# Schema-Agnostic Query Rewriting in SPARQL 1.1*

Stefan Bischof[1], Markus Krötzsch[2], Axel Polleres[3], and Sebastian Rudolph[2]

[1] Vienna University of Technology, Austria and Siemens AG Österreich, Austria
[2] Technische Universität Dresden, Germany
[3] Vienna University of Economics and Business, Austria

**Abstract.** SPARQL 1.1 supports the use of ontologies to enrich query results with logical entailments, and OWL 2 provides a dedicated fragment OWL QL for this purpose. Typical implementations use the OWL QL schema to rewrite a conjunctive query into an equivalent set of queries, to be answered against the non-schema part of the data. With the adoption of the recent SPARQL 1.1 standard, however, RDF databases are capable of answering much more expressive queries directly, and we ask how this can be exploited in query rewriting. We find that SPARQL 1.1 is powerful enough to "implement" a full-fledged OWL QL reasoner in a single query. Using additional SPARQL 1.1 features, we develop a new method of schema-agnostic query rewriting, where arbitrary conjunctive queries over OWL QL are rewritten into equivalent SPARQL 1.1 queries in a way that is fully independent of the actual schema. This allows us to query RDF data under OWL QL entailment without extracting or preprocessing OWL axioms.

## 1 Introduction

SPARQL 1.1, the recent revision of the W3C SPARQL standard, introduces significant extensions to the capabilities of the popular RDF query language [11]. Even at the very core of the query language, we can find many notable new features, including *property paths*, *value creation* (BIND), inline data (VALUES), negation, and extended filtering capabilities. In addition, SPARQL 1.1 now supports query answering over OWL ontologies, taking full advantage of ontological information in the data [9].

Query answering in the presence of ontologies is known as *ontology-based data access* (OBDA), and has long been an important topic in applied and foundational research. Even before SPARQL provided support for this feature, several projects have used ontologies to integrate disparate data sources or to provide views over legacy databases, e.g. [5,16,17,7,12]. The W3C OWL 2 Web Ontology Language includes the OWL QL language profile, which was specifically designed for this application [13]. With the arrival of SPARQL 1.1, every aspect of OBDA is thus supported by tailor-made W3C technologies.

In practice, however, SPARQL and OWL QL are rarely integrated. Most works on OBDA address the problem of answering *conjunctive queries* (CQs), which correspond to SELECT-PROJECT-JOIN queries in SQL, and (to some degree) to Basic Graph Patterns in SPARQL. The most common approach for OBDA is *query rewriting*, where

* Extended version of paper published at ISWC 2014.

a given CQ is rewritten into a (set of) CQs that fully incorporate the schema information of the ontology. The answers to the rewritten queries (obtained without considering the ontology) are guaranteed to agree with the answers of the original queries (over the ontology). This approach separates the ontology (used for query rewriting) from the rest of the data (used for query answering), and it is typical that the latter is stored in a relational database. Correspondingly, the rewritten queries are often transformed into SQL for query answering. SPARQL and RDF do not play a role in this.

In this paper, we thus take a fresh look on the problem of OBDA query rewriting with SPARQL 1.1 as our target query language. The additional expressive power of SPARQL 1.1 allows us to introduce a new paradigm of *schema-agnostic query rewriting*, where the ontological schema is not needed for rewriting queries. Rather, the ontology is stored *together with the data* in a single RDF database. This is how many ontologies are managed today, and it corresponds to the W3C view on OWL and RDF, which does not distinguish schema and data components. The fact that today's OBDA approaches separate both parts testifies to their focus on relational databases. Our work, somewhat ironically, widens the scope of OWL QL to RDF-based applications, which have hitherto focused on OWL RL as their ontology language of choice.

Another practical advantage of schema-agnostic query rewriting is that it supports frequent updates of both data and schema. The rewriting system does not need any information on the content of the database under query, while the SPARQL processor that executes the query does not need any support for OWL. This is particularly interesting if a database can only be accessed through a restricted SPARQL query interface that does not support reasoning. For example, we have used our approach to check the consistency of DBpedia under OWL semantics, using only the public Live DBpedia SPARQL endpoint[4] (it is inconsistent: every library is inferred to belong to the mutually disjoint classes "Place" and "Agent").

Our main contributions are as follows:

- We express the standard reasoning tasks for OWL QL, including consistency checking, classification, and instance retrieval, in *single, fixed* SPARQL 1.1 queries that are independent of the ontology. For this, we use SPARQL 1.1 property paths, which support a simple form of recursion that is powerful enough for OWL QL reasoning.
- We show how to rewrite arbitrary SPARQL Basic Graph Patterns (BGPs) into single SPARQL 1.1 queries of polynomial size. This task is simplified by the fact that SPARQL does not support "non-distinguished" variables as used in general CQs.
- We present a schema-agnostic rewriting of general CQs in SPARQL 1.1, again into single queries of polynomial size. This rewriting is more involved, and we use two additional features: inline data (VALUES) and (in)equality checks in filters.
- We show the limits of schema-agnostic rewriting in SPARQL 1.1 by proving that many other OWL features cannot be supported in this way. This includes even the most basic features OWL EL and OWL RL, and mild extensions of OWL QL.

Worst-case reasoning complexity remains the same in all cases, yet our approach is much more practical in the case of standard reasoning and BGP rewriting. For general CQs, the rewritten queries are usually too complex for today's RDF databases to handle.

[4] http://live.dbpedia.org/sparql

Nevertheless, we think that our "SPARQL 1.1 implementation" of OWL QL query answering is a valuable contribution, since it reduces the problem of supporting OWL QL in an RDF database to the task of optimizing a single (type of) query. Since OWL QL subsumes RDFS, one can also apply our insights to implement query answering under RDFS ontologies, which again leads to much simpler queries.

In Section 2, we start by giving a compact introduction to the parts of SPARQL 1.1 that we require. Thereafter, in Section 3, we introduce OWL QL and relate its semantics to a *chase* procedure. In Section 4, we develop queries for implementing basic QL reasoning in SPARQL 1.1, and in Section 5, we extend this into a schema-agnostic query rewriting procedure for conjunctive queries. Finally, we investigate the limits of schema-agnostic query rewriting, and present several negative results in Section 6. We close with a short discussion and outlook in Section 7. Omitted proofs can be found in the accompanying technical report [3].

## 2 Preliminaries: RDF and SPARQL 1.1

We consider RDF documents based on the set **IRI** of IRIs and **BN** of blank node identifiers; we do not consider literals, since they would complicate our exposition without adding technical insights (they can mostly be treated like named individuals in OWL QL). We use Turtle syntax for denoting RDF throughout this paper.

In addition to IRIs and blank nodes, SPARQL 1.1 queries use variables as constituents, which are indicated by a preceding question mark. For compatibility with the entailment regimes, we will consider SPARQL 1.1 under the set semantics, i.e., multiplicities of solutions will be ignored, as indicated by the DISTINCT keyword. Next, we introduce syntax and semantics of the SPARQL 1.1 fragment employed in this paper.

*Path expressions* are defined inductively as follows: (i) Every IRI is a property path. (ii) For $p$ and $q$ property paths, the following expressions are property paths as well: $(\hat{}p)$ for inverse, $(p\,/\,q)$ for sequence, $(p\,|\,q)$ for alternative, $(p^*)$ for Kleene star. As usual, parentheses can be omitted if there is no danger of confusion. *Triple expressions* are of the form $s\ p\ o$ where $s$ and $o$ are IRIs, blank nodes, or variables, whereas $p$ is an IRI, a variable, or a path expression. *Basic graph patterns* are defined as finite sequences of triple expressions separated by a period. *Values blocks* for inline data have the shape VALUES $(?x_1 \ldots ?x_n)\{(v_{1,1} \ldots v_{1,n}) \ldots (v_{k,1} \ldots v_{k,n})\}$ for natural numbers $n$ and $k$ with $v_{i,j} \in \mathbf{IRI} \cup \mathbf{BN}$. *Filter expressions* are of the form FILTER(*boolexp*) where *boolexp* is an algebraic expression encoding the application of filter functions to variables resulting in a Boolean value (for more details see [11]). *Graph patterns* are defined inductively: (i) any basic graph pattern is a graph pattern (ii) if $gp_1$ and $gp_2$ are graph patterns then $\{gp_1\}$ UNION $\{gp_2\}$ is a graph pattern (iii) any sequence of graph patterns, values blocks and filter expressions is again a graph pattern. A SELECT-DISTINCT *query* is a SPARQL 1.1 query of the shape SELECT DISTINCT *varlist* WHERE $\{gp\}$, where $gp$ is a graph pattern and *varlist* is a list of variables occurring in $gp$.

We now define the sematics of SPARQL 1.1 queries, without taking reasoning into account; this is known as *simple entailment* (as opposed to OWL DL entailment, where the OWL axioms are evaluated under OWL Direct Semantics [9]). We define the *evaluation* of path expressions w.r.t. $G$ as a binary relation over

**IRI** ∪ **BN** in an inductive way: $eval_G(p) = \{(u_1, u_2) \mid u_1\ p\ u_2 \in G)\}$ for $p \in$ **IRI**, inverse $eval_G(\hat{}p) = \{(u_2, u_1) \mid (u_1, u_2) \in eval_G(p)\}$, sequence $eval_G(p\ /\ q) = \{(u_1, u_3) \mid (u_1, u_2) \in eval_G(p), (u_2, u_3) \in eval_G(q)\}$, alternative $eval_G(p \mid q) = eval_G(p) \cup eval_G(q)$, Kleene star $eval_G(p^*) = \bigcup_{n \geq 0} eval_G(p^n)$ where $eval_G(p^0) = \{(u, u) \mid u \in$ **IRI** ∪ **BN** occurs in $G\}$ and $eval_G(p^{n+1}) = eval_G(p^n) \circ eval_G(p)$. The *evaluation* $eval_G(bgp)$ of a basic graph pattern *bgp* w.r.t. some RDF graph $G$ is the set of all partial mappings $\mu$ from variables in *bgp* to IRIs or blank nodes of $G$, such that there exists some mapping $\sigma$ from all blank nodes in *bgp* to terms of $G$ for which $\mu(\sigma(bgp)) \in G$. Moreover, $eval_G(\mathsf{VALUES}\ (?x_1 \ldots ?x_n)\{(v_{1,1} \ldots v_{1,n}) \ldots (v_{k,1} \ldots v_{k,n})\}) = \{\{?x_1 \mapsto v_{1,1}, \ldots, ?x_n \mapsto v_{1,n}\}, \ldots \{?x_1 \mapsto v_{k,1}, \ldots, ?x_n \mapsto v_{k,n}\}\}$ and $eval_G(\{gp_1\}\ \mathsf{UNION}\ \{gp_2\}) = eval_G(gp_1) \cup eval_G(gp_2)$. For graph patterns *gp* that are sequences of graph patterns, values blocks and filter expressions $\mathsf{FILTER}(boolexp_1), \ldots, \mathsf{FILTER}(boolexp_\ell)$ we let $eval_G(gp) = \{\mu \mid \mu \in J \wedge \mu(boolexp_1) = true \wedge \ldots \wedge \mu(boolexp_\ell) = true\}$ where $J$ is the join over all $eval_G(block)$ where *block* ranges over all graph patterns and values blocks of the sequence. We say a graph pattern *gp* has a *match* into a graph $G$ if $eval_G(gp) \neq \emptyset$. Finally, the set of *answers* of a SELECT-DISTINCT query SELECT DISTINCT *varlist* WHERE {*gp*} is the set obtained by restricting every partial function $\mu \in eval_G(gp)$ to the variables contained in *varlist*.

## 3 OWL QL: RDF Syntax and Rule-Based Semantics

OWL QL is one of the OWL 2 profiles, which restrict the OWL 2 DL ontology language to ensure that reasoning is tractable [13]. To ensure compatibility with SPARQL, we work only with the RDF representation of OWL QL here [14]. Like OWL 2 DL, OWL QL requires "standard use" of RDFS and OWL vocabulary, i.e., special vocabulary that is used to encode ontology axioms in RDF is strictly distinct from the ontology's vocabulary, and can only occur in specific triple patterns. Only a few special IRIs, such as owl:Thing, can also be used like ontology vocabulary in axioms.

