Closed-World Semantics for Conjunctive Queries with Negation over $\mathcal{ELH}_\bot$ Ontologies[⋆]

Stefan Borgwardt[0000−0003−0924−8478] and Walter Forkel[0000−0002−0343−5136]

Chair for Automata Theory, Technische Universität Dresden, Germany
firstname.lastname@tu-dresden.de

Abstract. Ontology-mediated query answering is a popular paradigm for enriching answers to user queries with background knowledge. For querying the *absence* of information, however, there exist only few ontology-based approaches. Moreover, these proposals conflate the closed-domain and closed-world assumption, and therefore are not suited to deal with the anonymous objects that are common in ontological reasoning. We propose a new closed-world semantics for answering conjunctive queries with negation over ontologies formulated in the description logic $\mathcal{ELH}_\bot$, which is based on the *minimal* canonical model. We propose a rewriting strategy for dealing with negated query atoms, which shows that query answering is possible in polynomial time in data complexity.

1 Introduction

Ontology-mediated query answering (OMQA) allows using background knowledge for answering user queries, supporting data-focused applications offering search, analytics, or data integration functionality. An *ontology* is a logical theory formulated in a decidable fragment of first-order logic, with a trade-off between the expressivity of the ontology and the efficiency of query answering. *Rewritability* is a popular topic of research, the idea being to reformulate ontological queries into database queries that can be answered by traditional database management systems [8, 10, 15, 21, 27].

Ontology-based systems do not use the *closed-domain* and *closed-world* semantics of databases. Instead, they acknowledge that unknown (*anonymous*) objects may exist (*open domain*) and that facts that are not explicitly stated may still be true (*open world*). Anonymous objects are related to *null* values in databases, but are not used explicitly; for example, if we know that every person has a mother, then first-order models include all mothers, even though they may not be mentioned in the input dataset. The open-world assumption ensures that, if the dataset does not contain an entry on, e.g. whether a person is male or female, then we do not infer that this person is neither male nor female, but rather consider all possibilities.

[⋆] This work was supported by the DFG grant BA 1122/19-1 (GOASQ) and grant 389792660 (TRR 248) (see https://perspicuous-computing.science).

The biomedical domain is a fruitful area for OMQA methods, due to the availability of large ontologies covering a multitude of topics[1] and the demand for managing large amounts of patient data, in the form of *electronic health records (EHRs)* [12]. For example, for the preparation of clinical trials[2] a large number of patients need to be screened for eligibility, and an important area of current research is how to automate this process [7, 23, 28, 29, 31].[3]

However, ontologies and EHRs mostly contain *positive* information, while clinical trials also require certain *exclusion criteria* to be absent in the patients. For example, we may want to select only patients that have *not* been diagnosed with cancer,[4] but such information cannot be entailed from the given knowledge. The culprit for this problem is the open-world semantics, which considers a cancer diagnosis possible unless it has been explicitly ruled out.

One possibility is to introduce (partial) closed-world semantics to ontology languages [1,24]. For example, one can declare the predicate *human* to be "closed", i.e. if an object is not explicitly listed as *human* in the dataset, then it is considered to be not human. However, such approaches fail to deal with anonymous objects; indeed, they conflate the open-world and open-domain assumptions by requiring that all closed information is restricted to the known objects. For example, even if we don't know the mother of a person, we still know that she is human, even though this may not be explicitly stated in the ontology (but entailed by it). Using the semantics of [1, 24] would hence enforce a partial *closed-domain* assumption as well, in that A's mother would have to be a known object from the dataset.

Epistemic logics are another way to give a closed-world-like semantics to negated formulas; e.g. one can formulate queries like "no cancer diagnosis is *known*" using the epistemic knowledge modality **K**. Such formalisms are also unable to deal with closed-world knowledge over anonymous objects [11, 32]. Most closely related to our proposal are Datalog-based semantics for negation, based on the (Skolem) chase construction [2, 18]. We compare all these existing semantics in detail in Section 3.

The contribution of this paper is a new closed-world semantics to answer *conjunctive queries with (guarded) negation* [6] over ontologies formulated in $\mathcal{ELH}_\bot$, an ontology language that covers many biomedical ontologies. Our semantics is based on the *minimal canonical model*, which encodes all inferences of the ontology in the most concise way possible. As a side effect, this means that standard CQs without negation are interpreted under the standard open-world semantics. In order to properly handle negative knowledge about anonymous objects, however, we have to be careful in the construction of the canonical model, in particular about the number and type of anonymous objects that are introduced. Since in general the minimal canonical model is infinite, we develop a rewriting technique, in the spirit of the combined approach of [22, 25], and most closely inspired by [8, 15], which allows us to evaluate conjunctive queries with

[1] https://bioportal.bioontology.org

[2] https://clinicaltrials.gov

[3] https://n2c2.dbmi.hms.harvard.edu

[4] An exclusion criterion in https://clinicaltrials.gov/ct2/show/NCT01463215

negation over a finite part of the canonical model, using traditional database techniques.

An extended version of this paper, including an appendix with full proofs, can be found at `https://tu-dresden.de/inf/lat/papers`.

2 Preliminaries

We recall the definitions of $\mathcal{ELH}_\bot$ and first-order queries, which are needed for our rewriting of conjunctive queries with negation.

The Description Logic $\mathcal{ELH}_\bot$. Let N_C, N_R, N_I be countably infinite sets of *concept*, *role*, and *individual names*, respectively. A *concept* is built according to the syntax rule $C ::= A \mid \top \mid \bot \mid C \sqcap C \mid \exists r.C$, where $A \in N_C$ and $r \in N_R$. An *ABox* is a finite set of *concept assertions* $A(a)$ and *role assertions* $r(a, b)$, where $a, b \in N_I$. A *TBox* is a finite set of *concept inclusions* $C \sqsubseteq D$ and *role inclusions* $r \sqsubseteq s$, where C, D are concepts and r, s are roles. In the following we assume the TBox to be in normal form, i.e. that it contains only inclusions of the form

$$A_1 \sqcap \dots \sqcap A_n \sqsubseteq B, \qquad A \sqsubseteq \exists r.B, \qquad \exists r.A \sqsubseteq B, \qquad r \sqsubseteq s$$

where $A_{(i)} \in N_C \cup \{\top\}$, $B \in N_C \cup \{\bot\}$, $r, s \in N_R$, and $n \geq 1$. A *knowledge base (KB)* (or *ontology*) $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ is a pair of a TBox $\mathcal{T}$ and an ABox $\mathcal{A}$. We refer to the set of individual names occurring in $\mathcal{K}$ by $\mathrm{Ind}(\mathcal{K})$. We write $C \equiv D$ to abbreviate the two inclusions $C \sqsubseteq D$, $D \sqsubseteq C$, and similarly for role inclusions.

The semantics of $\mathcal{ELH}_\bot$ is defined in terms of interpretations $I = (\Delta^I, \cdot^I)$ as usual [5]. In the following, we assume all KBs to be consistent and make the standard name assumption, i.e. that for every individual name a in any interpretation I we have $a^I = a$. An axiom α is *entailed* by $\mathcal{K}$ (written $\mathcal{K} \models \alpha$) if α is satisfied in all models of $\mathcal{K}$. We abbreviate $\mathcal{K} \models C \sqsubseteq D$ to $C \sqsubseteq_{\mathcal{T}} D$, and similarly for role inclusions; note that the ABox does not influence the entailment of inclusions. Entailment in $\mathcal{ELH}_\bot$ can be decided in polynomial time [4].

Query Answering. Let N_V be a countably infinite set of *variables*. The set of *terms* is $N_T := N_V \cup N_I$. A *first-order query* $\phi(\mathbf{x})$ is a first-order formula built from *concept atoms* $A(t)$ and *role atoms* $r(t, t')$ with $A \in N_C$, $r \in N_R$, and $t_i \in N_T$, using the boolean connectives $(\wedge, \vee, \neg, \rightarrow)$ and universal and existential quantifiers $(\forall x, \exists x)$. The free variables $\mathbf{x}$ of $\phi(\mathbf{x})$ are called *answer variables* and we say that ϕ is k-ary if there are k answer variables. The remaining variables are the *quantified variables*. We use $\mathrm{Var}(\phi)$ to denote the set of all variables in ϕ. A query without any answer variables is called a *Boolean query*.

Let $I = (\Delta, \cdot^I)$ be an interpretation. An *assignment* $\pi\colon \mathrm{Var}(\phi) \to \Delta$ *satisfies* ϕ in I, if $I, \pi \models \phi$ under the standard semantics of first-order logic. We write $I \models \phi$ if there is a satisfying assignment for ϕ in I. Let $\mathcal{K}$ be a KB. A k-tuple $\mathbf{a}$ of individual names from $\mathrm{Ind}(\mathcal{K})$ is an *answer* to ϕ in I if ϕ has a satisfying assignment π in I with $\pi(\mathbf{x}) = \mathbf{a}$; it is a *certain answer* to q over $\mathcal{K}$ if it is an answer to q in all models of $\mathcal{K}$. We denote the set of all answers to ϕ in I by $\mathrm{ans}(\phi, I)$, and the set of all certain answers to ϕ over $\mathcal{K}$ by $\mathrm{cert}(\phi, \mathcal{K})$.

