Towards Privacy-Preserving Ontology Publishing

Franz Baader, Adrian Nuradiansyah[⋆]
firstname.lastname@tu-dresden.de

Theoretical Computer Science, TU Dresden

Abstract. We make a first step towards adapting the approach of Cuenca Grau and Kostylev for privacy-preserving publishing of linked data to Description Logic ontologies. We consider the case where both the knowledge about individuals and the privacy policies are expressed using $\mathcal{EL}$ concepts. We introduce the notions of compliance of a concept with a policy and of safety of a concept for a policy, and show how optimal compliant (safe) generalizations of a given $\mathcal{EL}$ concept can be computed.

1 Introduction

When publishing information about individuals, one needs to ensure that certain privacy constraints are fulfilled. These constraints are encoded as *privacy policies*, and before publishing the information one needs to check whether the information is *compliant* with these policies [6]. We illustrate this setting using an example from [6]: when publishing information about hospitals, doctors, and patients, the policy may require that one should not be able to find out who are the cancer patients. In case the information to be published is not policy compliant, it first needs to be modified in a minimal way to make it compliant. However, compliance per se is not enough if a possible attacker can also obtain relevant information from other sources, which together with the published information might violate the privacy policy. *Safety* requires that the combination of the published information with any other compliant information is again compliant [6]. More information on privacy-preserving data publishing can be found in the survey [7].

In [6], this problem was investigated in a setting where the information to be published is given as a relational dataset with (labeled) null values, and the policy is given by a conjunctive query. In order to make a given dataset compliant or safe, one is basically allowed to replace constants (or null values) by new null values. The paper investigates the complexity of deciding compliance (Is a given modification of a dataset policy compliant?), safety (Is a given modification of a dataset safe w.r.t. a policy?), and optimality (Is a given modification of a dataset safe w.r.t. a policy and changes the dataset in a minimal way?). The obtained complexity results depend on whether combined or data complexity is considered, and whether closed- or open-world semantics are used. The paper

[⋆] Funded by DFG within the Research Training Group 1907 “RoSI”

does not consider the case where the information in the dataset is augmented by ontological knowledge.

In the present paper, we make a first step towards handling ontologies in this context, but consider a quite restricted setting, where information about an individual is given by a concept of the inexpressive Description Logic (DL) $\mathcal{EL}$. Basically, this is the setting where the ontology consists of an ABox containing only concept assertions of the form $C(a)$ for possibly complex concepts C, but no role assertion. In [8], such an ABox was called an instance store. In addition, we assume that there is no TBox, i.e., all the information about the individual a is given by the concept C.[1] A policy is then given by an instance query, i.e., by an $\mathcal{EL}$ concept D. A concept C (giving information about some individual a) is *compliant* with this policy, if it is not subsumed by D, i.e., if $C(a)$ does not imply $D(a)$. In our example, the policy could be formalized as the $\mathcal{EL}$ concept

$$D = \mathit{Patient} \sqcap \exists \mathit{seen\_by}.(\mathit{Doctor} \sqcap \exists \mathit{works\_in}.\mathit{Oncology}),$$

which says that one should not be able to find out who are the patients that are seen by a doctor that works for the oncology department. The concept

$$C = \mathit{Patient} \sqcap \mathit{Male} \sqcap \exists \mathit{seen\_by}.(\mathit{Doctor} \sqcap \mathit{Female} \sqcap \exists \mathit{works\_in}.\mathit{Oncology})$$

is not compliant with the policy D since $C \sqsubseteq D$. The concept

$$C' = \mathit{Male} \sqcap \exists \mathit{seen\_by}.(\mathit{Doctor} \sqcap \mathit{Female} \sqcap \exists \mathit{works\_in}.\mathit{Oncology})$$

is a compliant generalization of C, i.e., $C \sqsubseteq C'$ and $C' \not\sqsubseteq D$. However, it is not safe since $C' \sqcap \mathit{Patient} \sqsubseteq D$, i.e., if the attacker already knows that a is a patient then together with $C'(a)$ the hidden information D is revealed. In contrast,

$$C'' = \mathit{Male} \sqcap \exists \mathit{seen\_by}.(\mathit{Doctor} \sqcap \mathit{Female} \sqcap \exists \mathit{works\_in}.\top),$$

is a safe generalization of C, though it is less obvious to see this. This concept is, however, not optimal since more information than necessary is removed. In fact, the concept

$$\begin{aligned} C''' = \mathit{Male} \sqcap\ & \exists \mathit{seen\_by}.(\mathit{Doctor} \sqcap \mathit{Female} \sqcap \exists \mathit{works\_in}.\top) \\ & \exists \mathit{seen\_by}.(\mathit{Female} \sqcap \exists \mathit{works\_in}.\mathit{Oncology}) \end{aligned}$$

is a safe generalization of C that is more specific than C'', i.e. $C \sqsubseteq C''' \sqsubset C''$.