OWL classes, properties, and individuals are represented by RDF elements, where complex class and property expressions are represented by blank nodes. Whether an expression is represented by an IRI or a blank node does not have an impact on ontological entailment, so we ignore this distinction in most cases. OWL 2 DL allows us to use a single IRI to represent an individual, a class, and a property in the same ontology; owing to the restrictions of standard use, it is always clear which meaning applies in a particular case. Hence we will also work with one single set of IRIs.

Next, we define the constrains that an RDF graph has to satisfy to represent an OWL QL ontology. To this end, consider a fixed RDF graph $G$. A *property expression in G* is an IRI or a blank node _:b that occurs in a pattern {_:b owl:inverseOf *P*} with $P \in$ **IRI**. We use **PRP** for the set of all property elements in a given RDF graph. OWL QL further distinguishes two types of class expressions with different syntactic constraints. The set **SBC** of *subclasses in G* consists of all IRIs and all blank nodes _:b that occur in a pattern {_:b owl:onProperty *P*; owl:someValuesFrom owl:Thing}, where $P \in$ **PRP**. The set **SPC** of *superclasses in G* is defined recursively as follows. An element $x$ is in **SPC** if it is in **IRI**, or if it is in **BN** and $G$ contains one of the following patterns:

- {$x$ owl:onProperty **PRP**; owl:someValuesFrom $y$} where $y \in$ **SPC**;
- {$x$ owl:intersectionOf $(y_1, \ldots, y_n)$} where $y_1, \ldots, y_n \in$ **SPC**;

– {$x$ owl:complementOf $y$} where $y \in$ **SBC**.

$G$ is an *OWL QL ontology* may use the following triple patterns to encode *axioms*:

– {**IRI PRP IRI**}
– {**IRI** rdf:type **SPC**}
– {**SBC** rdfs:subClassOf **SPC**}
– {**SBC** owl:equivalentClass **SBC**}
– {**SBC** owl:disjointWith **SBC**}
– {**PRP** rdfs:range **SPC**}
– {**PRP** rdfs:domain **SPC**}
– {**PRP** rdfs:subPropertyOf **PRP**}
– {**PRP** owl:equivalentProperty **PRP**}
– {**PRP** owl:inverseOf **PRP**}
– {**PRP** owl:propertyDisjointWith **PRP**}
– {**IRI** owl:differentFrom **IRI**}
– {**BN** rdf:type owl:AllDisjointClasses; owl:members (**SBC**,...,**SBC**)}
– {**BN** rdf:type owl:AllDisjointProperties; owl:members (**PRP**,...,**PRP**)}
– {**BN** rdf:type owl:AllDifferent; owl:members (**IRI**,...,**IRI**)}

$G$ is an OWL QL ontology if every triple in $G$ is part of a unique axiom or a unique complex class or property definition used in such axioms. For simplicity, we ignore triples used in annotations or ontology headers. Moreover, we do not consider the OWL QL property characteristics symmetry, asymmetry, and global reflexivity. Asymmetry and reflexivity are not a problem, but their explicit treatment would inflate our presentation considerably. Symmetry, in contrast, cannot be supported with SPARQL 1.1, as we will show in Section 6. This is no major limitation of our approach, since symmetry can be expressed using inverses. This shows that rewritability of an ontology language does not depend on ontological expressiveness alone.

The semantics of OWL QL is inherited from OWL DL. However, since OWL QL does not support any form of disjunctive information, one can also describe the semantics by defining a *universal model*, i.e., a structure that realizes all entailments of an ontology but no additional entailments. Such a "least model" exactly captures the semantics of an ontology.

To define a universal model for OWL QL, we define a set of RDF-based inference rules, similar to the rules given for OWL RL in the standard [13]. In contrast to OWL RL, however, the application of rules can introduce new elements to an RDF graph, and the universal model that is obtained in the limit is not finite in general. Indeed, our goal is not to give a practical reasoning algorithm, but to define the semantics of OWL QL in a way that is useful for analyzing the correctness of the rewriting algorithms we introduce.

The main rules for reasoning in OWL QL are defined in Table 1. A rule is *applicable* if the premise on the left matches the current RDF graph and the conclusion on the right does not match the current graph; in this case, the conclusion is added to the graph. In case of rule (2), this requires us to create a fresh blank node. In all other cases, we only add new triples among existing elements. Rules like (3) are actually schemas for an infinite number of rules for lists of any length $n$ and any index $i \in \{1,\dots,n\}$. Rules (15)–(16) cover owl:Thing and owl:topObjectProperty, which lead to conclusions that are true for "all" individuals. To ensure standard use, we cannot simply assert $x$ rdf:type owl:Thing for *every* IRI $x$, and we restrict instead to IRIs that are used as individuals in the ontology. We define INDIVIDUAL($x$) to be the SPARQL pattern {$x$ rdf:type owl:NamedIndividual} UNION {$x$ rdf:type ?C . ?C rdf:type owl:Class} UNION {$x$ ?P ?Y . ?P rdf:type owl:ObjectProperty} UNION {?Y ?P $x$ . ?P rdf:type owl:ObjectProperty}. Note that this also covers any newly introduced individuals.

**Table 1.** RDF inference rules for OWL QL

| | |
|---|---|
| → [] rdf:type owl:Thing | (1) |
| ?X rdf:type [owl:onProperty ?P; owl:someValuesFrom ?C] → ?X ?P [rdf:type ?C] | (2) |
| ?X rdf:type [owl:intersectionOf (?C1, ..., ?C$i$, ..., ?C$n$)] → ?X rdf:type ?C$i$ | (3) |
| ?X rdf:type ?C . ?C rdfs:subClassOf ?D → ?X rdf:type ?D | (4) |
| ?X rdf:type ?C . ?C owl:equivalentClass ?D → ?X rdf:type ?D | (5) |
| ?X rdf:type ?C . ?D owl:equivalentClass ?C → ?X rdf:type ?D | (6) |
| ?X ?P ?Y . | |
| ?C owl:onProperty ?P; owl:someValuesFrom owl:Thing → ?X rdf:type ?C | (7) |
| ?X ?P ?Y . ?P rdfs:domain ?C → ?X rdf:type ?C | (8) |
| ?X ?P ?Y . ?P rdfs:range ?C → ?Y rdf:type ?C | (9) |
| ?X ?P ?Y . ?P owl:inverseOf ?Q → ?Y ?Q ?X | (10) |
| ?X ?P ?Y . ?Q owl:inverseOf ?P → ?Y ?Q ?X | (11) |
| ?X ?P ?Y . ?P rdfs:subPropertyOf ?Q → ?X ?Q ?Y | (12) |
| ?X ?P ?Y . ?P owl:equivalentProperty ?Q → ?X ?Q ?Y | (13) |
| ?X ?P ?Y . ?Q owl:equivalentProperty ?P → ?X ?Q ?Y | (14) |
| INDIVIDUAL(?X) → ?X rdf:type owl:Thing | (15) |
| ?X rdf:type owl:Thing . ?Y rdf:type owl:Thing → ?X owl:topObjectProperty ?Y | (16) |

**Definition 1.** *The* chase $G'$ *of an OWL QL ontology G is a possibly infinite RDF graph obtained from G by fair application of the rules of Tables 1, meaning that every rule that is applicable has eventually been applied.*

Finally, some features of OWL QL can only make the ontology inconsistent, but not introduce any other kinds of positive entailments. According patterns are shown in Table 2. If any of these match, the ontology is inconsistent, every OWL axiom is a logical consequence, and there is no universal model.

**Theorem 1.** *Consider an OWL QL ontology G with chase $G'$, and a basic graph pattern P. A variable mapping $\mu$ is a solution for P over G under the OWL DL entailment regime if and only if either*

1. *$\mu$ is a solution for P over $G'$ under simple entailment, or*
2. *one of the rules of Table 2 matches $G'$.*

## 4 QL Reasoning with SPARQL Property Expressions

Next, we define SPARQL 1.1 queries to solve standard reasoning tasks of OWL QL. We start with simple cases and then consider increasingly complex reasoning problems.

**Table 2.** RDF inference patterns for inconsistency in OWL QL

| | |
|---|---|
| ?X owl:bottomObjectProperty ?Y | (17) |
| ?X rdf:type owl:Nothing | (18) |
| ?X rdf:type ?C . ?X rdf:type [owl:complementOf ?C] | (19) |
| ?X rdf:type ?C . ?X rdf:type ?D . ?C owl:disjointWith ?D | (20) |
| ?X rdf:type ?C$i$ . ?X rdf:type ?C$j$. | |
| _:b rdf:type owl:AllDisjointClasses; owl:members (?C1, ..., ?C$i$, ..., ?C$j$, ..., ?C$n$) | (21) |
| ?X ?P ?Y . ?X ?Q ?Y . ?P owl:propertyDisjointWith ?Q | (22) |
| ?X ?P$i$ ?Y . ?X ?P$j$ ?Y. | |
| _:b rdf:type owl:AllDisjointProperties; owl:members (?P1, ..., ?P$i$, ..., ?P$j$, ..., ?P$n$) | (23) |
| ?X owl:differentFrom ?X | (24) |
| _:b rdf:type owl:AllDifferent; owl:members (?I1, ..., ?X, ..., ?X, ..., ?I$n$) | (25) |

We first focus on the property hierarchy. An axiom of the form $p$ rdfs:subPropertyOf $q$ is entailed by an ontology $G$ if, for newly introduced individuals $a$ and $b$, $G \cup \{a\ p\ b\}$ entails $\{a\ q\ b\}$. By Theorem 1, the rules of Section 3 represent all possibilities for deriving this information. In this particular case, we can see that only rules (10)–(14) in Table 1 can derive a triple of the form $a\ q\ b$, where q is a regular property. The case $q$ =owl:topObjectProperty is easy to handle, since $p$ rdfs:subPropertyOf owl:topObjectProperty is always true (which is also shown by rules (15) and (16)). In addition, it might be that $G \cup \{a\ p\ b\}$ is inconsistent, implied by rules of Table 2; we will ignore this case for now, since it requires more powerful reasoning.

**Definition 2.** *We introduce* sPO, invOf, *and* eqP *as abbreviations for* rdfs:subPropertyOf, owl:inverseOf, *and* owl:equivalentProperty, *respectively, and define the following composite property path expressions* SpoEqp := (sPO | eqP | ^eqP), Inv := (invOf | ^invOf), subPropertyOf := (SpoEqp | (Inv / SpoEqp* / Inv))*, *as well as* subInvPropertyOf := SpoEqp* / Inv / subPropertyOf. *Moreover, for an arbitrary term $x$, let* univProperty[$x$] *be the pattern* {owl:topObjectProperty (SpoEqp | Inv)* $x$}.

The pattern subPropertyOf does not check for property subsumption that is caused by the inconsistency rules in Table 2, but it can be used to check for subsumptions related to owl:topObjectProperty. This relies on the following correctness property of the pattern univProperty[$p$]. We provide a particularly detailed proof here, since many of our later correctness properties will rely on similar arguments.

**Lemma 1.** *Consider a consistent OWL QL ontology $G$ with property $p \in$* **PRP**. *Then $G$ entails* owl:topObjectProperty rdfs:subPropertyOf $p$ *iff the pattern* univProperty[$p$] *matches $G$.*

*Proof.* For the "if" direction, assume that the pattern univProperty[$p$] matches $G$. We need to show that $G$ entails owl:topObjectProperty rdfs:subPropertyOf $p$. Using Theorem 1, this is equivalent to the claim: the triple _:a $p$ _:b can be derived by applying the

deduction rules of Table 1 to $G \cup \{\_\text{:a owl:topObjectProperty \_:b}\}$. In particular, we know that the latter is consistent, since otherwise $G$ would clearly be inconsistent as well.