A *conjunctive query* (CQ) $q(\mathbf{x})$ is a first-order query of the form $\exists\mathbf{y}.\,\varphi(\mathbf{x},\mathbf{y})$, where φ is a conjunction of atoms. Abusing notation, we write $\alpha \in q$ if the atom α occurs in q, and conversely may treat a set of atoms as a conjunction. The *leaf variables* x in q are those that do not occur in any atoms of the form $r(x,y)$. Clearly, q is satisfied in an interpretation if there is a satisfying assignment for $\varphi(\mathbf{x},\mathbf{y})$, which is often called a *match* for q. A CQ is *rooted* if all variables are connected to an answer variable through role atoms.

CQ answering over $\mathcal{ELH}_\bot$ KBs is *combined first-order rewritable* [25]: For any CQ q and consistent KB $\mathcal{K} = (\mathcal{T},\mathcal{A})$ we can find a first-order query $q_\mathcal{T}$ and a finite interpretation $I_\mathcal{K}$ such that $\operatorname{cert}(q,\mathcal{K}) = \operatorname{ans}(q_\mathcal{T}, I_\mathcal{K})$. Importantly, $I_\mathcal{K}$ is independent of q, i.e. can be reused to answer many different queries, while $q_\mathcal{T}$ is independent of $\mathcal{A}$, i.e. each query can be rewritten without using the (possibly large) dataset. The rewritability results are based crucially on the *canonical model* property of $\mathcal{ELH}_\bot$: For any consistent KB $\mathcal{K}$ one can construct a model $I_\mathcal{K}$ that is homomorphically contained in any other model. This is a very useful property since any match in the canonical model corresponds to matches in all other models of $\mathcal{K}$, and therefore $\operatorname{cert}(q,\mathcal{K}) = \operatorname{ans}(q, I_\mathcal{K})$ holds for all CQs q.

3 Conjunctive Queries With Negation

We are interested in answering queries of the following form.

Definition 1. Conjunctive queries with (guarded) negation *(NCQs) are constructed by extending CQs with* negated concept atoms $\neg A(t)$ *and* negated role atoms $\neg r(t,t')$*, such that, for any negated atom over terms t (and t') the query contains at least one positive atom over t (and t').*

We first discuss different ways of handling the negated atoms, and then propose a new semantics that is based on a particular kind of *minimal* canonical model. For this, we consider an example based on real EHRs (ABoxes) from the MIMIC-III database [20], criteria (NCQs) from clinicaltrials.gov, and the large medical ontology SNOMED CT[5] (the TBox). We omit here the “role groups” used in SNOMED CT, which do not affect the example. We also simplify the concept names and their definitions for ease of presentation. We assume that the ABoxes have been extracted from EHRs by a natural language processing tool based, e.g. on existing concept taggers like [3,30]; of course, this extraction is an entire research field in itself, which we do not attempt to tackle in this paper.

Example 2. We consider three patients. Patient p_1 (patient 2693 in the MIMIC-III dataset) is diagnosed with breast cancer and an unspecified form of cancer (this often occurs when there are multiple mentions of cancer in a patient’s EHR, which cannot be resolved to be the same entity). Patient p_2 (patient 32304 in the MIMIC-III dataset) suffers from breast cancer and skin cancer (“[S]tage IV breast cancer with mets to skin, bone, and liver”). For p_3 (patient 88432 in the

[5] https://www.snomed.org/snomed-ct

MIMIC-III dataset), we know that p_3 has breast cancer that involves the skin ("Skin, left breast, punch biopsy: Poorly differentiated carcinoma").

Since SNOMED CT does not model patients, we add a special role name *diagnosedWith* that connects patients with their diagnoses. One can use this to express diagnoses in two ways. First, one can explicitly introduce individual names for diagnoses in assertions like $\mathrm{diagnosedWith}(p_1, d_1)$, $\mathrm{BreastCancer}(d_1)$, $\mathrm{diagnosedWith}(p_1, d_2)$, $\mathrm{Cancer}(d_2)$, implying that these diagnoses are treated as distinct entities under the standard name assumption. Alternatively, one can use complex assertions like $\exists\mathrm{diagnosedWith}.\mathrm{Cancer}(p_1)$, which allows the logical semantics to resolve whether two diagnoses actually refer to the same object. Since ABoxes only contain concept names, in this case one has to introduce auxiliary definitions like $\mathrm{CancerPatient} \equiv \exists\mathrm{diagnosedWith}.\mathrm{Cancer}$ into the TBox. We use both variants in our example, to illustrate their different behaviours.

We obtain the KB $\mathcal{K}_C$, containing knowledge about different kinds of cancers and cancer patients, together with information about the three patients. The information about cancers is taken from SNOMED CT (in simplified form):

$$\begin{aligned}\mathrm{SkinCancer} &\equiv \mathrm{Cancer} \sqcap \exists\mathrm{findingSite}.\mathrm{SkinStructure}\\ \mathrm{BreastCancer} &\equiv \mathrm{Cancer} \sqcap \exists\mathrm{findingSite}.\mathrm{BreastStructure}\\ \mathrm{SkinOfBreastCancer} &\equiv \mathrm{Cancer} \sqcap \exists\mathrm{findingSite}.\mathrm{SkinOfBreastStructure}\\ \mathrm{SkinOfBreastStructure} &\sqsubseteq \mathrm{BreastStructure} \sqcap \mathrm{SkinStructure}\end{aligned}$$

The EHRs are compiled into several assertions per patient:

$$\begin{aligned}&\text{Patient } p_1\text{: } \mathrm{BreastCancerPatient}(p_1),\ \mathrm{CancerPatient}(p_1)\\ &\text{Patient } p_2\text{: } \mathrm{SkinCancerPatient}(p_2),\ \mathrm{BreastCancerPatient}(p_2)\\ &\text{Patient } p_3\text{: } \mathrm{diagnosedWith}(p_3, c_3),\ \mathrm{SkinOfBreastCancer}(c_3)\end{aligned}$$

Additionally, we add the following auxiliary definitions to the TBox:

$$\begin{aligned}\mathrm{CancerPatient} &\equiv \exists\mathrm{diagnosedWith}.\mathrm{Cancer}\\ \mathrm{SkinCancerPatient} &\equiv \exists\mathrm{diagnosedWith}.\mathrm{SkinCancer}\\ \mathrm{BreastCancerPatient} &\equiv \exists\mathrm{diagnosedWith}.\mathrm{BreastCancer}\end{aligned}$$

For example, skin cancers and breast cancers are cancers occurring at specific parts of the body ("body structure" in SNOMED CT), and a breast cancer patient is someone who is diagnosed with breast cancer. This means that, in every model of $\mathcal{K}_C$, every object that satisfies BreastCancerPatient (in particular p_2) must have a diagnosedWith-connected object that satisfies BreastCancer, and so on.

For a clinical trial,[6] we want to find patients that have "breast cancer", but not "breast cancer that involves the skin." This can be translated into an NCQ:

$$\begin{aligned}q_B(x) := \exists y, z.\, &\mathrm{diagnosedWith}(x, y) \land \mathrm{Cancer}(y) \land \mathrm{findingSite}(y, z) \land\\ &\mathrm{BreastStructure}(z) \land \neg\mathrm{SkinStructure}(z)\end{aligned}$$

[6] https://clinicaltrials.gov/ct2/show/NCT01960803

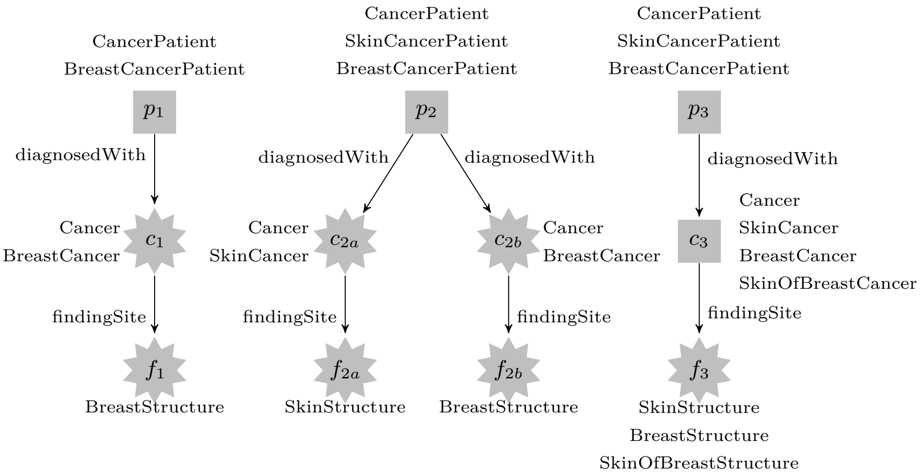

Fig. 1. The minimal canonical model $I_{\mathcal{K}_C}$. Named individuals are depicted by squares, anonymous objects by stars.

We know that p_1 is diagnosed with BreastCancer as well as Cancer. Since the former is more specific, we assume that the latter refers to the same BreastCancer. However, since we have no information about an involvement of the skin, p_1 should be returned as an answer to q_B.

