$ and $T$. The following definition provides the required notions to compare systems’ outputs for completeness evaluation purposes.

**Definition 7.** Let rew_i $(Q, T)$ = $(R_D^i, R_Q^i)$ for $i \in \{1, 2\}$. We say that rew_2 extends rew_1 for $Q$ and $T$ if the following condition holds for each instance $I$ mentioning only predicates from $T$:

$$\text{cert}(R_Q^1, R_D^2 \cup I) \subseteq \text{cert}(R_Q^2, R_D^2 \cup I)$$

(6)

If rew_2 extends rew_1 and the inclusion in Condition (6) is proper for some $I$, we say that rew_2 strictly extends rew_1.

Intuitively, if rew_2 strictly extends rew_1, then it is "at least as complete as" rew_1 for the given $Q$ and $T$. Thus, when evaluating several systems in practice, if one of them extends all the others, we have strong evidence to believe that such system is $(Q, T)$-complete. Furthermore, if a system rew_2 is $(Q, T)$-sound and strictly extends rew_1 for $Q$ and $T$, we can conclude that rew_1 is incomplete for $Q$ and $T$.

The following proposition provides sufficient conditions for checking whether rew_2 (strictly) extends rew_1; hence, when used in combination with Proposition 5, it allows us to draw conclusions about systems’ completeness.

**Proposition 8.** Let rew_i $(Q, T)$ = $(R_D^i, R_Q^i)$ for $i \in \{1, 2\}$. If both of the following conditions hold, then rew_2 extends rew_1 for $Q$ and $T$:

1. $H \sigma \in \text{chase}(R_D^2 \cup I^r_\sigma)$ for each $r \in R_D^1$.
2. $Q \sigma \in \text{chase}(R_Q^2 \cup I^r_\sigma)$ for each $r \in R_Q^1$.

Furthermore, if $R_D^2 = \emptyset$, rew_2 strictly extends rew_1 for $Q$ and $T$.

**Proof.** Assume that $H \sigma \in \text{chase}(R_D^2 \cup I^r_\sigma)$ for each $r \in R_D^1$; the properties of chase ensure that $R_D^1 \cup I^r_\sigma \models H \sigma$ for each $r \in R_D^1$; but then, since $I^r_\sigma$ is an injective instantiation of $r$, this implies that $R_D^1 \models \text{chase}(R_D^2 \cup I^r_\sigma)$. Similarly, if Condition 2 holds, then we have $R_D^2 \models \text{chase}(R_Q^2 \cup I^r_\sigma)$. As a result, (6) holds by the properties of entailment, as required.

Finally, assume that $R_D^1 = R_D^2 = \emptyset$ and $r \in R_Q^2$ exists such that $Q \sigma \not\in \text{chase}(R_D^2 \cup I^r_\sigma)$ for $I^r_\sigma$ an injective instantiation. The properties of the chase then ensure that $R_D^1 \cup I^r_\sigma \not\models Q \sigma$; since $R_D^2 \cup I^r_\sigma \models Q \sigma$, and rew_2 extends rew_1, the inclusion in (6) is indeed proper.

Note that $R_D^1$ and $R_Q^2$ consist of datalog rules; thus, the chase computations involved in Proposition 8 are guaranteed to terminate. Furthermore, if we determine that rew_2 strictly extends rew_1, we can track the application of chase rules in the corresponding chase expansion to detect the source of the systems’ disagreement in a (semi)automatic way. As described in the evaluation section, this allowed us to detect the sources of incompleteness in practice.

**Example 9.** Consider Q2, T, and rew_2 from Example 3 and let rew_2 $(Q_2, T) = (\emptyset, R_Q^2)$. Since rew_2 computes rew_2$(Q_2, T) = (\emptyset, R_Q^2)$ in Example 3, we have $Q'_2 \not\in R_Q^2$ for $Q'_2 = \text{Student}(x) \rightarrow Q(x)$. Let rew_3 be a sound system s.t. rew_3 $(Q_2, T) = (\emptyset, R_Q^2 \cup \{Q'_2\})$. rew_3 strictly extends rew_2 since $Q(a) \not\in \text{chase}(R_Q^2 \cup I^r_\sigma)$ for $I^r_\sigma = \{\text{Student}(a)\}$; thus, rew_2 is not $(Q_2, T)$-complete.