Thus assume a path $(\textsc{SpoEqp} \mid \textsc{Inv})^n$ of length $n \geq 0$ from owl:topObjectProperty to $p$. We show the claim by induction on $n$. For $n = 0$, $p$ = owl:topObjectProperty and the claim is immediate. For $n > 0$, let $p'$ be the element in the that is reached after $n-1$ steps in the path (and for which the claim was already shown by induction). We distinguish cases according to which of the optional properties $q$ in the pattern connects $p'$ to $p$:

- If $q$ = rdfs:subPropertyOf, then we can apply rule (12) to derive _:a $p$ _:b from _:a $p'$ _:b. Since the latter can be derived from $G \cup \{\_\text{:a owl:topObjectProperty \_:b}\}$ by the induction hypothesis, the claim follows.
- The cases $q$ = owl:equivalentProperty and $q$ = ^owl:equivalentProperty are similar using rules (13) and (14), respectively.
- If $q$ = owl:inverseOf, we can use the same argument as before to obtain a derivation of _:a $p$ _:b from $G \cup \{\_\text{:b owl:topObjectProperty \_:a}\}$, using rule (10) in the last step. Note that we apply the induction hypothesis to an input with _:a and _:b swapped. To get the desired derivation, we note that _:b owl:topObjectProperty _:a can be derived from _:a owl:topObjectProperty _:b by applying rule (15) to _:a and _:b, followed by rule (16).
- The case $q$ = ^owl:inverseOf is again similar, using rule (11).

For the "only if" direction, assume that _:a $p$ _:b can be derived from the ontology $G \cup \{\_\text{:a owl:topObjectProperty \_:b}\}$ by applying the deduction rules. This can only be accomplished by applying rules (12)–(16). Moreover, we can assume without loss of generality that (15) and (16) are only applied at the beginning of the derivation to obtain {_:b owl:topObjectProperty _:a} from {_:a owl:topObjectProperty _:b} (the latter being the only interesting derivation that rule (16) could produce here). Thus, to simplify our claim, consider a derivation of _:a $p$ _:b can be derived from $G \cup \{\_\text{:a owl:topObjectProperty \_:b}, \_\text{:b owl:topObjectProperty \_:a}\}$.

The proof is by induction on the length $\ell$ of this derivation We claim that there is a path of the form $(\textsc{SpoEqp} \mid \textsc{Inv})^n$ of length $n \geq 0$ from owl:topObjectProperty to $p$.

If $\ell = 0$, $p$ = owl:topObjectProperty and the claim is immediate (with $n = 0$). For $\ell > 0$, we distinguish cases according to the rule applied in the last step of the derivation:

- Rule (12), (13), or (14) applied to a previous consequence {_:a $p'$ _:b}. Then $G$ contains a triple $p'$ $q$ $p$ for $q$ = rdfs:subPropertyOf, $q$ = owl:equivalentProperty, or $q$ = ^owl:equivalentProperty, respectively. By the induction hypothesis, there is a path as in the claim from owl:topObjectProperty to $p'$. We can extend this path by $p'$ $q$ $p$.
- Rule (10) or (11) applied to a previous consequence {_:b $p'$ _:a}. Then $G$ contains a triple $p'$ $q$ $p$ for $q$ = owl:inverseOf or $q$ = ^owl:inverseOf, respectively. The induction hypothesis applies since we can always swap _:a and _:b in a derivation. Thus there is a path as in the claim from owl:topObjectProperty to $p'$. We can extend this path by $p'$ $q$ $p$.
- Rules (15) cannot occur by our assumption on the derivation. □

The following result shows the essential correctness property of subPropertyOf on consistent ontologies.

**Proposition 1.** *Consider an OWL QL ontology $G$ with properties $p,q \in$* **PRP** *such that $G \cup \{\_:a\ p\ \_:b\}$ is consistent. Then $G$ entails $p$* rdfs:subPropertyOf *$q$ iff the pattern $\{p$* subPropertyOf *$q\}$* UNION univProperty*$[q]$ matches $G$.*

*Proof.* For the "if" direction, we have to show that the above described calculus allows to derive the triple _:a $q$ _:b from $G \cup \{\_:a\ p\ \_:b\}$ whenever the pattern $\{p$ subPropertyOf $q\}$ UNION univProperty$[q]$ matches $G$. We consider both cases of the UNION expression.

First, let $p$ subPropertyOf $q$ be the matching pattern of the query, that is, we find some $n \in \mathbb{N}$ and a path from $p$ to $q$ matching the regular expression (SpoEqp | (Inv / SpoEqp* / Inv))$^n$. We show the claim via an induction over $n$. For $n = 0$ we obtain $p = q$, therefore _:a $q$ _:b holds by assumption.

For the induction step, assume the claim holds for $n$ and consider a path matching the expression (SpoEqp | (Inv / SpoEqp* / Inv))$^{n+1}$, which means that there is an individual $p'$ such that there is a path matching (SpoEqp | (Inv / SpoEqp* / Inv))$^n$ from $p$ to $p'$ and a path matching (SpoEqp | (Inv / SpoEqp* / Inv)) from $p'$ to $q'$. By induction hypothesis, there is a derivation for _:a $p'$ _:b (†). Now we further analyze the path from $p'$ to $q'$:

- If SpoEqp matches this path then, for each of the possible sub-cases of SpoEqp (viz. rdfs:subPropertyOf, owl:equivalentProperty, or ^ owl:equivalentProperty) we find an appropriate rule (namely rule (12), (13), or (14) of Table 1, respectively) to derive _:a $q$ _:b from _:a $p'$ _:b.
- If the path from $p'$ to $q$ is matched by (Inv / SpoEqp* / Inv), there are individuals $q'$ and $q''$, such that there are (i) an Inv path from $p'$ to $q'$, (ii) a path from $q'$ to $q''$ matching SpoEqp$^k$ for some $k \geq 0$, and (iii) an Inv path from $q''$ to $q$.
  From (†) and (i), we can obtain _:b $q'$ _:a (‡) via rule 10 or 11. Given (‡) and (ii), we can perform another induction over $k$, recalling the above argument regarding SpoEqp, to arrive at _:b $q''$ _:a. Now, exploiting (iii) and rule 10 or 11 once more, we finally obtain _:a $q'$ _:b as claimed.

For the second part of the UNION expression, we note that from _:a $p$ _:b, we can infer _:a owl:topObjectProperty _:b by means of rule 15 and 16. Then, we can invoke Lemma 1 to arrive at _:a $q$ _:b as claimed.

For the "only if" direction, assume $G$ is such that _:a $q$ _:b can be derived from $G \cup \{\_:a\ p\ \_:b\}$ by applying the deduction rules. If _:a $q$ _:b can be derived from $G \cup \{\_:a$ owl:topObjectProperty _:b$\}$, then the claim follows from Lemma 1. For the remaining case, we can restrict to derivations of _:a $q$ _:b using rules (10)–(14). Clearly, any such derivation is linear, with each rule applying to a triple in $G$ and a triple of the form _:a $q'$ _:b or _:b $q'$ _:b. Let $p = q_0, \ldots, q_n = q$ be the sequence of properties used in the latter. Only rules (10) and (11) can swap the order of _:a and _:b, hence there must be an even number of applications of these rules in the derivation. It is easy to see that the expression subPropertyOf hatches exactly these sequences of properties $q_0 \ldots q_n$. □

We will extend this to cover the inconsistent case in Theorem 2 below. First, however, we look at entailments of class subsumptions. In this case, the main rules are (2)–(9). However, several of these rules also depend on property triples derived by rules (10)–(14), and we apply our results on property subsumption to take this into account.

**Definition 3.** *Let* eqC *and* sCO *abbreviate* owl:equivalentClass *and* rdfs:subClassOf*, respectively. We define property path expressions*

- intListMember := (owl:intersectionOf / rdf:rest* / rdf:first),
- someProp := (owl:onProperty / subPropertyOf / (^owl:onProperty | rdfs:domain)),
- somePropInv := (owl:onProperty / subInvPropertyOf / rdfs:range),
- subClassOf := (sCO | eqC | ^eqC | intListMember | someProp | somePropInv)*.

*Moreover, we let* univClass[$x$] *denote the pattern* {owl:Thing subClassOf $x$} UNION {owl:topObjectProperty ((SpoEqp | Inv)* / (^owl:onProperty | rdfs:domain | rdfs:range)/ subClassOf) $x$}

We can use subClassOf to check if a superclass expression in $G$ is subsumed by a subclass expression in $G$; in particular, this applies to class names. As before, we exclude the possibility that one of the classes is incoherent (i.e., entailed to be equivalent to owl:Nothing).

**Proposition 2.** *Consider an OWL QL ontology $G$ with classes $c \in$ **SPC** and $d \in$ **SBC** such that $G \cup$ {_:a rdf:type $c$} is consistent. Then $G$ entails $c$* rdfs:subClassOf *$d$ iff the pattern* {$c$ subClassOf $d$} UNION univClass[$d$] *matches $G$.*

*Proof.* For the "if" direction, we have to show that the above described calculus allows to derive the triple _:a rdf:type $d$ from $G \cup$ {_:a rdf:type $c$} whenever the pattern {$c$ subClassOf $d$} UNION univClass[$d$] matches $G$. We consider both cases of the UNION expression.

First, let $c$ subClassOf $d$ be the matching pattern of the query, that is, we find some $n \in \mathbb{N}$ and a path from $c$ to $d$ matching the regular expression (sCO | eqC | ^eqC | intListMember | someProp | somePropInv)$^n$. We show the claim via an induction over $n$. For $n = 0$ we obtain $a = b$, therefore _:a rdf:type $d$ holds by assumption. For the induction step, assume the claim holds for $n$ and consider a path matching the expression (sCO | eqC | ^eqC | intListMember | someProp | somePropInv)$^{n+1}$ which means that there is a class $c'$ such that there is a path matching (sCO | eqC | ^eqC | intListMember | someProp | somePropInv)$^n$ from $c$ to $c'$ and a path matching (sCO | eqC | ^eqC | intListMember | someProp | somePropInv) from $c'$ to $d'$. By induction hypothesis, we can deduce that _:a rdf:type $c'$ must hold (†). Now we further analyze the path from $c'$ to $d'$ by separately considering the 6 disjunctive options:

- $c'$ sCO $d'$: we can use the induction hypothesis and rule (4) to infer _:a rdf:type $d'$.
- $c'$ eqC $d'$: we can use the induction hypothesis and rule (5) to infer _:a rdf:type $d'$.
- $c'$ ^eqC $d'$: we can use the induction hypothesis and rule (6) to infer _:a rdf:type $d'$.
- $c'$ intListMember $d'$: presuming $G$ to be a well-formed OWL QL graph, the regular expression intListMember does only match inside a structure of the shape $x$ owl:intersectionOf $(x_1, \ldots, x_n)$ and connects $x$ with some $x_i$. Then we can use the induction hypothesis and rule (3) to infer _:a rdf:type $d'$.
- $c'$ someProp $d'$: in this case there must exist $p$ and $q$ connected by a path matching subPropertyOf such that $G$ also contains the triples $c'$ owl:onProperty $p$ and either $d'$ owl:onProperty $p$ or $p$ rdfs:domain $d'$. Again assuming well-formedness of $G$ we can apply rule (2) to infer _:a $p$ _:b for some fresh bnode _:b. Consequently, due to

**Table 3.** Pattern EMPTYCLASS[$x$] for detecting empty classes.

```
x (sCO | eqC | ^eqC | INTLISTMEMBER | owl:someValuesFrom |
    (owl:onProperty / (INV | SPOEQP)* / (^owl:onProperty | rdfs:domain | rdfs:range))* ?C . {
   {?C SUBCLASSOF owl:Nothing} UNION
   {?C SUBCLASSOF ?D1 {{?C SUBCLASSOF ?D2} UNION UNIVCLASS[?D2]} {
      {?D1 DISJOINTCLASSES ?D2} UNION
      {?V rdf:type owl:AllDisjointClasses . TWOMEMBERS[?V, ?D1, ?D2]}
   }} UNION
   {?C (owl:onProperty / (INV | SPOEQP)*) ?P . {
      {?P SUBPROPERTYOF owl:bottomObjectProperty} UNION
      {?P SUBPROPERTYOF ?Q1 {{?P SUBPROPERTYOF ?Q2} UNION UNIVPROPERTY[?Q2]} {
         {?Q1 (owl:propertyDisjointWith | ^owl:propertyDisjointWith) ?Q2} UNION
         {?V rdf:type owl:AllDisjointProperties . TWOMEMBERS[?V, ?Q1, ?Q2]}
      }
   }}
}
```

Property 1, we can infer _:a $q$ _:b. Finally applying either rule (7) or rule (8), we arrive at _:a rdf:type $d'$.