We know that p_2 suffers from cancer in the skin and the breast, but not if the skin of the breast is also affected. Since neither location is implied by the other, we assume that they refer to distinct areas. p_2 should thus be an answer to q_B.

In the case of p_3, it is explicitly stated that it is the same cancer that is occurring (not necessarily exclusively) at the skin of the breast. In this case, the ABox assertions override the distinctness assumption we made for p_2. Thus, p_3 should not be an answer to q_B. ■

In practice, more complicated cases than in our example can occur: The nesting of anonymous objects will be deeper and more branched when using large biomedical ontologies. For example, in SNOMED CT it is possible to describe many details of a cancer, such as the kind of cancer, whether it is a primary or secondary cancer, and in which part of the body it is found. This means that even a single assertion can lead to the introduction of multiple levels of anonymous objects in the canonical model. In some ontologies there are even cyclic concept inclusions, which lead to infinitely many anonymous individuals, e.g. in the GALEN ontology[7]. We focus on Example 2 in this paper, to illustrate the relevant issues in a clear and easy to follow manner.

We now evaluate existing semantics on this example.

Standard Certain Answer Semantics as defined in Section 2 is clearly not suited here, because one can easily construct a model of $\mathcal{K}_C$ in which c_1 is

[7] http://www.opengalen.org/

also a skin cancer, and hence p_1 is not an element of $\mathrm{cert}(q_B, \mathcal{K}_C)$. Moreover, under certain answer semantics answering CQs with guarded negation is already coNP-complete [17], and hence not (combined) rewritable.

Epistemic Logic allows us to selectively apply closed-world reasoning using the modal knowledge operator $\mathbf{K}$. For a formula $\mathbf{K}\varphi$ to be true, it has to hold in all "connected worlds", which is often considered to mean all possible models of the KB, adopting an $S5$-like view [11]. For q_B, we could read $\neg\mathrm{SkinStructure}(z)$ as "not known to be a skin structure", i.e. $\neg\mathbf{K}\mathrm{SkinStructure}(z)$. Consider the model $I_{\mathcal{K}_C}$ in Figure 1 and the assignment $\pi = \{x \mapsto p_3, y \mapsto c_3, z \mapsto f_3\}$, for which we want to check whether it is a match for q_B. Under epistemic semantics, $\neg\mathbf{K}\mathrm{SkinStructure}(z)$ is considered true if $\mathcal{K}$ has a (different) model in which f_3 does not belong to SkinStructure. However, f_3 is an anonymous object, and hence its name is not fixed. For example, we can easily obtain another model by renaming f_3 to f_1 and vice versa. Then f_3 would not be a skin structure, which means that $\neg\mathbf{K}\mathrm{SkinStructure}(z)$ is true in the original model $I_{\mathcal{K}_C}$, which is not what we expected. This is a known problem with epistemic first-order logics [32].

Skolemization can enforce a stricter comparison of anonymous objects between models. The inclusion SkinOfBreastCancer $\sqsubseteq$ $\exists$findingSite.SkinOfBreast could be rewritten as the first-order sentence

$$\forall x.\ \Big(\mathrm{SkinOfBreastCancer}(x) \to \mathrm{findingSite}\big(x, f(x)\big) \land \mathrm{SkinOfBreast}\big(f(x)\big)\Big),$$

where f is a fresh function symbol. This means that c_3 would be connected to a finding site that has the unique name $f(c_3)$ in every model. Queries would be evaluated over Herbrand models only. Hence, for evaluating $\neg\mathbf{K}\mathrm{SkinStructure}(z)$ when z is mapped to $f(c_3)$, we would only be allowed to compare the behavior of $f(c_3)$ in other Herbrand models. The general behavior of this anonymous individual is fixed, however, since in all Herbrand models it is *the* finding site of c_3. While this improves the comparison by introducing pseudo-names for all anonymous individuals, it limits us in different ways: Since p_3 is inferred to be a BreastCancerPatient, the Skolemized version of BreastCancerPatient $\sqsubseteq$ $\exists$diagnosedWith.BreastCancer introduces a new successor $g(p_3)$ of p_3 satisfying BreastCancer, which, together with the definition of BreastCancer, means that p_3 is an answer to q_B since there is an additional breast cancer diagnosis that does not involve the skin.

Datalog-based Ontology Languages with negation [2, 18] are closely related to Skolemized ontologies, since their semantics is often based on the so-called *Skolem chase* [26]. This is closer to the semantics we propose in Section 3.1, in that a single canonical model is used for all inferences. However, it suffers from the same drawback of Skolemization described above, due to superfluous successors. To avoid this, our semantics uses a special minimal canonical model (see Definition 4), which is similar to the *restricted chase* [16] or the *core chase* [14], but always produces a unique model without having to merge domain elements. To the best of our knowledge, there exist no complexity results for Datalog-based languages with negation over the these other chase variants.

Closed Predicates are a way to declare, for example, the concept name SkinStructure as "closed", which means that all skin structures must be declared explicitly, and no other SkinStructure object can exist [1,24]. This provides a way to give answers to negated atoms as in q_B. However, as explained in the introduction, this mechanism is not suitable for anonymous objects since it means that only named individuals can satisfy SkinStructure. When applied to $\mathcal{K}_C$, the result is even worse: Since there is no (named) SkinStructure object, no skin structures can exist at all and $\mathcal{K}_C$ becomes inconsistent. Closed predicates are appropriate in cases where the KB contains a full list of all instances of a certain concept name, and no other objects should satisfy it; but they are not suitable to infer negative information about anonymous objects. Moreover, CQ answering with closed predicates in $\mathcal{ELH}_\bot$ is already coNP-hard [24].

3.1 Semantics for NCQs

We propose to answer NCQs over a special canonical model of the knowledge base. On the one hand, this eliminates the problem of tracking anonymous objects across different models, and on the other hand enables us to encode our assumptions directly into the construction of the model. In particular, we should only introduce the minimum necessary number of anonymous objects since, unlike in standard CQ answering, the precise shape and number of anonymous objects has an impact on the semantics of negated atoms.

Given $\mathcal{K}_C$, in contrast to the Skolemized semantics, we will not create both a generic "Cancer" and another "BreastCancer" successor for p_1, because the BreastCancer is also a Cancer, and hence the first object is redundant. Therefore, in the minimal canonical model of $\mathcal{K}_C$ depicted in Figure 1, for patient p_1 only one successor is introduced to satisfy the definitions of both BreastCancerPatient and CancerPatient at the same time. In contrast, p_2 has two successors, because BreastCancer and SkinCancer do not imply each other. Finally, for p_3 the ABox contains a single successor that is a SkinOfBreastCancer, which implies a single findingSite-successor that satisfies both SkinStructure and BreastStructure.

To detect whether an object required by an existential restriction $\exists r.A$ is redundant, we use the following notion of minimality.

Definition 3 (Structural Subsumption). *Let $\exists r.A$, $\exists t.B$ be concepts with $A, B \in N_C$ and $r, t \in N_R$. We say that $\exists r.A$ is* structurally subsumed *by $\exists t.B$ (written $\exists r.A \sqsubseteq^s_\mathcal{T} \exists t.B$) if $r \sqsubseteq_\mathcal{T} t$ and $A \sqsubseteq_\mathcal{T} B$.*

Given a set V of existential restrictions, we say that $\exists r.A \in V$ is minimal w.r.t. $\sqsubseteq^s_\mathcal{T}$ (in V) *if there is no $\exists t.B \in V$ such that $\exists t.B \sqsubseteq^s_\mathcal{T} \exists r.A$.*

A CQ $q_1(\mathbf{x})$ is structurally subsumed *by a CQ $q_2(\mathbf{x})$ with the same answer variables (written $q_1 \sqsubseteq^s_\mathcal{T} q_2$) if, for all $x, y \in \mathbf{x}$, it holds that*

$$\bigsqcap_{\alpha(x)\in q_1} \alpha \sqsubseteq_\mathcal{T} \bigsqcap_{\alpha(x)\in q_2} \alpha, \text{ and } \bigsqcap_{\alpha(x,y)\in q_1} \alpha \sqsubseteq_\mathcal{T} \bigsqcap_{\alpha(x,y)\in q_2} \alpha,$$

where role conjunction is interpreted in the standard way [5].

In contrast to standard subsumption, $\exists r.A$ is not structurally subsumed by $\exists t.B$ w.r.t. the TBox $\mathcal{T} = \{\exists r.A \sqsubseteq \exists t.B\}$, as neither $r \sqsubseteq_\mathcal{T} t$ nor $A \sqsubseteq_\mathcal{T} B$ hold. Similarly, structural subsumption for CQs considers all (pairs of) variables separately.

We use this notion to define the minimal canonical model.