### Experiments

We have implemented a test query generator (see Algorithm 1) that is applicable to TBoxes in the DL ECLvII.\(^3\) We then applied our query generator to the benchmark ontologies described in (Pérez-Urbina, Horrocks, and Motik 2009).

\(^3\)Available at http://code.google.com/p/sygenia/
For all test ontologies, except for A, AX, and S, the chase naturally terminated in just a few seconds. Instead of fixing an a priori integer bound for the “cyclic” ontologies (as required by Algorithm 1), our implementation annotates each skolem constant generated by the chase with the rules with existentials in the head that have been applied so far to generate them. If such a rule \( r \) is applicable to facts involving a skolem constant, but \( r \) has been used before in order to generate this particular skolem constant, then we detect a “cycle”; what we bound is the number of such detected cycles. Table 1 summarises our results. We can observe that test queries for all ontologies can be generated rather efficiently and the number of them is also relatively small.

<table>
<thead>
<tr>
<th>System</th>
<th>Ontology</th>
<th>#Queries</th>
<th>Issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requiem</td>
<td>AX</td>
<td>8</td>
<td>incomplete</td>
</tr>
<tr>
<td>CGLLR</td>
<td>U</td>
<td>36</td>
<td>unsound</td>
</tr>
<tr>
<td></td>
<td>UX</td>
<td>18</td>
<td>incomplete</td>
</tr>
<tr>
<td>IQAROS</td>
<td>S</td>
<td>25</td>
<td>incomplete</td>
</tr>
<tr>
<td>Nyaya</td>
<td>S, UX</td>
<td>82</td>
<td>incomplete</td>
</tr>
</tbody>
</table>

Table 2: Statistics about unsoundness and incompleteness

We used these queries to evaluate and compare CGLLR,\(^4\) REQUIEM,\(^5\) Rapid,\(^6\) IQAROS,\(^7\) and Nyaya;\(^8\) we didn’t evaluate Presto and QuOnto, which aren’t publicly available.

Correctness Evaluation

We ran all systems over all ontologies and all our test queries and analysed their output (UCQ) rewritings. Contrary to previous evaluations (Pérez-Urbina, Horrocks, and Motik 2009; Chortaras, Trivela, and Stamou 2011; Rosati and Almatelli 2010), which were based on the same ontologies, but which used “hand-crafted” test queries instead, we discovered important correctness issues in all these systems (see Table 2).

<table>
<thead>
<tr>
<th>System</th>
<th>Ontology</th>
<th>#queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requiem</td>
<td>AX</td>
<td>8</td>
</tr>
<tr>
<td>CGLLR</td>
<td>AX</td>
<td>8</td>
</tr>
<tr>
<td>Rapid</td>
<td>U</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>UX</td>
<td>18</td>
</tr>
<tr>
<td>IQAROS</td>
<td>S</td>
<td>25</td>
</tr>
<tr>
<td>Nyaya</td>
<td>S, UX</td>
<td>82</td>
</tr>
</tbody>
</table>

Table 1: Statistics of query generation

Table 3: Sum of rewriting times (seconds) for all test queries

<table>
<thead>
<tr>
<th>System</th>
<th>CG</th>
<th>Req</th>
<th>Rap</th>
<th>IQAROS</th>
<th>Nyaya</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>65.4</td>
<td>1.9</td>
<td>1.6</td>
<td>0.4</td>
<td>662.2</td>
</tr>
<tr>
<td>AX</td>
<td>5,680.0</td>
<td>10,442.2</td>
<td>586.8</td>
<td>64.2</td>
<td>37.5</td>
</tr>
<tr>
<td>P5</td>
<td>163.0</td>
<td>9.4</td>
<td>0.2</td>
<td>1.1</td>
<td>0.4</td>
</tr>
<tr>
<td>P5X</td>
<td>310.0</td>
<td>78.7</td>
<td>4.8</td>
<td>1.0</td>
<td>0.4</td>
</tr>
<tr>
<td>S</td>
<td>0.7</td>
<td>1.2</td>
<td>0.6</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>U</td>
<td>0.3</td>
<td>0.4</td>
<td>0.2</td>
<td>0.04</td>
<td>0.4</td>
</tr>
<tr>
<td>UX</td>
<td>0.3</td>
<td>0.6</td>
<td>0.3</td>
<td>0.06</td>
<td>0.6</td>
</tr>
<tr>
<td>V</td>
<td>2.3</td>
<td>3.4</td>
<td>0.7</td>
<td>0.8</td>
<td>5.7</td>
</tr>
</tbody>
</table>

Performance Evaluation

Performance was evaluated by using the systems’ latest versions; in the case of CGLLR, Requiem, Rapid, and IQAROS, the latest versions include fixes for all the correctness issues we identified in our experiments.