– $c'$ SOMEPROPINV $d'$: in this case there must exist $p$ and $q$ connected by a path matching SUBINVPROPERTYOF such that $G$ also contains the triples $c'$ owl:onProperty$p$ and $p$rdfs:range $d'$. We can apply rule (2) to infer _:a $p$ _:b for some fresh bnode _:b. Consequently, exploiting the argument in the proof of Property 1, we can infer _:b $q$ _:a. Finally applying rule (9), we arrive at _:a rdf:type $d'$.

The case for UNIVCLASS[$d$] being the matching pattern can be shown in a way analogous to the one above, additionally using Rule 15 and rule (16) for the base cases.

For the "only if" direction, we have to analyze all possible proofs. For this, it is helpful to distinguish two cases: one where a proof can be found that directly applies the rule (15) to all occurrences of proof-tree leafs carrying _:a rdf:type $d$. In such a case, a match to UNIVCLASS[$d$] can be constructed from the proof. In all other cases we can construct a match to {$c$ SUBCLASSOF $d$} in way very analogous to the argument in Proposition 2. □

It remains to identify classes that are incoherent, i.e., for which $c$ rdfs:subClassOf owl:Nothing is entailed. To do this, we need to consider the patterns of Table 2.

**Definition 4.** *For arbitrary terms x, y, and z, let* TWOMEMBERS[$x,y,z$] *be the pattern* {$x$ (owl:members / rdf:rest*) ?W . ?W rdf:first $y$ . ?W (rdf:rest$^+$ / rdf:first) $z$}*, and let* DISJOINTCLASSES *be the property path expression* (owl:disjointWith | ^owl:disjointWith | owl:complementOf | ^owl:complementOf)*. The query pattern* EMPTYCLASS[$x$] *is defined as in Table 3, and the query pattern* EMPTYPROPERTY[$x$] *is defined as in Table 4.*

As their name suggests, the patterns of the previous definition allow us to detect classes and properties that must be empty in every model of the ontology. To prove this, we first make some simpler observations:

**Table 4.** Pattern EMPTYPROPERTY[$x$] for detecting empty properties.

```
x (INV | SPOEQP | (^owl:onProperty /
    (sCO | eqC | ^eqC | INTLISTMEMBER | owl:someValuesFrom)* / owl:onProperty))* ?P . {
   {?P SUBPROPERTYOF owl:bottomObjectProperty} UNION
   {?P SUBPROPERTYOF ?Q1 {{?P SUBPROPERTYOF ?Q2} UNION UNIVPROPERTY[?Q2]} {
      {?Q1 (owl:propertyDisjointWith | ^owl:propertyDisjointWith) ?Q2} UNION
      {?V rdf:type owl:AllDisjointProperties . TWOMEMBERS[?V, ?Q1, ?Q2]}
   }} UNION
   {?P ((^owl:onProperty | rdfs:domain | rdfs:range) / SUBCLASSOF) ?C . {
      {?C SUBCLASSOF owl:Nothing} UNION
      {?C SUBCLASSOF ?D1 {{?C SUBCLASSOF ?D2} UNION UNIVCLASS[?D2]} {
         {?D1 DISJOINTCLASSES ?D2} UNION
         {?V rdf:type owl:AllDisjointClasses . TWOMEMBERS[?V, ?D1, ?D2]}
      }
   }}
}
```

**Lemma 2.** *The pattern* TWOMEMBERS[$x,y,z$] *matches an ontology $G$ iff $G$ contains an RDF list $x$ with two distinct elements $y$ and $z$.*

**Lemma 3.** *Consider a consistent OWL QL ontology $G$ with class $c$. Then $G \cup$* {_:a rdf:type $c$} *is inconsistent iff the pattern* EMPTYCLASS[$c$] *matches $G$.*

*Proof.* The general structure of the proof is as in Lemma 1, but with a lot more cases to consider. We sketch the arguments in order to avoid getting lost in details here.

First, we can show a property of the first two lines of the pattern in Table 3. Namely, the variable ?C in the pattern generally represents a class that must be non-empty whenever the class $x$ ($c$ in our claim) is non-empty. Formally: $G$ has a match for the pattern $c$ (rdfs:subClassOf | owl:equivalentClass | ^owl:equivalentClass | INTLISTMEMBER | owl:someValuesFrom | (owl:onProperty / (INV | SPOEQP)* / (^owl:onProperty | rdfs:domain | rdfs:range))* $d$ iff $G \cup \{d$ rdfs:subClassOf owl:Nothing$\}$ is consistent but $G \cup \{$_:a rdf:type $c, d$ rdfs:subClassOf owl:Nothing$\}$ is inconsistent.

This is shown by easy inductions as in Lemma 1. Most importantly, we need to observe that non-emptiness of classes can directly follow from the rules (1)–(9), and (15). The cases of (1) and (15) are not of interest, since they infer non-emptiness of owl:Thing: $d$ in our claim cannot be a superclass of owl:Thing as this would make $G \cup \{d$ rdfs:subClassOf owl:Nothing$\}$ inconsistent. Of the remaining rules, (2)–(6) are are covered by the options rdfs:subClassOf, owl:equivalentClass, ^owl:equivalentClass, INTLISTMEMBER, and owl:someValuesFrom in the pattern, respectively.

For the remaining cases, we need to take derivations of property assertion triples into account. The only relevant rule to derive such triples from premises of the form _:a rdf:type $c'$ is (2). After this, further property triples are inferred as in Proposition 1 using rules (10)–(14), corresponding (as shown before) to the expression (INV | SPOEQP)*. Again, owl:topObjectProperty is not of interest here since we assume $G \cup \{d$ rdfs:subClassOf owl:Nothing$\}$ to be consistent. Finally, property assertion triples

can be used to transfer new class assertion triples in rules (7)–(9), corresponding to the final options (^owl:onProperty | rdfs:domain | rdfs:range) in the pattern. This correspondence of rules and pattern can be exploited to obtain the desired result by two inductions, as before.

The remaining parts of the pattern in Table 3 lists relevant cases in which the non-emptiness of the class represented by ?C would lead to inconsistency. These cases correspond to the patterns in Table 2. Cases (18)–(21) are covered by the patterns in the third to seventh line of Table 2, where we use (reasoning similar to) Proposition 2 for the essential correctness of subpatterns univClass[?D2] and subClassOf . Cases (17), (22) and (23) are covered by the remaining lines, where we use Lemma 1 and Proposition 1 for the essential correctness of subpatterns univProperty[?Q2] and subPropertyOf. Cases (24) and (25) can only be violated by $G$ initially and thus do not require checking here.

Lemma 2 provides the essential correctness of the pattern twoMembers$[x,y,z]$ used in several places. For cases (17), (22) and (23), we need to consider property assertion triples that are derived from the non-emptiness of ?C; the pattern used to find a non-empty property ?P is the same pattern that already occurred on the second line, and the same reasoning applies.

Using the, by now obvious, correspondences between inference rules and patterns, we can thus prove the overall claim. □

**Lemma 4.** *Consider a consistent OWL QL ontology $G$ with property $p$. Then $G \cup$ {_:a $p$ _:b} is inconsistent iff the pattern* emptyProperty$[p]$ *matches $G$.*

*Proof.* The proof follows exactly the same arguments as the proof of Lemma 3. Indeed, many of the subpatterns used are the same, with the main difference being that we now start the derivation from property assertion triples rather than from class assertion triples. □

We can now completely express OWL QL schema reasoning in SPARQL 1.1:

**Theorem 2.** *An OWL QL ontology $G$ is inconsistent iff it has a match for the pattern*

```
{?X rdf:type ?C . emptyClass[?C]} UNION {?X ?P ?Y . emptyProperty[?P]} UNION
{?X owl:differentFrom ?X} UNION
{?V rdf:type owl:AllDifferent . twoMembers[?V, ?X, ?X]}.
```
(26)

*$G$ entails $c$* rdfs:subClassOf *$d$ for $c \in$* **SPC** *and $d \in$* **SBC** *iff $G$ is either inconsistent or has a match for the pattern*

```
{c subClassOf d} UNION univClass[d] UNION emptyClass[c].
```
(27)

*$G$ entails $x$* rdf:type *$c$ iff $G$ is either inconsistent or has a match for the pattern*

```
{{x (rdf:type / subClassOf) c} UNION
 {x ?P ?Y . ?P (subPropertyOf / (^owl:onProperty | rdfs:domain) / subClassOf) c} UNION
 {?Y ?P x . ?P (subPropertyOf / rdfs:range / subClassOf) c}
} UNION univClass[c]
```
(28)

*G entails p* rdfs:subPropertyOf *q for* $p,q \in$ **PRP** *iff G is either inconsistent or has a match for the pattern*

$$\{p \text{ subPropertyOf } q\} \text{ UNION } \text{univProperty}[q] \text{ UNION } \text{emptyProperty}[p]. \tag{29}$$

*G entails x p y iff G is either inconsistent or has a match for the pattern*

$$\{x \text{ ?R } y \text{ . ?R subPropertyOf } p\} \text{ UNION } \{y \text{ ?R } x \text{ . ?R subInvPropertyOf } p\} \\ \text{UNION } \text{univProperty}[p]. \tag{30}$$

*Proof.* In each of the cases, we can show correctness using similar techniques as in the proof of Lemma 1 and the subsequent proofs shown in this section. We consider each case individually.

For (26), correctness is an easy consequence of Lemma 3 and Lemma 4, together with the observation that the two last lines of (26) correspond to the cases (24) and (25) in Table 2. We need to use rules (1) (for the first time) and (16) to see that (26) also covers the cases where owl:Thing is a subclass of owl:Nothing, or where owl:topObjectProperty is a subproperty of owl:bottomObjectProperty.

For (27), correctness follows from Proposition 2 and Lemma 3.

For (28), we see from Proposition 2 why the first and last line are correct. However, Proposition 2 only covers derivations that start from a class assertion triple. When checking for the type of an individual in (28), the derivation might also start at property assertion triples given in the ontology. Our arguments in the proof of Proposition 2 covered property assertion triples, but only as an intermediate stage of the derivation. It is not hard to see that the second and third line of (28) are similar to the respective expressions someProp and somePropInv in Definition 3, and correctness is shows using the same reasoning as in the proof of Proposition 2.

For (29), correctness follows from Proposition 1 and Lemma 4.

For (30), correctness is an easy consequence of the same reasoning as in the proof of Proposition 1, together with the easy observation that subInvPropertyOf is similar to subPropertyOf but swaps the sides. Note that only the rules (10)–(14) are relevant for normal derivations (not involving owl:topObjectProperty, which is covered by Lemma 1). □

## 5 OWL QL Query Rewriting with SPARQL 1.1

We now turn towards query answering over OWL QL ontologies using SPARQL 1.1. Research in OWL QL query answering typically considers the problem of answering *conjunctive queries* (CQs), which are conjunctions of OWL property and class assertions that use variables only in the place of individuals, not in the place of properties or classes. Conjunction can easily be represented by a Basic Graph Pattern in SPARQL, yet CQs are not a subset of SPARQL, since they also support existential quantification of variables. Normal query variables are called *distinguished* while existentially quantified variables are called *non-distinguished*. Distinguished variables can only bind to elements of the ontology, whereas for non-distinguished variables it suffices if the ontology implies that some binding must exist.