Definition 4 (Minimal Canonical Model). *Let $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ be an $\mathcal{ELH}_\bot$ KB. We construct the* minimal canonical model $I_\mathcal{K}$ *of $\mathcal{K}$ as follows:*

1. *Set $\Delta^{I_\mathcal{K}} := N_I$ and $a^{I_\mathcal{K}} := a$ for all $a \in N_I$.*
2. *Define $A^{I_\mathcal{K}} := \{a \mid \mathcal{K} \models A(a)\}$ for all $A \in N_C$ and $r^{I_\mathcal{K}} := \{(a,b) \mid \mathcal{K} \models r(a,b)\}$ for all $r \in N_R$.*
3. *Repeat:*
   - *(a) Select an element $d \in \Delta^{I_\mathcal{K}}$ that has not been selected before and let $V := \{\exists r.B \mid d \in A^{I_\mathcal{K}}$ and $d \notin (\exists r.B)^{I_\mathcal{K}}$ with $A \sqsubseteq_\mathcal{T} \exists r.B,\ A, B \in N_C\}$.*
   - *(b) For each $\exists r.B \in V$ that is minimal w.r.t. $\sqsubseteq^s_\mathcal{T}$, add a fresh element e to $\Delta^{I_\mathcal{K}}$, for each $B \sqsubseteq_\mathcal{T} A$ add e to $A^{I_\mathcal{K}}$, and for each $r \sqsubseteq_\mathcal{T} s$ add (d,e) to $s^{I_\mathcal{K}}$.*

By $I_\mathcal{A}$ we denote the restriction of $I_\mathcal{K}$ to named individuals, i.e. the result of applying only Steps 1 and 2, but not Step 3.

If Step 3 is applied fairly, i.e. such that each new domain element that is created in (b) is eventually also selected in (a), then $I_\mathcal{K}$ is indeed a model of $\mathcal{K}$ (if $\mathcal{K}$ is consistent at all). In particular, all required existential restrictions are satisfied at each domain element, because the existential restrictions that are minimal w.r.t. $\sqsubseteq^s_\mathcal{T}$ entail all others.

Moreover, $I_\mathcal{K}$ satisfies the properties expected of a canonical model [15,25]: it can be homomorphically embedded into any other model of $\mathcal{K}$, and therefore $\mathrm{cert}(q, \mathcal{K}) = \mathrm{ans}(q, I_\mathcal{K})$ holds for all CQs q. We now define the semantics of NCQs as described before, i.e. by evaluating them as first-order formulas over the minimal canonical model $I_\mathcal{K}$, which ensures that our semantics is compatible with the usual certain-answer semantics for CQs.

Definition 5 (Minimal-World Semantics). *The* (minimal-world) answers *to an NCQ q over a consistent $\mathcal{ELH}_\bot$ KB $\mathcal{K}$ are* $\mathrm{mwa}(q, \mathcal{K}) := \mathrm{ans}(q, I_\mathcal{K})$.

For Example 2, we get $\mathrm{mwa}(q_B, \mathcal{K}_C) = \{p_1, p_2\}$ (see Figure 1), which is exactly as intended. Unfortunately, in general the minimal canonical model is infinite, and we cannot evaluate the answers directly. Hence, we employ a rewriting approach to reduce NCQ answering over the minimal canonical model to (first-order) query answering over $I_\mathcal{A}$ only.

4 A Combined Rewriting for NCQs

We show that NCQ answering is combined first-order rewritable. As target representation, we obtain first-order queries of a special form.

Definition 6 (Filtered query). *Let $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ be an $\mathcal{ELH}_\bot$ KB. A* filter *on a variable z is a first-order expression $\psi(z)$ of the form*

$$\big(\exists z'.\psi^+(z,z')\big) \to \big(\exists z'.\psi^+(z,z') \land \psi^-(z,z') \land \Psi\big) \qquad (1)$$

where $\psi^+(z,z')$ is a conjunction of atoms of the form $A(z')$ or $r(z,z')$, that contains at least one role atom, $\psi^-(z,z')$ is a conjunction of negated atoms $\neg A(z')$ or $\neg r(z,z')$, and Ψ is a (possibly empty) set of filters on z'.

A filtered query *ϕ is of the form $\exists \mathbf{y}.\big(\varphi(\mathbf{x},\mathbf{y}) \land \Psi\big)$ where $\exists \mathbf{y}.\varphi(\mathbf{x},\mathbf{y})$ is an NCQ and Ψ is a set of filters on leaf variables in φ. It is* rooted *if $\exists \mathbf{y}.\varphi(\mathbf{x},\mathbf{y})$ is rooted.*

Note that every NCQ is a filtered query where the set of filters Ψ is empty.

We will use filters to check for the existence of "typical" successors, i.e. role successors that behave like the ones that are introduced by the canonical model construction to satisfy an existential restriction. In particular, a typical successor does not satisfy any superfluous concept or role atoms. For example, in Figure 1 the element c_1 introduced to satisfy ∃diagnosedWith.BreastCancer for p_1 is a typical successor, because it satisfies only BreastCancer and Cancer and not, e.g. SkinCancer. In contrast, the diagnosedWith-successor c_3 of p_3 is atypical, since the ontology does not contain an existential restriction ∃diagnosedWith.SkinOfBreastCancer that could have introduced such a successor in the canonical model.

The idea of the rewriting procedure is to not only rewrite the positive part of the query, as in [8, 15], but to also ensure that no critical information is lost. This is accomplished by rewriting the negative parts and by saving the structure of the eliminated part of the query in the filter. A filter on z ensures that the rewritten query can only be satisfied by mapping z to an anonymous individual in the canonical model, or to a named individual that behaves in a similar way.

Definition 7 (Rewriting). *Let $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ be a KB and $\phi = \exists \mathbf{y}.\varphi(\mathbf{x},\mathbf{y}) \land \Psi$ be a rooted filtered query. We write $\phi \to_\mathcal{T} \phi'$ if ϕ' can be obtained from ϕ by applying the following steps:*

(S1) Select a quantified leaf variable $\hat{x}$ in φ. Let $\hat{y}$ be a fresh variable and select

$$\begin{aligned}
\mathsf{Pred} &:= \{y \mid r(y,\hat{x}) \in \varphi\} \cup \{y \mid \neg r(y,\hat{x}) \in \varphi\} && \textit{(predecessors of } \hat{x}\textit{)},\\
\mathsf{Pos} &:= \{A(\hat{x}) \in \varphi\} \cup \{r(\hat{y},\hat{x}) \mid r(y,\hat{x}) \in \varphi\} && \textit{(positive atoms for } \hat{x}\textit{)},\\
\mathsf{Neg} &:= \{\neg A(\hat{x}) \in \varphi\} \cup \{\neg r(\hat{y},\hat{x}) \mid \neg r(y,\hat{x}) \in \varphi\} && \textit{(negative atoms for } \hat{x}\textit{)}.
\end{aligned}$$

(S2) Select some $M \sqsubseteq_\mathcal{T} \exists s.N$ with $M, N \in N_C$ that satisfies all of the following:
- *(a) $s(\hat{y},\hat{x}) \land N(\hat{x}) \sqsubseteq^s_\mathcal{T} \mathsf{Pos}$, and*
- *(b) $s(\hat{y},\hat{x}) \land N(\hat{x}) \not\sqsubseteq^s_\mathcal{T} \alpha$ for all $\neg\alpha \in \mathsf{Neg}$.*

(S3) Let $\mathcal{M}'$ be the set of all $M' \in N_C$ such that $M' \sqsubseteq_\mathcal{T} \exists s'.N'$ with $N' \in N_C$,
- *(a) $\exists s'.N' \sqsubseteq^s_\mathcal{T} \exists s.N$ (where $\exists s.N$ was chosen in (S2)), and*
- *(b) $s'(\hat{y},\hat{x}) \land N'(\hat{x}) \sqsubseteq^s_\mathcal{T} \alpha$ for some $\neg\alpha \in \mathsf{Neg}$.*

(S4) Drop from φ every atom that contains $\hat{x}$.

(S5) Replace all variables $y \in \mathsf{Pred}$ in φ with $\hat{y}$.

(S6) Add the atoms $M(\hat{y})$ *and* $\{\neg M'(\hat{y}) \mid M' \in \mathcal{M}'\}$ *to* φ.
(S7) Set the new filters to $\Psi' := \Psi \cup \{\psi^*(\hat{y})\} \setminus \Psi_{\hat{x}}$, *where* $\Psi_{\hat{x}} := \{\psi(\hat{x}) \in \Psi\}$ *and*

$$\psi^*(\hat{y}) := \big(\exists \hat{x}.\, s(\hat{y}, \hat{x}) \wedge N(\hat{x})\big) \to \big(\exists \hat{x}.\, s(\hat{y}, \hat{x}) \wedge N(\hat{x}) \wedge \mathsf{Neg} \wedge \Psi_{\hat{x}}\big).$$

We write $\phi \to^*_{\mathcal{T}} \phi'$ *if there exists a finite sequence* $\phi \to_{\mathcal{T}} \cdots \to_{\mathcal{T}} \phi'$. *Furthermore, let* $\mathrm{rew}_{\mathcal{T}}(\phi) := \{\phi' \mid \phi \to^*_{\mathcal{T}} \phi'\}$ *denote the finite set of all rewritings of* ϕ.

There can only be a finite number of rewritings for a given query since there is only a finite number of possible subsumptions $M \sqsubseteq_{\mathcal{T}} \exists s.N$ that can be used for rewriting steps. Additionally, in every step one variable ($\hat{x}$) is eliminated from the NCQ part of the filtered query. Since the query is rooted, there always exists at least one predecessor that is renamed to $\hat{y}$, hence the introduction of $\hat{y}$ never increases the number of variables. Finally, it is easy to see that rewriting a rooted query always yields a rooted query.