In this paper, we will show how to compute optimal compliant and optimal safe generalizations of $\mathcal{EL}$ concepts C with $\mathcal{EL}$ policies, but instead of only one policy concept we allow for a finite set of $\mathcal{EL}$ concepts as policy, where a concept C' is compliant with the policy $\{D_1, \ldots, D_p\}$ iff it is compliant with each element of this set, i.e., $C \not\sqsubseteq D_i$ holds for all $i = 1, \ldots, p$. But first, we need to introduce the DL $\mathcal{EL}$ and the notions of compliance, safety, etc. in a more formal way.

[1] Since $\mathcal{EL}$ concepts are closed under conjunction, we can assume that the ABox contains only one assertion for a.

2 Preliminaries

A wide range of DLs of different expressive power haven been investigated in the literature [2]. Here, we only introduce the DL $\mathcal{EL}$, for which reasoning is tractable [3,5,1]. Let N_C and N_R be mutually disjoint sets of *concept* and *role names*, respectively. Then $\mathcal{EL}$ *concepts* over these names are constructed from concept names using the constructors top concept ($\top$), conjunction ($C \sqcap D$), and existential restriction ($\exists r.C$). The *size* of an $\mathcal{EL}$ concept C is the number of occurrences of $\top$ as well as concept and role names in C, and its *role depth* is the maximal nesting of existential restrictions.

The semantics of $\mathcal{EL}$ is defined through *interpretations* $\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})$, where $\Delta^\mathcal{I}$ is a non-empty set, called the *domain*, and $\cdot^\mathcal{I}$ is the *interpretation function*, which maps every $A \in N_C$ to a set $A^\mathcal{I} \subseteq \Delta^\mathcal{I}$ and every $r \in N_R$ to a binary relation $r^\mathcal{I} \subseteq \Delta^\mathcal{I} \times \Delta^\mathcal{I}$. This function $\cdot^\mathcal{I}$ is extended to arbitrary $\mathcal{EL}$ concepts by setting $\top^\mathcal{I} := \Delta^\mathcal{I}$, $(C \sqcap D)^\mathcal{I} := C^\mathcal{I} \cap D^\mathcal{I}$, and $(\exists r.C)^\mathcal{I} := \{\delta \in \Delta^\mathcal{I} \mid \exists \eta \in C^\mathcal{I}.(\delta, \eta) \in r^\mathcal{I}\}$.

The $\mathcal{EL}$ concept C is *subsumed by* the $\mathcal{EL}$ concept D (written $C \sqsubseteq D$) if $C^\mathcal{I} \subseteq D^\mathcal{I}$ holds for all interpretations $\mathcal{I}$. Strict subsumption (written $C \sqsubset D$) holds if $C \sqsubseteq D$ and $D \not\sqsubseteq C$, and we say that C is *equivalent* to D (written $C \equiv D$) if $C \sqsubseteq D$ and $D \sqsubseteq C$. Subsumption between $\mathcal{EL}$ concepts can be decided in polynomial time [3]. The following recursive characterization of subsumption in $\mathcal{EL}$ was shown in [4].

Proposition 1. *Let*

$$\begin{aligned} C &= A_1 \sqcap \ldots \sqcap A_k \sqcap \exists r_1.C_1 \sqcap \ldots \sqcap \exists r_m.C_m, \textit{ and} \\ D &= B_1 \sqcap \ldots \sqcap B_\ell \sqcap \exists s_1.D_1 \sqcap \ldots \sqcap \exists s_n.D_n, \end{aligned}$$

where $A_1, \ldots, A_k, B_1, \ldots, B_\ell \in N_C$ *and* $r_1, \ldots, r_m, s_1, \ldots, s_n \in N_R$. *Then we have* $C \sqsubseteq D$ *iff*

- $\{B_1, \ldots, B_\ell\} \subseteq \{A_1, \ldots, A_k\}$, *and*
- *for all* $i \in \{1, \ldots, n\}$ *there is* $j \in \{1, \ldots, m\}$ *such that* $s_i = r_j$ *and* $C_j \sqsubseteq D_i$.