*Example 1.* Consider an OWL ontology with the assertion :peter rdf:type :Person and the axiom :Person rdfs:subClassOf [owl:onProperty :father; owl:someValuesFrom :Person]. This implies that :peter has some :father but the ontology may not contain any element of which we know that it plays this role. In this case, the SPARQL pattern {?X :father ?Y} would not have a match with ?X = :peter under OWL DL entailment. In contrast, if the variable ?Y were non-distinguished, the query would match with ?X = :peter (and ?Y would not receive any binding).

SPARQL can only express CQs where all variables are distinguished. To define this fragment of SPARQL, recall that the OWL DL entailment regime of SPARQL 1.1 requires every variable to be *declared* for a certain type (individual, object property, datatype property, or class) [9]. This requirement is the analogue of "standard use" on the level of query patterns, and it allows us to focus on instance retrieval here. We thus call a Basic Graph Pattern *P CQ-pattern* if: (1) *P* does not contain any OWL, RDF, or RDFS URIs other than rdf:type in property positions, (2) all variables in *P* are declared as required by the OWL DL entailment regime, (3) property variables occur only in predicate positions, and (4) class variables occur only in object positions of triples with predicate rdf:type. Rewriting CQ-patterns is an easy application of Theorem 2:

**Definition 5.** *For a triple pattern e* rdf:type *c, the rewriting* $[\![x$ rdf:type $c]\!]$ *is the graph pattern* (28) *as in Theorem 2; for a triple pattern x p y, the rewriting* $[\![x\ p\ y]\!]$ *is the graph pattern* (30)*. The rewriting* $[\![P]\!]$ *of a CQ-pattern P is obtained by replacing every triple pattern s p o in P by* $\{[\![s\ p\ o]\!]\}$.

**Theorem 3.** *If G is the RDF graph of a consistent OWL QL ontology, then the matches of a CQ-pattern P on G under OWL DL entailment are exactly the matches of* $[\![P]\!]$ *on G under simple entailment.*

### 5.1 Rewriting General Conjunctive Queries

We now explain the additional aspects that we need to take into account for computing answers to CQs with non-distinguished variables, and give an intuitive overview of our rewriting approach. A general challenge that we have to address is that classical query rewriting for OWL QL may lead to exponentially many queries, owing to the fact that many non-deterministic choices have to be made to find a query match. Some of these choices depend on the ontology, e.g., on the depth of the class hierarchy, and are naturally represented in (small) SPARQL 1.1 queries in our approach. Other choices, however, depend on the query, e.g., the decision which variables should be identified (query factorization). It is not immediately clear how to represent these choices in a polynomial query, even when using path expressions. Our solution depends on the creative use of the VALUES feature of SPARQL 1.1.

As explained before, non-distinguished variables can be matched to inferred individuals that are not named in the ontology. The chase introduced in Section 3 still captures this more general notion of query answering. The only rule to infer new individuals is (2), which introduces fresh bnodes that we call *anonymous individuals*. The elements of the original ontology (bnode or not) are *named individuals*. It is well known that a QL ontology *G* entails a CQ *q* if and only if there is a match from *q* to the (possibly

infinite) chase of $G$ such that all distinguished variables are mapped to named individuals. Non-distinguished variables can be mapped to either named or anonymous individuals.

To represent the match of a query variable $x$ in the rewritten query, we introduce a SPARQL variable ?Mx. For named individuals, ?Mx can bind to the individual in the RDF graph. However, if $x$ is non-distinguished, then it could match to anonymous individuals, which are not represented by any individual in RDF. In this case, we bind ?Mx to the bnode _:b representing the OWL property restriction _:b owl:onProperty ?P; owl:someValuesFrom ?C that was used in rule (2) to generate the anonymous individual. Indeed, all class and property assertions that are derived for the anonymous individual can be deduced from ?P and ?C only, so this binding allows us to check query conditions.

However, the bnode _:b does not determine the identity of the anonymous individual, since infinitely many anonymous individuals can be generated from the same OWL property restriction. Example 1 illustrates this: every person has another person as is its father, *ad infinitum*. Nevertheless, the query :peter :father ?Z . ?Z :father ?Z should not have a match, even if ?Z is non-distinguished. Disregarding universal property assertions that follow from rule (16), anonymous individuals can only be related to their parent individual (represented by ?X in rule (2)) or to their children (which have the anonymous element as their parent). Therefore, to check if a triple pattern ?X $p$ ?Y can match, we may need to know if ?X is the parent of ?Y. We capture this with auxiliary variables ?Pxy which we bind to one of two possible values (interpreted as *true* and *false*).

We thus introduce variables ?Pxy for every pair of CQ variables $x$ and $y$ where $y$ is non-distinguished. This completely specifies the parenthood of the matches. Together with the generating OWL restriction represented by ?Mx, this gives us enough information to verify property assertions. To find all matches of a CQ, one has to allow for the possibility that several query variables represent the same element of the chase. To capture this, we introduce variables ?Exy that tell us if the values of $x$ and $y$ are equal; again we use two possible values to represent *true* and *false*. Additional conditions in our query will ensure that there are no cycles in the parenthood relation, and that equal values are indeed equal. Many of these can be encoded in propositional logic, as explained next.

### 5.2 Expressing Propositional Logic in SPARQL 1.1

Our intuitive explanation above uses "Boolean" variables like ?Pxy and ?Exy, which can have one of two values. Moreover, the bindings of these variables should obey further constraints. For example, if $x$ is the parent of $y$ and $y$ is identified with $z$, then $x$ is the parent of $z$. This corresponds to a propositional logic implication ?Pxz $\wedge$ ?Eyz $\rightarrow$ ?Pxz.

We express this using the VALUES feature of SPARQL 1.1, which allows us to assign a fixed set of bindings to a list of variables. For example, the pattern VALUES (?Pxy){(<http://example.org/true>)(<http://example.org/false>)} has exactly two solutions, binding ?Pxy to one of the given URIs. The URIs used here are irrelevant, and it does not even matter if they occur in the data; we thus use the abbreviations T and F to denote two distinct URIs that we use to represent Boolean values. Propositional logic formulae can now be represented by encoding their truth table using VALUES. For example, the implication ?Pxz $\wedge$ ?Eyz $\rightarrow$ ?Pxz can be expressed as:

VALUES (?Pxy ?Eyz ?Pxz){(F F F)(T F F)(F T F)(F F T)(T F T)(F T T)(T T T)}. (31)

We denote this pattern as $[\![\mathsf{?Pxz} \wedge \mathsf{?Eyz} \rightarrow \mathsf{?Pxz}]\!]$, and similarly for any other propositional logic formula over SPARQL variables. The solutions to (31) are exactly the truth assignments under which the implication holds. In particular, every solution requires each of the three variables to be bound to T or F (and thus to never be undefined).

### 5.3 A Schema-Agnostic Rewriting for Conjunctive Queries

We now specify the complete rewriting of CQs in SPARQL 1.1, which consists of rewritings for the individual triple patterns and several additional patterns to ensure that the bindings of all (auxiliary) variables are as intended. Consider a CQ $q$ with variables $\mathsf{Var}(q)$, partitioned into the set $\mathsf{Var}_d(q)$ of distinguished variables and $\mathsf{Var}_n(q)$ of non-distinguished variables. Our encoding uses the following sets of SPARQL variables:

- for every $x \in \mathsf{Var}(q)$, a variable ?Mx (encoding the "match for $x$").

In addition, we use the following propositional SPARQL variables:

- for every $x \in \mathsf{Var}(q)$, a variable ?Nx ("$x$ is a named individual").
- for every pair $x, y \in \mathsf{Var}(q)$, a variable ?Exy ("$x$ is equal to $y$");
- for every pair $x \in \mathsf{Var}(q)$ and $y \in \mathsf{Var}_n(q)$, a variable ?Pxy ("$x$ is the parent of $y$");
- for every pair $x, y \in \mathsf{Var}_n(q)$, a variable ?Axy ("$x$ is an ancestor of $y$");

The variables ?Axy are used to encode the transitive closure over the parent relations on non-distinguished variables; this is necessary to preclude cyclic ancestries. We use PropConstraints($q$) to denote the SPARQL encoding of all of the following implications (for every possible combination of the above variables, if no other condition is given):

$$\text{for } x \in \mathsf{Var}_d(q)\text{: } \mathsf{T} \rightarrow \mathsf{?Nx}$$

$$\begin{array}{rcr} \mathsf{?Exy} \rightarrow \mathsf{?Eyx} & \mathsf{?Exy} \wedge \mathsf{?Nx} \rightarrow \mathsf{?Ny} & \mathsf{?Pxy} \rightarrow \mathsf{?Axy} \\ \mathsf{?Exy} \wedge \mathsf{?Eyz} \rightarrow \mathsf{?Exz} & \mathsf{?Exy} \wedge \mathsf{?Pxz} \rightarrow \mathsf{?Pyz} & \mathsf{?Axy} \wedge \mathsf{?Ayz} \rightarrow \mathsf{?Axz} \\ \mathsf{?Pxz} \wedge \mathsf{?Pyz} \rightarrow \mathsf{?Exy} & \mathsf{?Exy} \wedge \mathsf{?Pzx} \rightarrow \mathsf{?Pzy} & \mathsf{?Axx} \rightarrow \mathsf{F} \end{array}$$

The previous conditions do not ensure yet that the bindings for ?Mx and ?My are the same whenever ?Exy is true. This cannot be encoded using VALUES. Instead, we define EqualityFilter($q$) to be the condition of the following filter conditions:

$$\mathsf{FILTER(?Exy = F \,||\, ?Mx = ?My)} \qquad x, y \in \mathsf{Var}(q)$$

We can now define the rewriting of the actual query conditions. For readability, we use $[\![\mathsf{?V} := u]\!]$ to abbreviate VALUES (?V){($u$)}. The triple pattern $x$ rdf:type $c$ is rewritten into the following pattern, denoted Rewrite($x$ rdf:type $c$):

```
      {[[?Nx := T]] . [[?Mx rdf:type c]]}
UNION {univClass[c]}
UNION {[[?Nx := F]] . ?E subClassOf c
       {{?Mx owl:someValuesFrom ?E} UNION
        {?Mx (owl:onProperty / subPropertyOf / rdfs:range) ?E} UNION
        {?Mx (owl:onProperty / subInvPropertyOf / (^owl:onProperty | rdfs:domain)) ?E}}
```

A triple pattern $x\ p\ y$ is rewritten into the following pattern, denoted REWRITE($x\ p\ y$):

```
        {⟦?Nx := T⟧ . ⟦?Ny := T⟧ . ⟦?Mx p ?My⟧}
UNION {UNIVPROPERTY[p]}
UNION {⟦?Ny := F⟧ . ⟦?Pxy := T⟧ . ?My (owl:onProperty / SUBPROPERTYOF) p
       {REWRITE(x rdf:type ?My)}
UNION {⟦?Nx := F⟧ . ⟦?Pyx := T⟧ . ?Mx (owl:onProperty / SUBINVPROPERTYOF) p
       {REWRITE(y rdf:type ?Mx)}
```

Note that the parenthood relationship ?Pyx is only relevant for checking certain triple patterns. In each of these cases, we verify that the parent element is really capable of creating the required child. This ensures that all assumed parenthoods that are relevant to prove the query are really derived. In addition, we still need to check that all anonymous elements are really derived (from some original ancestor element in the ontology).

*Example 2.* Consider an OWL ontology with the assertion :peter rdf:type :Person and the axiom :Person rdfs:subClassOf [owl:onProperty :mother; owl:someValuesFrom :Woman]. Then the query {?X rdf:type :Woman} with ?X non-distinguished has a match. However, if we remove the triple :peter rdf:type :Person, then the query does not have a match. In contrast, our pattern REWRITE($x$ rdf:type :Mother) could match in either case.