The rewriting of Neg to the new negated atoms (via $\mathcal{M}'$ in (S6)) ensures that we do not lose important exclusion criteria, which may result in too many answers. Similarly, the filters exclude atypical successors in the ABox that may result in spurious answers. Both of these constructions are necessary.

Example 8. Consider the query q_B from Example 2. Using Definition 7, we obtain the first-order queries $\phi_B = q_B$, ϕ'_B, and ϕ''_B, where

$$\begin{aligned}\phi'_B = {} & \exists y.\, \mathrm{diagnosedWith}(x,y) \wedge \mathrm{BreastCancer}(y) \wedge \neg\mathrm{SkinOfBreastCancer}(y) \wedge \\ & \Big(\big(\exists z.\, \mathrm{findingSite}(y,z) \wedge \mathrm{BreastStructure}(z)\big) \to \\ & \quad \big(\exists z.\, \mathrm{findingSite}(y,z) \wedge \mathrm{BreastStructure}(z) \wedge \neg\mathrm{SkinStructure}(z)\big)\Big)\end{aligned}$$

results from choosing z in (S1), BreastCancer $\sqsubseteq_{\mathcal{K}_C}$ $\exists$findingSite.BreastStructure in (S2), and computing $\mathcal{M}' = \{\mathrm{SkinOfBreastCancer}\}$ in (S3), and

$$\begin{aligned}\phi''_B = {} & \mathrm{BreastCancerPatient}(x) \wedge \\ & \Big((\exists y.\, \mathrm{diagnosedWith}(x,y) \wedge \mathrm{BreastCancer}(y)) \to \\ & \quad (\exists y.\, \mathrm{diagnosedWith}(x,y) \wedge \mathrm{BreastCancer}(y) \wedge \neg\mathrm{SkinOfBreastCancer}(y)) \wedge \\ & \quad \big((\exists z.\, \mathrm{findingSite}(y,z) \wedge \mathrm{BreastStructure}(z)) \to \\ & \quad\quad (\exists z.\, \mathrm{findingSite}(y,z) \wedge \mathrm{BreastStructure}(z) \wedge \neg\mathrm{SkinStructure}(z))\big)\Big)\end{aligned}$$

is obtained due to BreastCancerPatient $\sqsubseteq_{\mathcal{K}_C}$ $\exists$diagnosedWith.BreastCancer. We omitted the redundant atoms Cancer(y) for clarity.

The finite interpretation $I_{\mathcal{A}_C}$ can be seen in Figure 1 by ignoring all star-shaped nodes. When computing the answers over $I_{\mathcal{A}_C}$, we obtain

$$\mathrm{ans}(\phi_B, I_{\mathcal{A}_C}) = \emptyset,\ \mathrm{ans}(\phi'_B, I_{\mathcal{A}_C}) = \emptyset,\ \text{and}\ \mathrm{ans}(\phi''_B, I_{\mathcal{A}_C}) = \{p_1, p_2\}.$$

For ϕ'_B, the conjunct $\neg$SkinOfBreastCancer(y) is necessary to exclude p_3 as an answer. In ϕ''_B, p_3 is excluded due to the filter that detects c_3 as an atypical successor, because it satisfies not only BreastCancer, but also SkinOfBreastCancer. Hence, both (S6) and (S7) are necessary steps in our rewriting. ■

4.1 Correctness

In Definition 7, the new filter $\psi^*(\hat{y})$ may end up inside another filter expression after applying subsequent rewriting steps, i.e. by rewriting w.r.t. $\hat{y}$. In this case, however, the original structure of the rewriting is preserved, including all internal filters as well as the atoms $M(\hat{y})$, which are included implicitly by $\exists s.N \sqsubseteq M$, and $\{\neg M'(\hat{y}) \mid M' \in \mathcal{M}'\}$, which are included in Neg. We exploit this behavior to show that, whenever a rewritten query is satisfied in the finite interpretation $I_\mathcal{A}$, then it is also satisfied in $I_\mathcal{K}$. This is the most interesting part of the correctness proof, because it differs from the known constructions for ordinary CQs, for which this step is trivial.

Lemma 9. *Let $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ be a consistent $\mathcal{ELH}_\bot$ KB and ϕ be a rooted NCQ. Then, for all $\phi' \in \mathrm{rew}_\mathcal{T}(\phi)$,*

$$\mathrm{ans}(\phi', I_\mathcal{A}) \subseteq \mathrm{mwa}(\phi', \mathcal{K}).$$

Proof. Let $\phi' = \exists \mathbf{y}.(\varphi(\mathbf{x}, \mathbf{y}) \land \Psi)$ and π be an assignment of $\mathbf{x}, \mathbf{y}$ to N_I such that $I_\mathcal{A}, \pi \models \varphi(\mathbf{x}, \mathbf{y})$. Since $I_\mathcal{A}$ and $I_\mathcal{K}$ coincide on the domain N_I, we also have $I_\mathcal{K}, \pi \models \varphi(\mathbf{x}, \mathbf{y})$. Consider any filter $\psi(z) = \exists z'.\psi^+(z, z') \to \exists z'.(\beta(z, z') \land \Psi^*)$ in Ψ, where $\beta(z, z') := \psi^+(z, z') \land \psi^-(z, z')$. Then $\psi(z)$ was introduced at some point during the rewriting, suppose by selecting $M \sqsubseteq_\mathcal{T} \exists s.N$ in (S2). This means that φ contains the atom $M(z)$, and hence $d := \pi(z)$ is a named individual that is contained in $M^{I_\mathcal{A}} \subseteq M^{I_\mathcal{K}}$. By (S2), this means that $I_\mathcal{K}, \pi \models \exists z'.\psi^+(z, z')$, and we have to show that $I_\mathcal{K}, \pi \models \exists z'.(\beta(z, z') \land \Psi^*)$:

1. If $I_\mathcal{A}, \pi \models \exists z'.\beta(z, z')$, then $I_\mathcal{K}, \pi \models \exists z'.\beta(z, z')$ by the same argument as for $\varphi(\mathbf{x}, \mathbf{y})$ above, and we can proceed by induction on the structure of the filters to show that the inner filters Ψ^* are satisfied by the assignment π (extended appropriately for z').
2. If $I_\mathcal{A}, \pi \not\models \exists z'.\beta(z, z')$, then we cannot use a named individual to satisfy the filter $\psi(z)$ in $I_\mathcal{K}$. Moreover, since $I_\mathcal{A}$ satisfies $\psi(z)$, we also know that $I_\mathcal{A}, \pi \not\models \exists z'.\psi^+(z, z')$. Since $\psi^+(z, z') = s(z, z') \land N(z')$, this implies that $d \notin (\exists s.N)^{I_\mathcal{A}}$. Hence, $\exists s.N$ is included in the set V constructed in Step 3(a) of the canonical model construction for the element $d = \pi(z)$. Thus, there exists $M' \sqsubseteq_\mathcal{T} \exists s'.N'$ such that $d \in (M')^{I_\mathcal{A}}$, $d \notin (\exists s'.N')^{I_\mathcal{A}}$, and $\exists s'.N' \sqsubseteq^s_\mathcal{T} \exists s.N$. By Step 3(b), $I_\mathcal{K}$ must contain an element d' such that $d' \in A^{I_\mathcal{K}}$ iff $N' \sqsubseteq_\mathcal{T} A$ and $(d, d') \in r^{I_\mathcal{K}}$ iff $s' \sqsubseteq_\mathcal{T} r$. Since $N' \sqsubseteq_\mathcal{T} N$ and $s' \sqsubseteq_\mathcal{T} s$, we obtain that $I_\mathcal{K}, \pi \cup \{z' \mapsto d'\} \models \psi^+(z, z')$.

   We show that the assignment $\pi \cup \{z' \mapsto d'\}$ also satisfies $\psi^-(z, z') = \mathsf{Neg}$. Assume to the contrary that there is $\neg A(z') \in \mathsf{Neg}$ such that $d' \in A^{I_\mathcal{K}}$ (the case of negated role atoms is again analogous). Then we have $N' \sqsubseteq_\mathcal{T} A$, which shows that all conditions of (S3) are satisfied, and hence M' must be included in $\mathcal{M}'$. Since the atoms $\{\neg M'(\hat{y}) \mid M' \in \mathcal{M}'\}$ are contained in φ, we know that they are satisfied by π in $I_\mathcal{K}$, i.e. $d \notin (M')^{I_\mathcal{K}}$ and hence also $d \notin (M')^{I_\mathcal{A}}$, which is a contradiction.

   It remains to show that the inner filters Ψ^* are satisfied by the assignment $\pi \cup \{z' \mapsto d'\}$ in $I_\mathcal{K}$. Since we are now dealing with an anonymous domain element d', we can use similar, but simpler, arguments as above to prove this by induction on the structure of the filters. This is possible because the atoms $s(\hat{y}, \hat{x})$, $N(\hat{x})$ implied by $M(\hat{y})$ and the negated atoms induced by $\mathcal{M}'$ are present in the query even if the filter is integrated into another filter during a subsequent rewriting step. □

We can use this lemma to show correctness of our approach, i.e. the answers returned for the *union* of queries given by $\mathrm{rew}_\mathcal{T}(\phi)$ over $I_\mathcal{A}$ are exactly the answers of the original NCQ ϕ over $I_\mathcal{K}$. The proof, which can be found in the extended version, is based on existing proofs for ordinary CQs [8, 15], extended appropriately to deal with the filters.