We are now ready to define the important notions regarding privacy-preserving publishing of ontological information that will be investigated in this paper. As mentioned in the introduction, policies are finite sets of $\mathcal{EL}$ concepts. We assume in the following, that the concepts occurring in the policy are not equivalent to top since otherwise there would not be compliant concepts.

Definition 1. *A* policy *is a finite set* $\mathcal{P} = \{D_1, \ldots, D_p\}$ *of* $\mathcal{EL}$ *concepts such that* $\top \not\equiv D_i$ *for* $i = 1, \ldots, p$. *Given an* $\mathcal{EL}$ *concept* C *and a policy* $\mathcal{P} = \{D_1, \ldots, D_p\}$, *the* $\mathcal{EL}$ *concept* C' *is*

- compliant *with* $\mathcal{P}$ *if* $C' \not\sqsubseteq D_i$ *holds for all* $i = 1, \ldots, p$;
- safe *for* $\mathcal{P}$ *if* $C' \sqcap C''$ *is compliant with* $\mathcal{P}$ *for all* $\mathcal{EL}$ *concepts* C'' *that are compliant with* $\mathcal{P}$;
- *a* $\mathcal{P}$-compliant generalization *of* C *if* $C \sqsubseteq C'$ *and* C' *is compliant with* $\mathcal{P}$;

- *an* optimal $\mathcal{P}$-compliant generalization *of C if it is a $\mathcal{P}$-compliant generalization of C and there is no $\mathcal{P}$-compliant generalization C'' of C such that $C'' \sqsubset C'$;*
- *a $\mathcal{P}$*-safe generalization *of C if $C \sqsubseteq C'$ and C' is safe for $\mathcal{P}$;*
- *an* optimal $\mathcal{P}$-safe generalization *of C if it is a $\mathcal{P}$-safe generalization of C and there is no $\mathcal{P}$-safe generalization C'' of C such that $C'' \sqsubset C'$.*

It is easy to see that safety implies compliance since the top concept is always compliant: if C' is safe for $\mathcal{P}$, then $\top \sqcap C' \equiv C'$ is compliant.

3 Characterizing compliance

In this section, we characterize the concepts that are compliant with a given policy $\mathcal{P}$, and use this to develop an algorithm that computes all optimal $\mathcal{P}$-compliant generalizations of a given $\mathcal{EL}$ concept C.

But first, we need to introduce some more notations. We call an $\mathcal{EL}$ concept an *atom* if it is a concept name or an existential restriction. Given an $\mathcal{EL}$ concept C, we denote the set of atoms occurring in its top-level conjunction with $\mathtt{con}(C)$. For example, if $C = A \sqcap \exists r.(B \sqcap \exists s.A)$, then $\mathtt{con}(C) = \{A, \exists r.(B \sqcap \exists s.A)\}$. As a special case of Proposition 1, subsumption between atoms can be characterized as follows. If E, F are atoms, then $E \sqsubseteq F$ iff

- $E = F \in N_C$ or
- E, F are existential restrictions of the form $E = \exists r.E', F = \exists r.F'$ such that $E' \sqsubseteq F'$.

Definition 2. *Let S,T be sets of atoms. Then we say that S* covers *T if for every $F \in T$ there is $E \in S$ such that $E \sqsubseteq F$.*

With this notation, Proposition 1 can be reformulated as follows: $C \sqsubseteq D$ iff $\mathtt{con}(C)$ covers $\mathtt{con}(D)$. The following (polynomial-time decidable) characterization of compliance is thus an immediate consequence of Proposition 1.