To fix this, we introduce, for every non-distinguished variable $x$, an additional pattern MATCHEXISTS($x$) that verifies that an element of the assumed type is actually derived. This pattern also ensures that named individuals are always bound to individuals. Anonymous individuals may be inferred from our assumption that the domain is not empty, or they must be derived from a named individual, which we represent by a bnode:

```
        {⟦?Nx := T⟧ . INDIVIDUAL(?Mx)}
UNION {⟦?Nx := F⟧ . ⟦?Mx := owl:Thing⟧}
UNION {⟦?Nx := F⟧ . ⟦_:b rdf:type ?E⟧ . ?E (rdfs:subClassOf | INTLISTMEMBER |
       (owl:onProperty / (INV | SPOEQP)* / (^owl:onProperty | rdfs:domain | rdfs:range)) |
       ^owl:equivalentClass | owl:equivalentClass | owl:someValuesFrom)* ?Mx}
```

We do not need to check that this derivation agrees with the guessed parenthood relations, since the check is only relevant for the elements that do not have a parent represented by a query variable. Also note that the pattern INDIVIDUAL(?Mx), which applies to both named and anonymous individuals in the chase, will only match existing (i.e., named) individuals when applied to the ontology.

**Definition 6.** *The rewriting* REWRITE($q$) *a CQ $q$ with distinguished variables $x_1, \ldots, x_n$ is the following SPARQL 1.1 query:*

```
SELECT DISTINCT ?Mx1, ..., ?Mxn WHERE {
  PROPCONSTRAINTS(q)
  REWRITE(x rdf:type c) for each condition x rdf:type c in q
  REWRITE(x p y) for each condition x p y in q
  MATCHEXISTS(x) for each variable x in q
  EQUALITYFILTER(q)
}
```

**Theorem 4.** *The answers of a conjunctive query $q$ over an OWL QL ontology $G$ are exactly the answers of the SPARQL 1.1 query* $\textsc{Rewrite}(q)$ *over $G$ under simple entailment.*

To prove the previous theorem, we establish a number of correctness properties for the individual parts of the query. The following result mainly states that the auxiliary Boolean SPARQL variables define a congruence relation $E$ and an acyclic relation $P$.

**Lemma 5.** *If $\mu$ is a solution mapping for the pattern* $\textsc{PropConstraints}(q)$*, then the following hold:*

*(a) $\mu$ maps every variable of the form* ?Nx*,* ?Pxy*,* ?Exy*, and* ?Axy *to either* T *or* F*.*
*(b) The set $N \subseteq \mathsf{Var}(q)$ defined as $x \in N$ iff $\mu(\mathsf{?Nx}) = \mathsf{T}$ contains all distinguished variables of $q$.*
*(c) The relation $E$ on $\mathsf{Var}(q)$ defined as $\langle x, y\rangle \in E$ iff $\mu(\mathsf{?Exy}) = \mathsf{T}$ is an equivalence relation.*
*(d) The relation $A$ on $\mathsf{Var}(q)$ defined as $\langle x, y\rangle \in A$ iff $\mu(\mathsf{?Axy}) = \mathsf{T}$ is an strict order (i.e., transitive and irreflexive).*
*(e) $A$ contains the binary relation $P$ on $\mathsf{Var}(q)$ defined as $\langle x, y\rangle \in P$ iff $\mu(\mathsf{?Pxy}) = \mathsf{T}$.*
*(f) $P$ is an acyclic, irreflexive relation.*
*(g) The following hold:*

$$x \in N \wedge \langle x, y\rangle \in E \rightarrow y \in N$$
$$\langle x, z\rangle \in P \wedge \langle x, y\rangle \in E \rightarrow \langle y, z\rangle \in P$$
$$\langle z, x\rangle \in P \wedge \langle x, y\rangle \in E \rightarrow \langle z, y\rangle \in P$$
$$\langle x, z\rangle \in P \wedge \langle y, z\rangle \in P \rightarrow \langle x, y\rangle \in E$$

*(h) If $\mu$ satisfies the filters* $\textsc{EqualityFilter}(q)$*, then $\langle x, y\rangle \in E$ implies $\mu(\mathsf{?Mx}) = \mu(\mathsf{?My})$.*

*Proof.* (a) is clear from the fact that every variable occurs in one of the implications. (b) is immediate from the first implication.

For (c), symmetry and transitivity are immediate from the implications; reflexivity follows from the implication $\mathsf{?Pxz} \wedge \mathsf{?Pxz} \rightarrow \mathsf{?Exx}$ and (a).

Item (d) and (e) are immediate from the implications. From this, (f) follows. Item (g) is again immediate from the implications, and (h) is immediate from the definition of $\textsc{EqualityFilter}(q)$. □

For the remainder of the proof, we relate the matches found in the rewritten query to the structure of the chase, to construct a match of the original CQ. For this we assume that one particular sequence of rule applications has been fixed for deriving the chase, so that we can argue by induction over this derivation. This is what we mean when referring to *the chase* in the rest of this section.

Inspired by the syntax of description logics, we use $\exists p.c$ to refer to any blank node _:b with triples _:b owl:onProperty $p$; owl:someValuesFrom $c$ in the ontology.

**Lemma 6.** *The following are equivalent:*

*(i) There is a solution $\mu$ for* $\textsc{MatchExists}(x)$ *such that $\mu(\mathsf{?Nx}) = \mathsf{F}$ and $\mu(\mathsf{?Mx}) = c$.*
*(ii) $c$ is an OWL class that is not empty in the chase.*

*In particular if* $\mu(?Nx) = F$ *and* $\mu(?Mx)$ *is a bnode of the form* $\exists p.c$*, then the chase contains an element in this class* $\exists p.c$*, and rule* (2) *is applicable.*

*Proof.* We first show that (i) implies (ii). If $\mu(?Nx) = F$ then $\mu$ is a solution for the second or third graph pattern in MatchExists($x$). Solutions of the second pattern are such that $\mu(?Mx) =$ owl:Thing, and the claim clearly holds. For solutions of the third pattern, we show the claim by induction over the number $n$ of iterations of the property path expression used for this solution. If $n = 0$ then $\mu(?Mx) = \mu(?E)$ and $\mu$ is a solution for $[\![$_:b rdf:type ?Mx$]\!]$. By Theorem 2, the ontology entails _:b rdf:type $\mu(?Mx)$, which implies the claim.

For the case $n > 0$, let $d$ be the element reached after $n = 1$ iterations. By the hypothesis, $d$ is non-empty class in the chase. It is easy to see that each of the optional parts in the property path of the query pattern propagate non-emptiness.

We now show that (ii) implies (i). This is done by induction over the construction of the chase. If $c$ is non-empty, then there is a rule application that derives a triple $x$ rdf:type $c$ in the $n$th step of the chase construction. Assume by way of induction that the claim holds for all triples of the form $y$ rdf:type $d$ that are derived in the first $n \geq 0$ steps of the chase. For $n = 0$ this is certainly true since $[\![$_:b rdf:type $c]\!]$ matches any triple of the form $x$ rdf:type $c$ that is given in the ontology.

For $n > 0$, $x$ rdf:type $c$ might be derived by an application of one of rules (2)–(9) and (15). Note that (10)–(14) cannot produce rdf:type triples in a well-formed OWL QL ontology. Moreover, the rule (15) only produces triples of the form $x$ rdf:type owl:Thing; the claim is immediate for these, since already $[\![$_:b rdf:type owl:Thing$]\!]$ matches any ontology (this is immediate from the definitions, but also a consequence of Theorem 2).

Of the remaining rules that can produce rdf:type triples, rules (2)–(6) are infer a triple $x$ rdf:type $c$ from a premise of the form $y$ rdf:type $d$ such that there is a direct path of triples from $d$ to $x$. This path is described by owl:someValuesFrom, intListMember, rdfs:subClassOf, owl:equivalentClass, and Inv owl:equivalentClass, respectively. Since $d$ is non-empty, by the induction hypothesis we find a solution $\mu'$ for MatchExists($x$) such that $\mu'(?Nx) = F$ and $\mu'(?Mx) = d$. The required $\mu$ is the solution that agrees with $\mu'$ on all variables other than possibly ?Mx, which is mapped to $c$ now. This yields a solution since the path from $\mu'(?E)$ to $\mu'(?Mx)$ can be extended to $\mu(?Mx)$ by any of the above properties, all of which appear in the property path pattern.

The argument for the remaining rules (7)–(9) is similar, but requires us to also consider the derivation of property assertion triples $y\ p\ z$ in the chase. These triples can be derived by rules (2), (10)–(14), and (16). Since rules (7)–(9) only derive rdf:type statements about the individuals that occur in a triple in the premise, and since rules (10)–(14) preserve these individuals, every rdf:type triples that follows from a property assertion in the ontology is the type of an individual in the ontology (namely one of the individuals used in the property assertion). Thus, by Theorem 2, such rdf:type triples match $[\![$_:b rdf:type $c]\!]$. This also covers all rdf:type triples that are derived from the conclusion of rule (16).

We can therefore restrict attention to rdf:type triples derived by applying rules (7)–(9) to a triple obtained from rule (2), with possible intermediate applications of rules (10)–(14). It is easy to verify that this is exactly what the property expres-

sion (owl:onProperty / (INV | SPOEQP)* / (^owl:onProperty | rdfs:domain | rdfs:range)) detects. The application of the induction hypothesis to show the claim is as before.

The rest of the claim is immediate. □

We have thus shown that all hypothesized anonymous individuals required for a CQ query match do really exist in the chase.

**Lemma 7.** *The atom x* rdf:type *c has a match in the chase iff one of the following holds (for some p and d):*

*(i) The chase contains the triple e* rdf:type *c where e is a named individual.*
*(ii) The chase contains the triple* _:b rdf:type *c where* _:b *is a bnode that was generated by applying rule* (1).
*(iii) The chase contains the triple* _:b rdf:type *c where* _:b *is a bnode that was generated by applying rule* (2) *to* $\exists p.d$.

*These conditions are equivalent to the following three cases, respectively:*

*(i') There is a solution* $\mu$ *for* REWRITE(*x* rdf:type *c*) *and* MATCHEXISTS(*x*) *such that* $\mu$(?Nx) = T *and* $\mu$(?Mx) = *e*.
*(ii') There is a solution* $\mu$ *for* REWRITE(*x* rdf:type *c*) *and* MATCHEXISTS(*x*) *such that* $\mu$(?Nx) = F *and* $\mu$(?Mx) = owl:Thing.
*(iii') There is a solution* $\mu$ *for* REWRITE(*x* rdf:type *c*) *and* MATCHEXISTS(*x*) *such that* $\mu$(?Nx) = F *and* $\mu$(?Mx) = $\exists p.d$.

*Moreover, every solution* $\mu$ *for* REWRITE(*x* rdf:type *c*) *and* MATCHEXISTS(*x*) *satisfies one of the conditions (i')–(iii').*

*Proof.* The first part of the claim is immediate from Theorem 1 together with the observation that every element of the chase is either named (i), introduced in rule (1) (ii), or introduced in rule (2) (iii).

It remains to show the three claimed equivalences. The equivalence of (i) and (i') is an immediate consequence of Theorem 2 and the fact that ⟦?Mx rdf:type *c*⟧ is used in REWRITE(*x* rdf:type *c*).

(ii) ⇒ (ii'). By Lemma 6, MATCHEXISTS(*x*) has a solution mapping $\mu$ with $\mu$(?Nx) = F and $\mu$(?Mx) = owl:Thing. It remains to show that this mapping is a solution for REWRITE(*x* rdf:type *c*). This is obtained by a simple solution over the steps of the chase. Indeed, it has been argued before that the pattern UNIVCLASS[*c*] matches exactly if *c* is a superclass of owl:Thing, which in turn holds exactly if REWRITE(_ : *b* rdf:type *c*) holds for the bnode _ : *b* generated in rule (1).