Lemma 10. *Let $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ be a consistent $\mathcal{ELH}_\bot$ KB and let $\phi(\mathbf{x})$ be a rooted NCQ. Then, for all $\phi' \in \mathrm{rew}_\mathcal{T}(\phi)$,*

$$\mathrm{mwa}(\phi, \mathcal{K}) = \bigcup_{\phi' \in \mathrm{rew}_\mathcal{T}(\phi)} \mathrm{ans}(\phi', I_\mathcal{A}).$$

We obtain the claimed complexity result.

Theorem 11. *Checking whether a given tuple* $\mathbf{a}$ *is a closed-world answer to an NCQ ϕ over a consistent $\mathcal{ELH}_\bot$ KB $\mathcal{K}$ can be done in polynomial time in data complexity.*

Under data complexity assumptions, ϕ and $\mathcal{T}$, and hence $\mathrm{rew}_\mathcal{T}(\phi)$, are fixed, and $I_\mathcal{A}$ is of polynomial size in the size of $\mathcal{A}$. However, if we want to use complex assertions in $\mathcal{A}$, as in Example 2, this leads to the introduction of additional acyclic definitions $\mathcal{T}'$, which are not fixed. The complexity nevertheless remains the same: Since $\mathcal{T}$ does not use the new concept names in $\mathcal{T}'$, we can apply the rewriting only w.r.t. $\mathcal{T}$, and extend $I_\mathcal{A}$ by a polynomial number of new elements that result from applying Definition 4 only w.r.t. $\mathcal{T}'$.

What is more important than the complexity result is that this approach can be used to evaluate NCQs using standard database methods, e.g. using views to define the finite interpretation $I_\mathcal{A}$ based on the input data given in $\mathcal{A}$, and SQL queries to evaluate the elements of $\mathrm{rew}_\mathcal{T}(\phi)$ over these views [22].

5 Conclusion

Dealing with the absence of information is an important and at the same time challenging task. In many real-world scenarios, it is not clear whether a piece of information is missing because it is unknown or because it is false. EHRs mostly talk about positive diagnoses and it would be impossible to list all the negative diagnoses, i.e. the diseases a patient does not suffer from. We showed that such a setting cannot be handled adequately by existing logic-based approaches, mostly because they do not deal with closed-world negation over anonymous objects. We introduced a novel semantics for answering conjunctive queries with negation and showed that it is well-behaved also for anonymous objects. Moreover, we demonstrated combined first-order rewritability, which allows us to answer NCQs by using conventional relational database technologies.

We are working on an optimized implementation of this method with the aim to deal with queries over large ontologies such as SNOMED CT. On the theoretical side, we will further develop our approach to also represent temporal and numeric information, such as the precise order and duration of a patient's illnesses and treatments, and the dosage of medications. Such information is important for evaluating the eligibility criteria of clinical trials [9,13,19].

References

1. Ahmetaj, S., Ortiz, M., Simkus, M.: Polynomial datalog rewritings for expressive description logics with closed predicates. In: Kambhampati, S. (ed.) Proc. of the 25th Int. Joint Conf. on Artificial Intelligence (IJCAI'16). pp. 878–885. AAAI Press (2016), `https://www.ijcai.org/Abstract/16/129`
2. Arenas, M., Gottlob, G., Pieris, A.: Expressive languages for querying the semantic web. In: Hull, R., Grohe, M. (eds.) Proc. of the 33rd Symp. on Principles of Database Systems (PODS'14). pp. 14–26. ACM (2014). https://doi.org/10.1145/2594538.2594555
3. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In: Proceedings of the AMIA Symposium. pp. 17–21. American Medical Informatics Association (2001)
4. Baader, F., Brandt, S., Lutz, C.: Pushing the $\mathcal{EL}$ envelope. In: Kaelbling, L.P., Saffiotti, A. (eds.) Proc. of the 19th Int. Joint Conf. on Artificial Intelligence (IJCAI'05). pp. 364–369. Professional Book Center (2005), `http://ijcai.org/Proceedings/09/Papers/053.pdf`
5. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, 2 edn. (2007)
6. Bárány, V., ten Cate, B., Otto, M.: Queries with guarded negation. Proc. of the VLDB Endowment **5**(11), 1328–1339 (2012). https://doi.org/10.14778/2350229.2350250
7. Besana, P., Cuggia, M., Zekri, O., Bourde, A., Burgun, A.: Using semantic web technologies for clinical trial recruitment. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) Proc. of the 9th Int. Semantic Web Conf. (ISWC'10). Lecture Notes in Computer Science, vol. 6497, pp. 34–49. Springer (2010). https://doi.org/10.1007/978-3-642-17749-1_3
8. Bienvenu, M., Ortiz, M.: Ontology-mediated query answering with data-tractable description logics. In: Faber, W., Paschke, A. (eds.) Reasoning Web 11th International Summer School. Lecture Notes in Computer Science, vol. 9203, pp. 218–307. Springer (2015). https://doi.org/10.1007/978-3-319-21768-0_9
9. Bonomi, L., Jiang, X.: Patient ranking with temporally annotated data. Journal of Biomedical Informatics **78**, 43–53 (2018). https://doi.org/10.1016/j.jbi.2017.12.007
10. Calvanese, D., Cogrel, B., Komla-Ebri, S., Kontchakov, R., Lanti, D., Rezk, M., Rodriguez-Muro, M., Xiao, G.: Ontop: Answering SPARQL queries over relational databases. Semantic Web **8**, 471–487 (2017)
11. Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R.: Epistemic first-order queries over description logic knowledge bases. In: Parsia, B., Sattler, U.,

Toman, D. (eds.) Proc. of the 19th Int. Workshop on Description Logics (DL'06). CEUR Workshop Proceedings, vol. 189, pp. 51–61 (2006)

12. Cresswell, K.M., Sheikh, A.: Inpatient clinical information systems. In: Sheikh, A., Cresswell, K.M., Wright, A., Bates, D.W. (eds.) Key Advances in Clinical Informatics, chap. 2, pp. 13–29. Academic Press (2017). https://doi.org/10.1016/B978-0-12-809523-2.00002-9
13. Crowe, C.L., Tao, C.: Designing ontology-based patterns for the representation of the time-relevant eligibility criteria of clinical protocols. AMIA Joint Summits on Translational Science Proc. **2015**, 173–177 (2015), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4525239/
14. Deutsch, A., Nash, A., Remmel, J.B.: The chase revisited. In: Lenzerini, M., Lembo, D. (eds.) Proc. of the 27th ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems (PODS'08). pp. 149–158. ACM (2008). https://doi.org/10.1145/1376916.1376938
15. Eiter, T., Ortiz, M., Šimkus, M., Tran, T.K., Xiao, G.: Query rewriting for horn-$\mathcal{SHIQ}$ plus rules. In: Hoffmann, J., Selman, B. (eds.) Proc. of the 26th AAAI Conf. on Artificial Intelligence (AAAI'12). pp. 726–733. AAAI Press (2012), http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/4931
16. Fagin, R., Kolaitis, P.G., Miller, R.J., Popa, L.: Data exchange: Semantics and query answering. Theoretical Computer Science **336**(1), 89–124 (2005). https://doi.org/10.1016/j.tcs.2004.10.033
17. Gutiérrez-Basulto, V., Ibáñez-García, Y., Kontchakov, R., Kostylev, E.V.: Queries with negation and inequalities over lightweight ontologies. Journal of Web Semantics **35**, 184–202 (2015). https://doi.org/10.1016/j.websem.2015.06.002
18. Hernich, A., Kupke, C., Lukasiewicz, T., Gottlob, G.: Well-founded semantics for extended datalog and ontological reasoning. In: Hull, R., Fan, W. (eds.) Proc. of the 32nd Symp. on Principles of Database Systems (PODS'13). pp. 225–236. ACM (2013). https://doi.org/10.1145/2463664.2465229
19. Hripcsak, G., Zhou, L., Parsons, S., Das, A.K., Johnson, S.B.: Modeling electronic discharge summaries as a simple temporal constraint satisfaction problem. Journal of the American Medical Informatics Association **12**(1), 55–63 (2005). https://doi.org/10.1197/jamia.m1623
20. Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific Data **3**(160035), 1–9 (2016). https://doi.org/10.1038/sdata.2016.35
21. Kharlamov, E., Hovland, D., Skjæveland, M.G., Bilidas, D., Jiménez-Ruiz, E., Xiao, G., Soylu, A., Lanti, D., Rezk, M., Zheleznyakov, D., Giese, M., Lie, H., Ioannidis, Y., Kotidis, Y., Koubarakis, M., Waaler, A.: Ontology based data access in Statoil. J. Web Semantics **44**, 3–36 (2017). https://doi.org/10.1016/j.websem.2017.05.005
22. Kontchakov, R., Lutz, C., Toman, D., Wolter, F., Zakharyaschev, M.: The combined approach to ontology-based data access. In: Walsh, T. (ed.) Proc. of the 22nd Int. Joint Conf. on Artificial Intelligence (IJCAI'11). pp. 2656–2661. AAAI Press (2011). https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-442
23. Köpcke, F., Prokosch, H.U.: Employing computers for the recruitment into clinical trials: A comprehensive systematic review. Journal of Medical Internet Research **16**(7), e161 (2014). https://doi.org/10.2196/jmir.3446
24. Lutz, C., Seylan, I., Wolter, F.: Ontology-based data access with closed predicates is inherently intractable (sometimes). In: Rossi, F. (ed.) Proc. of the 23rd Int. Joint Conf. on Artificial Intelligence (IJCAI'13). pp. 1024–1030. AAAI Press (2013), https://www.ijcai.org/Abstract/13/156