Proposition 2. *The $\mathcal{EL}$ concept C' is compliant with the policy $\mathcal{P} = \{D_1, \ldots, D_p\}$ iff $\mathtt{con}(C')$ does not cover $\mathtt{con}(D_i)$ for any $i = 1, \ldots, p$, i.e., for every $i = 1, \ldots, p$, at least one of the following two properties holds:*

- *there is a concept name $A \in \mathtt{con}(D_i)$ such that $A \notin \mathtt{con}(C')$; or*
- *there is an existential restriction $\exists r.D \in \mathtt{con}(D_i)$ such that $C \not\sqsubseteq D$ for all existential restrictions of the form $\exists r.C \in \mathtt{con}(C')$.*

Now assume that we are given an $\mathcal{EL}$ concept C and a policy $\mathcal{P} = \{D_1, \ldots, D_p\}$, and we want to construct a $\mathcal{P}$-compliant generalization C' of C. For C' to satisfy the condition of Proposition 2, there needs to exist for every $i = 1, \ldots, p$ an element of $\mathtt{con}(D_i)$ that is not covered by any element of $\mathtt{con}(C')$. In case $\mathtt{con}(C)$ contains elements covering such an atom, we need to remove or generalize them appropriately.

Definition 3. *We say that* $H \subseteq \mathtt{con}(D_1) \cup \ldots \cup \mathtt{con}(D_p)$ *is a* hitting set *of* $\mathtt{con}(D_1), \ldots, \mathtt{con}(D_p)$ *if* $H \cap \mathtt{con}(D_i) \neq \emptyset$ *for every* $i = 1, \ldots, p$*. This hitting set is* minimal *if there is no other hitting set strictly contained in it.*

Basically, the idea is now to choose a hitting set H of $\mathtt{con}(D_1), \ldots, \mathtt{con}(D_p)$ and use H to guide the construction of a compliant generalization of C. In order to make this generalization as specific as possible, we use minimal hitting sets. In case the policy contains concepts D_i with which C is already compliant (i.e., $C \not\sqsubseteq D_i$ holds), nothing needs to be done w.r.t. these concepts. This is why, in the following definition, $\mathtt{con}(D_i)$ does not take part in the construction of the hitting set if $C \not\sqsubseteq D_i$.

Definition 4. *Let* C *be an* $\mathcal{EL}$*-concept and* $\mathcal{P} = \{D_1, \ldots, D_p\}$ *a policy. The set* $SCG(C, \mathcal{P})$ *of* specific compliant generalizations *of* C *w.r.t.* $\mathcal{P}$ *consists of the concepts that can be constructed from* C *as follows:*

- *If* C *is compliant with* $\mathcal{P}$*, then* $SCG(C, \mathcal{P}) = \{C\}$*.*
- *Otherwise, choose a minimal hitting set* H *of* $\mathtt{con}(D_{i_1}), \ldots, \mathtt{con}(D_{i_q})$ *where* $i_1, \ldots, i_q$ *are exactly the indices for which* $C \sqsubseteq D_i$*. Note that* $q \geq 1$ *since we are in the case where* C *is not compliant with* $\mathcal{P}$*. In addition, according to our definition of a policy, none of the concepts* D_i *is equivalent to* $\top$*, and thus the sets* $\mathtt{con}(D_{i_j})$ *are non-empty. Consequently, at least one minimal hitting set exists. Each minimal hitting set yields a concept in* $SCG(C, \mathcal{P})$ *by removing or modifying atoms in the top-level conjunction of* C *in the following way:*
  - *For every concept name* $A \in \mathtt{con}(C)$*, remove* A *from the top-level conjunction of* C *if* $A \in H$*;*
  - *For every existential restriction* $\exists r_i.C_i \in \mathtt{con}(C)$*, consider the set*

    $$\mathcal{P}_i := \{G \mid \text{ there is } \exists r_i.G \in H \text{ such that } C_i \sqsubseteq G\}.$$

    - *If* $\mathcal{P}_i = \emptyset$ *then leave* $\exists r_i.C_i$ *as it is.*
    - *If* $\top \in \mathcal{P}_i$*, then remove* $\exists r_i.C_i$*.*
    - *Otherwise, replace* $\exists r_i.C_i$ *with* $\bigsqcap_{F \in SCG(C_i, \mathcal{P}_i)} \exists r_i.F$*.*

We will show below that every element of $SCG(C, \mathcal{P})$ is a compliant generalization of C, and that all optimal compliant generalizations of C belong to $SCG(C, \mathcal{P})$. However, $SCG(C, \mathcal{P})$ may also contain compliant generalizations of C that are not optimal, as illustrated by the following example.