(ii') ⇒ (ii). Conversely, if $\mu$(?Mx) = owl:Thing is a solution for REWRITE(*x* rdf:type *c*), then the pattern UNIVCLASS[*c*] must match. In all other cases, $\mu$(?Mx) = owl:Thing cannot be a match on a well-formed OWL DL ontology. Then it is again easy to show that this implies that *c* is a superclass of owl:Thing, which in turn implies (ii).

(iii) ⇒ (iii'). By Lemma 6, MATCHEXISTS(*x*) has a solution mapping $\mu$ with $\mu$(?Nx) = F and $\mu$(?Mx) = $\exists p.d$. It remains to show that $\mu$ is a solution for REWRITE(*x* rdf:type *c*). This, again, is done by induction over the chase steps that have been used to derive _:b rdf:type *c*. This argument follows the same structure as our earlier proofs and we omit

the details. It is important to note that the third pattern in the union Rewrite($x$ rdf:type $c$) does not cover all possible ways of deriving _:b rdf:type $c$; the remaining cases are covered by the second pattern univClass[$c$]. The argumentation is very similar to the case of $[\![x \text{ rdf:type } c]\!]$ in Theorem 2.

(iii') ⇒ (iii). In case (iii'), $\mu$ matches either the second or third part of the union Rewrite($x$ rdf:type $c$). In either case, it is easy to see that the possible patters really imply that _:b rdf:type $c$ for the blank node generated by $\exists p.d$. We omit the details.

The final claim that the cases (i')–(iii') cover all possible ways of finding a solution for the given patterns is clear from the definition. In fact, already pattern MatchExists($x$) alone ensures that solutions have one of the required forms. □

In summary, the previous lemma shows that $x$ rdf:type $c$ has a match in the chase iff Rewrite($x$ rdf:type $c$) and MatchExists($x$) have a corresponding solution mapping in the ontology. This is already an important ingredient for the overall proof. It remains to consider the case of property assertions.

**Lemma 8.** *The atom $x$ $p$ $y$ has a match in the chase iff one of the following holds (for some $s$ and $c$):*

*(i) The chase contains the triple $e$ $p$ $f$ where $e$ and $f$ are named individuals.*
*(ii) The chase contains the triple _:b $p$ _:b where _:b is a bnode that was generated by applying rule* (1)*.*
*(iii) The chase contains the triple $e$ $p$ _:b where _:b is a bnode that was generated by applying rule* (2) *to $e$* rdf:type $\exists s.c$*.*
*(iv) The chase contains the triple _:b $p$ $f$ where _:b is a bnode that was generated by applying rule* (2) *to $f$* rdf:type $\exists s.c$*.*

*These conditions are equivalent to the following cases, respectively:*

*(i') There is a solution $\mu$ for* Rewrite($x$ $p$ $y$)*,* MatchExists($x$)*, and* MatchExists($y$) *such that $\mu$*(?Nx) = T*, $\mu$*(?Mx) = *e, and $\mu$*(?My) = *$f$*.
*(ii') There is a solution $\mu$ for* Rewrite($x$ $p$ $y$)*,* MatchExists($x$)*, and* MatchExists($y$) *such that $\mu$*(?Nx) = $\mu$(?Ny) = F *and $\mu$*(?Mx) = $\mu$(?My) = owl:Thing.
*(iii') There is a solution $\mu$ for* Rewrite($x$ $p$ $y$)*,* MatchExists($x$)*, and* MatchExists($y$) *such that $\mu$*(?Ny) = F*, $\mu$*(?My) = $\exists s.c$*, and $\mu$ is a solution of* Rewrite($x$ rdf:type ?My).
*(iv') There is a solution $\mu$ for* Rewrite($x$ $p$ $y$)*,* MatchExists($x$)*, and* MatchExists($y$) *such that $\mu$*(?Nx) = F*, $\mu$*(?Mx) = $\exists s.c$*, and $\mu$ is a solution of* Rewrite($y$ rdf:type ?Mx).

*Moreover, every solution $\mu$ for* Rewrite($x$ $p$ $y$)*,* MatchExists($x$)*, and* MatchExists($y$) *satisfies one of the conditions (i')–(iv').*

*Proof.* To show the first part of the claim, consider the chase rules that can derive property assertion triples of the form $v$ $s$ $w$ where $s \neq$rdf:type. These rules are (2), (10)–(14), and (16). Among these, the rules (12)–(14) require a premise $v$ $s'$ $w$ with the same subject and object; rules (10) and (11) require a premise $w$ $s'$ $v$ with subject and object swapped. New relationships between hitherto unrelated elements can therefore only be

established by rules (2) and (16). All other relationships must be present in the ontology initially.

Thus, a pair of elements $v$ and $w$ can only be related by a property in the chase if one of the following holds: (i) the ontology already contained a triple $v\ s\ w$ or $w\ s\ v$; (ii) the relationship of $v$ and $w$ is universal, i.e., based on rule (16); (iii) a relationship from $v$ to $w$ was created by rule (2); (iv) a relationship from $w$ to $v$ was created by rule (2). It is easy to see that these cases correspond to the cases in the claim. In particular, for (ii) we can restrict to triples that follow from the universal property alone, without requiring any other information – these triples are derived for all pairs of individuals, and thus also hold between the bnode generated in rule (1) and itself.

For the second part of the claim, we need to show the mutual correspondence of the four cases. The proof for this is very similar to the proof of Lemma 7. We use Lemma 7 to relate the matches of the subpatterns of the form Rewrite($x$ rdf:type ?My) to the semantic conditions stated for the chase. We omit the details of the by now standard induction arguments.

The final claim that the cases (i')–(iv') cover all possible ways of finding a solution for the given patterns is again easy to verify from the definition of the patterns. Indeed, the four cases correspond to the four components of the union Rewrite($x\ p\ y$). As in Lemma 7, MatchExists($x$), and MatchExists($y$) ensure that bindings ?M_ are always defined and are restricted to named individuals, owl:Thing, and expressions $\exists s.c$. □

To complete the proof of Theorem 4, we need to bring these results together.

*Proof (of Theorem 4).* First, for soundness, assume that Rewrite($q$) has a solution $\mu$ over $G$ under simple entailment. By Lemma 6, whenever a variable ?Mx is bound to a class $\exists s.c$, this class must be non-empty, i.e., the anonymous individual represented by ?Mx exists in the chase. By Lemmas 7 and 8, all individual query patterns are supported by matches in the chase. Since our query pattern is a join over all of the triple pattern rewritings, these matches are consistent with each other. By Lemma 5, the parenthood relation is acyclic (under the assumed equality); this ensures that the triples of the chase that exist according to Lemma 8 do also form a chain in the chase. Indeed, it is clear from the chase rules that every anonymous individual that was generated by a class $\exists s.p$ has the same tree of children elements in the chase. Therefore, as long as the parenthood relation is acyclic, it is always possible to find a sequence of triples in the chase that matches the assumed sequence of anonymous individuals in the query match. Finally, Lemma 5 also asserts that the query match respects the assumed equality ((c), (g), (h)), and that distinguished variables are mapped to named individuals only (b). These properties ensure that CQ $q$ has a match to the chase of $G$ in which distinguished variables are mapped to named individuals.

For correctness, assume that the CQ $q$ has a match $\nu$ over $G$ under OWL DL entailment. This implies that $q$ has a match to the chase of $G$ where distinguished variables are mapped to named individuals. We define a solution $\mu$ for Rewrite($q$) in the obvious way:

- $\mu$(?Nx) = T if $\nu(x)$ is a named individual, and $\mu$(?Nx) = F otherwise;
- $\mu$(?Mx) = $\nu(x)$ if $\nu(x)$ is a named individual, $\mu$(?Mx) = $\exists s.c$ if $\nu(x)$ is an anonymous individual generated by applying rule (2) to $\exists s.c$, and $\mu$(?Mx) = owl:Thing otherwise;

- $\mu(?Exy) = \mathsf{T}$ if $\nu(x) = \nu(y)$, and $\mu(?Exy) = \mathsf{F}$ otherwise;
- $\mu(?Pxy) = \mathsf{T}$ if $\nu(y)$ was generated by chase rule (2) from individual $\nu(x)$, and $\mu(?Pxy) = \mathsf{F}$ otherwise;
- $\mu(?Axy) = \mathsf{T}$ if $\langle x, y \rangle$ is in the transitive closure of relation $P$ as defined in Lemma 5, and $\mu(?Axy) = \mathsf{F}$ otherwise.

It is easy to check that this assignment satisfies all patterns in Rewrite($q$). The non-trivial cases follow from Lemmas 6, 7, and 8. □

## 6 Limits of Schema-Agnostic Query Rewriting

We have seen that schema-agnostic query rewriting works for (almost) all of OWL QL, so it is natural to ask how far this approach can be extended. In this section, we outline the natural limits of SPARQL 1.1 as a query rewriting language, point out extensions to overcome these limits.

In Section 3, we excluded owl:SymmetricProperty from our considerations. Indeed, schema-agnostic SPARQL 1.1 queries cannot support this feature. This might be surprising, given that one can write $p$ rdf:type owl:SymmetricProperty as $p$ rdfs:subPropertyOf [owl:inverseOf $p$]. To see why this problem occurs, consider the following ontology:

```
:c rdfs:subClassOf [rdf:type owl:Restriction; owl:onProperty :p; owl:someValuesFrom owl:Thing] .
[] rdf:type owl:Restriction; owl:onProperty [owl:inverseOf :p]; owl:someValuesFrom owl:Thing;
   rdfs:subClassOf :d .
:p rdf:type owl:SymmetricProperty .
```

This ontology states: every :c has an outgoing :p property; everything with an incoming :p property is a :d; and :p is symmetric. Clearly, this implies that :c is a subclass of :d. We call this ontology $G$(:c, :p, :d). Now assume that we have a chain of such ontologies $G_n := G(:c1, :p1, :c2), \ldots, G(:cn, :pn, :d)$. Clearly, $G_n$ implies that :c1 rdfs:subClassOf :d, but there is no SPARQL 1.1 graph pattern with property paths that recognizes this triple structure in an ontology. The intuitive explanation is that $G_n$ contains a property path of length $4n$ that matches the following expression:

$$(\text{rdfs:subClassOf} \,/\, \text{owl:onProperty} \,/\, \hat{}\text{owl:onProperty} \,/\, \text{rdfs:subClassOf})^*$$

A SPARQL query that matches $G_n$ for any $n$ needs to use such a path expression; no other feature in SPARQL 1.1 can navigate arbitrary distances. However, it is impossible to verify that each :p$i$ on this path is of type owl:SymmetricProperty. To give a formal proof of this intuitive argument, we need to study the expressive power of SPARQL 1.1. It suffices to look at queries without selected variables (ASK queries in SPARQL). Any such query characterizes an infinite set of RDF graphs for which it has a match. We want to show that no such query can characterize the set of all OWL QL ontologies (with symmetry) that entail :c1 rdfs:subClassOf :d. To simplify the argument, we can restrict attention to a subset of "typical" RDF graphs that are matched.

**Definition 7.** *A graph $G$ is* more general *than graph $G'$ if there is a node-renaming function $r$ with $G(r) \subseteq G'$ (i.e., $r$ is a graph homomorphisms). A set of RDF graphs $\mathscr{G}$ is a* covering *of an* ASK *query $Q$ if, for every graph $G$ matched by $Q$, there is a graph $G' \in \mathscr{G}$ that is more general than $G$.*

Intuitively speaking, more general graphs match fewer queries. Coverings are simple for SPARQL 1.0 patterns, but tend to be infinite as soon as property paths are added. For example, for any $\ell \geq 0$, the query ?X :p* ?X is covered by the set of graphs that describe :p-loops of length greater than $\ell$; each of these coverings is infinite and none of them is minimal. The example illustrates that coverings nevertheless can reveal interesting structural properties of query matches. We use this to prove the next proposition.

**Proposition 3.** *There is no SPARQL 1.1* ASK *query that matches (under simple entailment) all RDF graphs of OWL QL ontologies that entail* :c1 rdfs:subClassOf :d.