25. Lutz, C., Toman, D., Wolter, F.: Conjunctive query answering in the description logic $\mathcal{EL}$ using a relational database system. In: Boutilier, C. (ed.) Proc. of the 21st Int. Joint Conf. on Artificial Intelligence (IJCAI'09). pp. 2070–2075. AAAI Press (2009)
26. Marnette, B.: Generalized schema mappings: From termination to tractability. In: Paredaens, J., Su, J. (eds.) Proc. of the 28th Symp. on Principles of Database Systems (PODS'09). pp. 13–22. ACM (2009). https://doi.org/10.1145/1559795.1559799
27. Mugnier, M.L., Thomazo, M.: An introduction to ontology-based query answering with existential rules. In: Reasoning Web International Summer School. pp. 245–278 (2014). https://doi.org/10.1007/978-3-319-10587-1_6
28. Ni, Y., Wright, J., Perentesis, J., Lingren, T., Deleger, L., Kaiser, M., Kohane, I., Solti, I.: Increasing the efficiency of trial-patient matching: Automated clinical trial eligibility pre-screening for pediatric oncology patients. BMC Medical Informatics and Decision Making **15**, 1–10 (2015). https://doi.org/10.1186/s12911-015-0149-3
29. Patel, C., Cimino, J., Dolby, J., Fokoue, A., Kalyanpur, A., Kershenbaum, A., Ma, L., Schonberg, E., Srinivas, K.: Matching patient records to clinical trials using ontologies. In: Aberer, K., Choi, K.S., Noy, N., Allemang, D., Lee, K.I., Nixon, L., Goldbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) Proc. of the 6th Int. Semantic Web Conf. (ISWC'07). Lecture Notes in Computer Science, vol. 4825, pp. 816–829. Springer (2007). https://doi.org/10.1007/978-3-540-76298-0_59
30. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. Journal of the American Medical Informatics Association **17**(5), 507–513 (2010)
31. Tagaris, A., Andronikou, V., Chondrogiannis, E., Tsatsaronis, G., Schroeder, M., Varvarigou, T., Koutsouris, D.D.: Exploiting Ontology Based Search and EHR Interoperability to Facilitate Clinical Trial Design, pp. 21–42. Springer (2014). https://doi.org/10.1007/978-3-319-06844-2_3
32. Wolter, F.: First order common knowledge logics. Studia Logica **65**(2), 249–271 (2000). https://doi.org/10.1023/A:1005271815356

A Proof of Lemma 10

We prove the two set inclusions separately.

Lemma 12. *Let $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ be a consistent $\mathcal{ELH}_\bot$ KB and let $\phi(\mathbf{x})$ be a rooted NCQ. Then, for all $\phi' \in \mathrm{rew}_\mathcal{T}(\phi)$,*

$$\mathrm{ans}(\phi', I_\mathcal{A}) \subseteq \mathrm{mwa}(\phi, I_\mathcal{K}).$$

Proof. By Lemma 9, we have $\mathrm{ans}(\phi', I_\mathcal{A}) \subseteq \mathrm{mwa}(\phi', \mathcal{K}) = \mathrm{ans}(\phi', I_\mathcal{K})$. Furthermore, there exists a sequence $\phi_0 \to_\mathcal{T} \cdots \to_\mathcal{T} \phi_n$ $(n > 0)$ with $\phi = \phi_0$ and $\phi' = \phi_n$. Hence it is sufficient to show that $\mathrm{ans}(\phi_i, I_\mathcal{K}) \subseteq \mathrm{ans}(\phi_{i-1}, I_\mathcal{K})$ for all $i, 1 \leq i \leq n$. Suppose the queries are of the following forms:

$$\phi_i = \exists \mathbf{y}_i.(\varphi_i(\mathbf{x}_i, \mathbf{y}_i) \wedge \Psi_i) \tag{2}$$

$$\phi_{i-1} = \exists \mathbf{y}_{i-1}.(\varphi_{i-1}(\mathbf{x}_{i-1}, \mathbf{y}_{i-1}) \wedge \Psi_{i-1}) \tag{3}$$

Let π_i be a satisfying assignment for $\varphi_i(\mathbf{x}_i, \mathbf{y}_i) \wedge \Psi_i$ in $I_\mathcal{K}$. Suppose $\phi_{i-1} \to_\mathcal{T} \phi_i$ by

1. selecting variable $\hat{x}$ and introducing $\hat{y}$ in (S1) and
2. selecting $M \sqsubseteq_\mathcal{T} \exists s.N$ in (S2).

Let $\pi_i(\hat{y}) = d$. By Step (S6), $M(\hat{y}) \in \varphi_i$ and since π_i satisfies φ_i, it has to hold that $d \in M^{I_\mathcal{K}}$. This implies that $d \in (\exists s.N)^{I_\mathcal{K}}$. Since π_i satisfies the new filter $\psi_i^*(\hat{y})$ that is constructed in (S7), and by (S2) the precondition of $\psi_i^*(\hat{y})$ is satisfied by π_i in $I_\mathcal{K}$, there has to be an assignment $\pi_i \cup \{\hat{x} \mapsto d'\}$ that satisfies the conclusion of $\psi_i^*(\hat{y})$.

We define the assignment π_{i-1} of the variables of φ_{i-1} as follows

$$\pi_{i-1}(z) := \begin{cases} d' & \text{if } z = \hat{x} \\ d & \text{if } z \in \mathsf{Pred} \\ \pi_i(z) & \text{otherwise.} \end{cases} \tag{4}$$

Then π_{i-1} is a satisfying assignment for ϕ_{i-1} in $I_\mathcal{K}$. To see this, first consider an atom α in φ_{i-1}. We show that π_{i-1} satisfies α in $I_\mathcal{K}$.

If α contains $\hat{x}$, it can be of the following forms: $A(\hat{x})$, $\neg A(\hat{x})$, $r(y, \hat{x})$ or $\neg r(y, \hat{x})$ with $y \in \mathsf{Pred}$. For all of these cases, we know by Step (S7) that they are either implied by $s(\hat{y}, \hat{x}) \wedge N(\hat{x})$ or contained in Neg, with y replaced by $\hat{y}$. By the choice of d', we know that π_{i-1} satisfies each such atom.

If α does not contain $\hat{x}$, then φ_i contains the atom α' that is obtained from α by replacing all of the variables from Pred with $\hat{y}$. By construction, we know that $\pi_{i-1}(y) = \pi_i(\hat{y})$ for all $y \in \mathsf{Pred}$ and $\pi_{i-1}(z) = \pi_i(z)$ otherwise. Since α' is satisfied by π_i in $I_\mathcal{K}$, α is satisfied by π_{i-1} in $I_\mathcal{K}$.

What remains to show is that π_{i-1} satisfies Ψ_{i-1}. Consider any $\psi(z) \in \Psi_{i-1}$, and distinguish the following cases:

1. If $z = \hat{x}$, then $\psi(\hat{x}) \in \Psi_{\hat{x}}$. Since $I_\mathcal{K}, \pi_i \cup \{\hat{x} \mapsto d'\} \models \Psi_{\hat{x}}$, we also have $I_\mathcal{K}, \{\hat{x} \mapsto d'\} \models \psi(\hat{x})$. Therefore, since $\pi_{i-1}(\hat{x}) = d'$, it holds that π_{i-1} satisfies $\psi(\hat{x})$ in $I_\mathcal{K}$.
2. If $z \in \mathsf{Pred}$ we know that $\pi_{i-1}(z) = \pi_i(\hat{y}) = d$. Since $I_\mathcal{K}, \pi_i \models \psi(\hat{y})$, it also holds that $I_\mathcal{K}, \pi_{i-1} \models \psi(z)$.
3. Otherwise the filter is present in Ψ_i. Then we know that $I_\mathcal{K}, \pi_i \models \psi(z)$ and $\pi_i(z) = \pi_{i-1}(z)$. Hence, it must also hold that $I_\mathcal{K}, \pi_{i-1} \models \psi(z)$. □

Lemma 13. *Let $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ be a consistent $\mathcal{ELH}_\bot$ KB and let $\phi(\mathbf{x})$ be a rooted NCQ. Then*

$$\mathrm{mwa}(\phi, \mathcal{K}) \subseteq \bigcup_{\phi' \in \mathrm{rew}_\mathcal{T}(\phi)} \mathrm{ans}(\phi', I_\mathcal{A}).$$

Proof. Suppose $\mathbf{a} \in \mathrm{mwa}(\phi, \mathcal{K}) = \mathrm{ans}(\phi, I_\mathcal{K})$. We have to show that there exists $\phi' \in \mathrm{rew}_\mathcal{T}(\phi)$ and a satisfying assignment π for ϕ' in $I_\mathcal{A}$ such that $\mathbf{a} = \pi(\mathbf{x})$. To do this, we assign a *degree* (a natural number) to each satisfying assignment (including the existentially quantified variables of the NCQ part) such that a satisfying assignment with degree 0 does not use any anonymous individuals. We then show that for each satisfying assignment with a degree greater than 0, we can find a rewriting for which a satisfying assignment yielding the same answer, but with a lower degree, exists. In addition, for every such assignment π and for all filters $\psi(y)$ in ϕ' it should hold that,

$$\text{if } \pi(y) \in N_I, \text{ then } I_{\mathcal{A}} \models \psi(\pi(y)), \tag{†}$$

i.e. all filters (at any stage of the rewriting) are satisfied within the confines of $I_{\mathcal{A}}$.