For each system, we measured the time to compute a UCQ rewriting for each test query and for each ontology. Due to space limitations we cannot present results for each test; hence, for illustration purposes, Table 3 provides for each ontology and each system the total time that the system took to compute a rewriting for all test queries. We can observe that Rapid and IQAROS were significantly faster than the other systems.\(^9\) Moreover, Nyaya could not finish rewriting the test queries for ontologies AX and P5X after 5 hours.

Finally, we have evaluated systems’ redundancy elimination mechanisms by measuring the size of the computed UCQ rewritings (i.e., the number of CQs they contain). Table 4 presents, for each ontology and each system, the sum of the sizes of all (UCQ) rewritings computed for all the test queries.\(^10\) We can observe that CGLLR presents the highest degrees of redundancy in its output, whereas Rapid and IQAROS computed the most succinct rewritings.

\(^4\)http://www.cs.ox.ac.uk/projects/requiem/C.zip
\(^5\)http://www.cs.ox.ac.uk/projects/requiem/home.html
\(^6\)http://www.image.ece.ntua.gr/~achort/rapid.zip
\(^7\)http://code.google.com/p/iqaros/
\(^8\)http://mais.dia.uniroma3.it/Nyaya/Home.html

\(^9\)We do not present results for ontology P1 since it was trivial (under 50 milliseconds) for all systems.

\(^10\)See Table 1 for total number of test queries for each ontology.
<table>
<thead>
<tr>
<th>( \tau )</th>
<th>CG</th>
<th>Req</th>
<th>Rap</th>
<th>IQAR</th>
<th>Nyaya</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>70,350</td>
<td>4,645</td>
<td>4,133</td>
<td>5,445</td>
<td>4,073</td>
</tr>
<tr>
<td>AX</td>
<td>815,921</td>
<td>976,151</td>
<td>369,817</td>
<td>159,252</td>
<td>-</td>
</tr>
<tr>
<td>P5</td>
<td>33,363</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>P5X</td>
<td>74,552</td>
<td>33,292</td>
<td>13,599</td>
<td>8,269</td>
<td>-</td>
</tr>
<tr>
<td>S</td>
<td>5,148</td>
<td>4,493</td>
<td>857</td>
<td>835</td>
<td>2,288</td>
</tr>
<tr>
<td>U</td>
<td>2,856</td>
<td>1,933</td>
<td>489</td>
<td>486</td>
<td>1,640</td>
</tr>
<tr>
<td>UX</td>
<td>3,060</td>
<td>2,949</td>
<td>705</td>
<td>702</td>
<td>2,561</td>
</tr>
<tr>
<td>V</td>
<td>13,439</td>
<td>13,356</td>
<td>3,737</td>
<td>3,737</td>
<td>8,609</td>
</tr>
</tbody>
</table>

Table 4: Sum of rewriting sizes for all test queries.

**Related Work and Conclusions**

To the best of our knowledge all benchmarks for ontology-based systems rely on hand-crafted test queries (Pérez-Urbina, Motik, and Horrocks 2010; Guo, Pan, and Hefflin 2005; Ma et al. 2006). The work that is closest to ours is the formal study of incompleteness of ontology reasoners (Stoilos, Cuenca Grau, and Horrocks 2010; Cuenca Grau and Stoilos 2011), where the authors describe techniques for synthetic query generation; however, these were limited to very lightweight DLs (\( \mathcal{EL} \) and DL-Lite) under very strict acyclicity conditions, and the queries were used to evaluate incomplete materialisation-based reasoners, such as Jena, Sesame, and OWLim.

In this paper, we presented a test query generation algorithm that can be applied to Horn ontologies. We also presented correctness evaluation techniques which have already been proved useful to uncover serious implementation errors in state of the art rewriting systems. These were not revealed by existing benchmarks and hence our techniques have already proved valuable to systems’ developers.

**Acknowledgements**

This work was supported by the EU FP7 project SEALS and by the EPSRC projects ConDOR, ExODA, and LogMap. B. Cuenca Grau is supported by a Royal Society University Research Fellowship. Giorgos Stoilos is supported by a Marie Curie FP7-Reintegration-Grants within European Union’s Seventh Framework Programme (FP7/2007-2013).

**References**