*Proof.* We define the *degree* of a node in an RDF graph to be the number of its distinct neighbors in the graph. Every property path expression has a covering of linear graphs (where each node has degree $\leq 2$). Thus, every SPARQL 1.1 query pattern with $\ell$ variables has a covering of graphs where at most $\ell$ nodes have a degree $> 2$.

On the other hand, each graph pattern $G_n$ as defined above has $n-1$ nodes $(p_i)_{i=1}^{n-1}$ of degree 3. Suppose for a contradiction that there is a SPARQL 1.1 query $Q$ as in the claim, which has $\ell$ variables. Then $Q$ must have a covering that does not include $G_m$. Since $Q$ matches $G_m$, there must be a graph $G'$ in the covering of $Q$ that is strictly more general than $G_m$ and which $Q$ also matches. However, it is not hard to see that any graph that is strictly more general that $G_m$ fails to entail :c1 rdfs:subClassOf :d – a contradiction. □

While this limitation is hardly more than a syntactic inconvenience, one might ask if there are query languages that can deal with this type of encoding. Indeed, one possible approach is nSPARQL, which has been proposed as an extension of SPARQL 1.0 with a form of path expressions that can test for the presence of certain side branches in property paths [15]. Similar test expressions have been considered in OBDA recently [2]. These query languages can handle the RDF encoding of symmetric properties.

Besides such "syntactic" limitations, schema-agnostic query rewriting is also restricted by complexity theoretic arguments. Simply put, the reasoning task solved in this way can not be harder (computationally speaking) than the data complexity of the underlying query language. The data complexity of the subset of SPARQL used in this paper is NLogSpace: SPARQL 1.1 patters are a variant of positive regular path queries [8], which have NLogSpace data complexity (by translation to linear Datalog [10]); inline data (VALUES) does not affect data complexity; and final filtering with equality checks can clearly be implemented in logarithmic space.

Thus, we cannot expect to obtain SPARQL 1.1 queries for any ontology language of (combined) complexity above NLogSpace. In particular, P is widely assumed to be strictly harder than NLogSpace (though no proof has been given yet), which lets us exclude most extensions of OWL QL as well as other lightweight OWL profiles:

**Theorem 5.** *If* P *is strictly harder than* NLogSpace*, then reasoning for the following ontology languages cannot be expressed in SPARQL 1.1 using property paths,* UNION*,* VALUES *and (in)equality filters:*

- *any subset of OWL with* owl:intersectionOf *in subclass positions, especially OWL EL and OWL RL;*
- *any subset of OWL with unrestricted* owl:someValuesFrom *in subclasses and superclasses (not limited to* owl:Thing*);*

– *the extension of OWL QL with regular property chain axioms.*

*Proof.* For the case of unrestricted owl:intersectionOf, the result is immediate from the known P-hardness of propositional Horn logic [6].

For the case of unrestricted owl:someValuesFrom, we can encode propositional Horn logic as follows. Consider a propositional implication $r : p_1 \wedge p_2 \rightarrow p_3$. We introduce class names $C_1$, $C_2$, and $C_3$, a class name $T$ (for "true"), and property names $r_1$ and $r_2$. The rule is now expressed by the axioms:

$C_3$ rdfs:subClassOf [rdf:type owl:Restriction; owl:onProperty $r_1$; owl:someValuesFrom $C_1$] .
[] rdf:type owl:Restriction; owl:onProperty $r_1$; owl:someValuesFrom $T$;
rdfs:subClassOf [rdf:type owl:Restriction; owl:onProperty $r_2$; owl:someValuesFrom $C_2$] .
[] rdf:type owl:Restriction; owl:onProperty $r_2$; owl:someValuesFrom $T$;
rdfs:subClassOf $T$ .

A fact $p_1$ is expressed by the axiom $C_1$ rdfs:subClassOf $T$. We can express any set of propositional facts and implications in this way, using the same class names $C_i$ for every use of $p_i$, but different properties $r_1$ and $r_2$ for each rule. The resulting ontology entails $C_i$ rdfs:subClassOf $T$ if and only if the original Horn logic theory entails $p_i$.

For the final case of OWL QL with property chain axioms is similar to the second case. However, instead of $T$, we now use an outgoing property $t$ to signify that a proposition is true:

$T$ rdfs:subClassOf [rdf:type owl:Restriction; owl:onProperty $t$; owl:someValuesFrom owl:Thing] .

Moreover, in addition to the properties $r_1$ and $r_2$, we introduce properties $r'_1$ and $r'_2$ for each rule. The axioms for expressing a rule $r : p_1 \wedge p_2 \rightarrow p_3$ then are:

$C_3$ rdfs:subClassOf [rdf:type owl:Restriction; owl:onProperty $r_1$; owl:someValuesFrom $C_1$] .
$r'_1$ owl:propertyChainAxiom ($r_1$ $t$) .
[] rdf:type owl:Restriction; owl:onProperty $r'_1$; owl:someValuesFrom owl:Thing;
rdfs:subClassOf [rdf:type owl:Restriction; owl:onProperty $r_2$; owl:someValuesFrom $C_2$] .
$r'_2$ owl:propertyChainAxiom ($r_2$ $t$) .
[] rdf:type owl:Restriction; owl:onProperty $r'_2$; owl:someValuesFrom owl:Thing;
rdfs:subClassOf $T$ .

Facts can be represented as before. It is easy to see that this captures propositional Horn logic. The property chain axioms are regular, since the $r'_i$ properties, which are implied by the axioms, never appear as a subproperty in any axiom. □

These complexity-theoretic limitations can only be overcome by using a more complex query language. Many query languages with P-complete data complexity can be found in the Datalog family of languages, which are supported by RDF databases like OWLIM and Oracle 11g that include rule engines.

## 7 Conclusions and Outlook

To the best of our knowledge, our work is the first to present a query rewriting approach for ontology-based data access in OWL QL that is completely independent of the ontology. The underlying paradigm of schema-agnostic query rewriting appears to be a promising approach that can be applied in many other settings. Indeed, two previous works, nSPARQL [15] and PSPARQL [1], independently proposed query-based mechanisms for reasoning in RDFS. While these works have not considered SPARQL 1.1, OWL QL, or arbitrary conjunctive queries, they still share important underlying ideas. We think that a common name is very useful to denote this approach to query rewriting.

In this paper, we have focused on laying the foundations for this new reasoning procedure. An important next step is to study its practical implementation and optimization. Considering the size of some of the queries we obtain, one would expect them to be challenging for RDF stores. We have started to implement our approach in a prototype system [3], and initial experiments confirm this expectation. Encouragingly, however, executing rewritten queries seems to be feasible, even in the raw, unoptimized form they have in this paper. Future work will be concerned with developing further optimizations that can be used in practical evaluations.

Indeed, while the queries we obtain might be challenging for current RDF stores, large parts of the queries are fixed and can be optimized for. Our work thus reduces the problem of adding OWL QL reasoning support to RDF stores to a query optimization problem. This can also guide future work in stores, such as OWLIM, which implement reasoning with inference rules: rather than trying to materialize (part of) an infinite OWL QL chase [4], they could materialize (sub)query results to obtain a sound and complete procedure. This provides completely new perspectives on the use of OWL QL in areas that have hitherto been reserved to OWL RL and RDFS.

Finally, our work also points into several interesting directions for foundational research, as mentioned in Section 6. Promising approaches include development of schema-agnostic rewriting procedures for languages like OWL EL that cannot be captured by SPARQL 1.1, and the development of query languages that suit this task [18].

**Acknowledgements** This work has been funded by the Vienna Science and Technology Fund (WWTF, project ICT12-015), and by the DFG in project DIAMOND (Emmy Noether grant KR 4381/1-1).

## References

1. Alkhateeb, F.: Querying RDF(S) with Regular Expressions. Ph.D. thesis, Université Joseph Fourier – Grenoble 1 (2008)
2. Bienvenu, M., Calvanese, D., Ortiz, M., Šimkus, M.: Nested regular path queries in description logics. CoRR abs/1402.7122 (2014)
3. Bischof, S., Krötzsch, M., Polleres, A., Rudolph, S.: Schema-agnostic query rewriting in SPARQL 1.1: Technical report. http://stefanbischof.at/publications/iswc14/ (2014)
4. Bishop, B., Bojanov, S.: Implementing OWL 2 RL and OWL 2 QL rule-sets for OWLIM. In: Dumontier, M., Courtot, M. (eds.) Proc. 8th Int. Workshop on OWL: Experiences and Directions (OWLED'11). CEUR Workshop Proceedings, vol. 796. CEUR-WS.org (2011)

5. Calvanese, D., Giacomo, G.D., Lembo, D., Lenzerini, M., Rosati, R.: Tractable reasoning and efficient query answering in description logics: The DL-Lite family. J. Automated Reasoning 39(3), 385–429 (2007)
6. Dantsin, E., Eiter, T., Gottlob, G., Voronkov, A.: Complexity and expressive power of logic programming. ACM Computing Surveys 33(3), 374–425 (2001)
7. Di Pinto, F., Lembo, D., Lenzerini, M., Mancini, R., Poggi, A., Rosati, R., Ruzzi, M., Savo, D.F.: Optimizing query rewriting in ontology-based data access. In: Proceedings of the 16th International Conference on Extending Database Technology. pp. 561–572. ACM (2013)
8. Florescu, D., Levy, A., Suciu, D.: Query containment for conjunctive queries with regular expressions. In: Mendelzon, A.O., Paredaens, J. (eds.) Proc. 17th Symposium on Principles of Database Systems (PODS'98). pp. 139–148. ACM (1998)
9. Glimm, B., Ogbuji, C. (eds.): SPARQL 1.1 Entailment Regimes. W3C Recommendation (21 March 2013), available at http://www.w3.org/TR/sparql11-entailment/
10. Gottlob, G., Papadimitriou, C.H.: On the complexity of single-rule datalog queries. Inf. Comput. 183(1), 104–122 (2003)
11. Harris, S., Seaborne, A. (eds.): SPARQL 1.1 Query Language. W3C Recommendation (21 March 2013), available at http://www.w3.org/TR/sparql11-query/
12. Kontchakov, R., Rodriguez-Muro, M., Zakharyaschev, M.: Ontology-based data access with databases: A short course. In: Rudolph, S., Gottlob, G., Horrocks, I., van Harmelen, F. (eds.) Reasoning Web, LNCS, vol. 8067, pp. 194–229. Springer, Mannheim, Germany (2013)
13. Motik, B., Cuenca Grau, B., Horrocks, I., Wu, Z., Fokoue, A., Lutz, C. (eds.): OWL 2 Web Ontology Language: Profiles. W3C Recommendation (27 October 2009), available at http://www.w3.org/TR/owl2-profiles/
14. Patel-Schneider, P.F., Motik, B. (eds.): OWL 2 Web Ontology Language: Mapping to RDF Graphs. W3C Recommendation (27 October 2009), available at http://www.w3.org/TR/owl2-mapping-to-rdf/
15. Pérez, J., Arenas, M., Gutierrez, C.: nSPARQL: A navigational language for RDF. J. Web Semantics 8, 255–270 (2010)
16. Pérez-Urbina, H., Motik, B., Horrocks, I.: Tractable query answering and rewriting under description logic constraints. J. Applied Logic 8(2), 186–209 (2010)
17. Rodriguez-Muro, M., Kontchakov, R., Zakharyaschev, M.: Ontology-based data access: Ontop of databases. In: Alani, H., Kagal, L., Fokoue, A., Groth, P.T., Biemann, C., Parreira, J.X., Aroyo, L., Noy, N.F., Welty, C., Janowicz, K. (eds.) Proc. 12th Int. Semantic Web Conf. (ISWC'13). LNCS, vol. 8218, pp. 558–573. Springer (2013)
18. Rudolph, S., Krötzsch, M.: Flag & check: Data access with monadically defined queries. In: Hull, R., Fan, W. (eds.) Proc. 32nd Symposium on Principles of Database Systems (PODS'13). pp. 151–162. ACM (2013)