For any element $d \in \Delta^{I_{\mathcal{K}}}$, we denote by $|d|$ the minimal number of role connections required to reach d from an element in N_I, with $|d| = 0$ iff $d \in N_I$. Additionally, for any assignment π' in $I_{\mathcal{K}}$, let

$$\deg(\pi') := \sum_{y \in \mathrm{dom}(\pi')} |\pi'(y)|. \tag{5}$$

Since $\phi \in \mathrm{rew}_{\mathcal{T}}(\phi)$, to prove the claim it suffices to show that whenever there is $\phi_1 = \exists \mathbf{y}.\varphi_1(\mathbf{x}, \mathbf{y}) \wedge \Psi \in \mathrm{rew}_{\mathcal{T}}(\phi)$ such that φ_1 has a match π_1 in $I_{\mathcal{K}}$ with $\mathbf{a} = \pi_1(\mathbf{x})$, $\deg(\pi_1) > 0$, and Equation (†) holds for π_1 and the filters in Ψ, then there exist ϕ_2 and π_2 with the same properties, but $\deg(\pi_2) < \deg(\pi_1)$.

Assume $\phi_1 \in \mathrm{rew}_{\mathcal{T}}(\phi)$ as above, and let π_1 be a match of φ_1. Since $\deg(\pi_1) > 0$ by assumption, there must exist a variable $\hat{x}$ of φ_1 such that $\pi_1(\hat{x}) \notin N_I$. Select $\hat{x}$ such that it is a leaf node in the subforest of $I_{\mathcal{K}}$ induced by π_1. Note that $\hat{x}$ cannot be an answer variable.

We know that $\pi_1(\hat{x}) = d_{\hat{x}}$ was induced by some axiom $\alpha = M \sqsubseteq_{\mathcal{T}} \exists s.N$ and element $d_p \in M^{I_{\mathcal{K}}}$ in Definition 4. By the construction of $I_{\mathcal{K}}$, we know that

(i) $d_{\hat{x}}$ has just the one predecessor d_p, and
(ii) $d_{\hat{x}} \in A^{I_{\mathcal{K}}}$ iff $N \sqsubseteq_{\mathcal{T}} A$ and $(d_p, d_{\hat{x}}) \in r^{I_{\mathcal{K}}}$ iff $s \sqsubseteq_{\mathcal{T}} r$.

We obtain the query ϕ_2 from ϕ_1 through rewriting, by selecting $\hat{x}$ and introducing $\hat{y}$ in (S1), and selecting α in (S2). Let Pred denote the set of predecessor variables of $\hat{x}$ as defined in (S1). To see that this is a valid choice, the conditions in (S2) need to be verified:

(S2a) For any $A(\hat{x}) \in \varphi_1$, we have $d_{\hat{x}} = \pi_1(\hat{x}) \in A^{I_{\mathcal{K}}}$, and hence $N \sqsubseteq_{\mathcal{T}} A$ by (ii). Consider any role atom $r(y, \hat{x}) \in \varphi_1$. From (i), the construction of $I_{\mathcal{K}}$ (no inverse edges), and the fact that π_1 is a satisfying assignment for $r(y, \hat{x})$ in $I_{\mathcal{K}}$, the only possibility is that $\pi_1(y) = d_p$. Therefore $(d_p, d_{\hat{x}}) = (\pi_1(y), \pi_1(\hat{x})) \in r^{I_{\mathcal{K}}}$. By (ii), this implies that $s \sqsubseteq_{\mathcal{T}} r$.

(S2b) Consider any $\neg A(\hat{x}) \in \varphi_1$, for which we must have $d_{\hat{x}} \notin A^{I_{\mathcal{K}}}$. From (ii) we know that $N \not\sqsubseteq_{\mathcal{T}} A$. Consider any $\neg r(y, \hat{x}) \in \varphi_1$. Since this is guarded by a positive role atom as above, again the only possibility is that $\pi_1(y) = d_p$. Hence $(d_p, d_{\hat{x}}) \notin r^{I_{\mathcal{K}}}$. By (ii), this implies that $s \not\sqsubseteq_{\mathcal{T}} r$.

We obtain a satisfying assignment π_2 for ϕ_2 in $I_{\mathcal{K}}$ such that $\mathbf{a} \in \pi_2(\mathbf{x})$ (and $\deg(\pi_2) < \deg(\pi_1)$) by setting for all $z \in \mathrm{Var}(\varphi_2)$:

$$\pi_2(z) := \begin{cases} \pi_1(z) & \text{if } z \in \mathrm{Var}(\varphi_1) \\ d_p & \text{if } z = \hat{y}. \end{cases}$$

To see that π_2 satisfies ϕ_2, we argue why it satisfies the new atoms and filter from (S6) and (S7); the old atoms (possibly with renamed variables) remain satisfied.

The new atom $M(\hat{y})$ is satisfied since $\pi_2(\hat{y}) = d_p \in M^{I_\mathcal{K}}$. Consider now an atom $\neg M'(\hat{y})$ with $M' \in \mathcal{M}'$ as specified in (S6); we have to show that $d_p \notin (M')^{I_\mathcal{K}}$. Assume to the contrary that $d_p \in (M')^{I_\mathcal{K}}$. By (S3), we know that $M' \sqsubseteq_\mathcal{T} \exists s'.N' \sqsubseteq^s_\mathcal{T} \exists s.N$. Moreover, $\exists s'.N'$ must be included in the set V in Step 3(a) of Definition 4, because otherwise we would already have $d_p \in (\exists s'.N')^{I_\mathcal{A}}$, i.e. there would be a named individual b such that $(d_p, b) \in (s')^{I_\mathcal{A}}$ and $b \in (N')^{I_\mathcal{A}}$. Since $s' \sqsubseteq_\mathcal{T} s$ and $N' \sqsubseteq_\mathcal{T} N$, this would imply $(d_p, b) \in s^{I_\mathcal{A}}$ and $b \in N^{I_\mathcal{A}}$, i.e. $d_p \in (\exists s.N)^{I_\mathcal{A}}$, which shows that the anonymous object $d_{\hat{x}}$ would not have been created. Since $\exists s'.N'$ is included in V and we assumed that $\exists s.N$ is minimal w.r.t. $\sqsubseteq^s_\mathcal{T}$, we must have $s \equiv_\mathcal{T} s'$ and $N \equiv_\mathcal{T} N'$. But then (S3b) directly contradicts (S2b).

We now consider the filters in ϕ_2. Suppose that Equation (†) holds for π_1 and all filters in ϕ_1. For the ones that are only copied from ϕ_1 (modulo renaming some variables to $\hat{y}$), the property is clearly preserved. For the new filter $\psi^*(\hat{y})$, assume that $\pi_2(\hat{y}) \in N_I$, and hence we need to show that $I_\mathcal{A} \models \pi_2(\psi^*(\hat{y}))$. Assume that there exists an element $d' \in N_I$ such that $(d_p, d') \in s^{I_\mathcal{A}}$ and $d' \in N^{I_\mathcal{A}}$. But then in Step 3(a) in Definition 4, $\exists s.N$ could not have been added to V since $d \in (\exists s.N)^{I_\mathcal{K}}$ already holds. Hence, the element $d_{\hat{x}}$ would have never been introduced, which is a contradiction. Therefore, in $I_\mathcal{A}$ the precondition of $\psi^*(\hat{y})$ is never met, which makes the filter trivially satisfied.

Finally, observe that $\mathrm{Var}(\varphi_2) \setminus \{\hat{y}\} \subset \mathrm{Var}(\varphi_1) \setminus \{\hat{x}\}$. From the facts that $|\pi_2(z)| = |\pi_1(z)|$ for all $z \in \mathrm{Var}(\varphi_2) \setminus \{\hat{y}\}$, $|\pi_2(\hat{y})| = |\pi_1(y)|$ for all $y \in \mathsf{Pred}$, $\mathsf{Pred} \neq \emptyset$ since φ_1 is connected, and $|\pi_1(\hat{x})| \geq 1$, we obtain that $\deg(\pi_2) < \deg(\pi_1)$. □