Example 1. Let $C = \exists r.(A_1 \sqcap A_2 \sqcap A_3 \sqcap A_4)$ and $\mathcal{P} = \{D_1, D_2\}$, where

$$D_1 = \exists r.A_1 \sqcap \exists r.(A_2 \sqcap A_3) \quad \text{and} \quad D_2 = \exists r.A_2 \sqcap \exists r.A_4.$$

We have $C \sqsubseteq D_1$ and $C \sqsubseteq D_2$, and thus C is not compliant with $\mathcal{P}$. Consequently, the elements of $SCG(C, \mathcal{P})$ are obtained by considering the minimal hitting sets of $\{\exists r.A_1, \exists r.(A_2 \sqcap A_3)\}$ and $\{\exists r.A_2, \exists r.A_4\}$.

If we take the minimal hitting set $H = \{\exists r.(A_2 \sqcap A_3), \exists r.A_2\}$ and consider the only existential restriction in $\mathtt{con}(C)$, the corresponding set $\mathcal{P}_i$ consists of

$A_2 \sqcap A_3$ and A_2. It is easy to see that $SCG(A_1 \sqcap A_2 \sqcap A_3 \sqcap A_4, \mathcal{P}_i) = \{A_1 \sqcap A_3 \sqcap A_4\}$ since the only minimal hitting set of $\{A_1, A_2\}$ and $\{A_2\}$ is $\{A_2\}$. Thus, we obtain $C' := \exists r.(A_1 \sqcap A_3 \sqcap A_4)$ as an element of $SCG(C, \mathcal{P})$.

However, if we take the minimal hitting set $H' = \{\exists r.A_1, \exists r.A_2\}$ instead, then the set $\mathcal{P}'_i$ corresponding to the only existential restriction in $\texttt{con}(C)$ is $\{A_1, A_2\}$. Consequently, in this case $SCG(A_1 \sqcap A_2 \sqcap A_3 \sqcap A_4, \mathcal{P}'_i) = \{A_3 \sqcap A_4\}$ since the only minimal hitting set of $\{A_1\}$ and $\{A_2\}$ is $\{A_1, A_2\}$. This yields $C'' := \exists r.(A_3 \sqcap A_4)$ as another element of $SCG(C, \mathcal{P})$. Since $C' \sqsubset C''$, the element C'' cannot be optimal.

Next, we show that the elements of $SCG(C, \mathcal{P})$ are compliant generalizations of C.

Proposition 3. *Let C be an $\mathcal{EL}$-concept and $\mathcal{P} = \{D_1, \ldots, D_p\}$ a policy. If $C' \in SCG(C, \mathcal{P})$, then C' is a $\mathcal{P}$-compliant generalization of C.*

Proof. In case C is already compliant with $\mathcal{P}$, then $C = C'$ and we are done. Thus, assume that C is not compliant with $\mathcal{P}$. We show that C' is a compliant generalization of C by induction on the role depth of C.

First, we show that C' is a generalization of C, i.e., $C \sqsubseteq C'$. This is an easy consequence of the fact that, when constructing C' from C, atoms from the top-level conjunction of C are left unchanged, are removed, or are replaced by a conjunction of more general atoms. The only non-trivial case is where we replace an existential restriction $\exists r_i.C_i$ with the conjunction $\bigsqcap_{F \in SCG(C_i, \mathcal{P}_i)} \exists r_i.F$. By induction, we know that $C_i \sqsubseteq F$ for all $F \in SCG(C_i, \mathcal{P}_i)$, and thus $\exists r_i.C_i \sqsubseteq \bigsqcap_{F \in SCG(C_i, \mathcal{P}_i)} \exists r_i.F$.

Second, we show that C' is compliant with $\mathcal{P}$, i.e., $C' \not\sqsubseteq D_i$ holds for $i = 1, \ldots, p$. For the indices i with $C \not\sqsubseteq D_i$, we clearly also have $C' \not\sqsubseteq D_i$ since $C \sqsubseteq C'$. Now, consider one of the remaining indices $i_j \in \{i_1, \ldots, i_q\}$, where $i_1, \ldots, i_q$ are exactly the indices for which $C \sqsubseteq D_i$. The concept C' was constructed by taking some minimal hitting set H of $\texttt{con}(D_{i_1}), \ldots, \texttt{con}(D_{i_q})$. If the element in H hitting $\texttt{con}(D_{i_j})$ is a concept name, then this concept name does not occur in $\texttt{con}(C')$, and thus $C' \not\sqsubseteq D_{i_j}$. Thus, assume that it is an existential restriction $\exists r_i.G$. But then each existential restriction $\exists r_i.C_i$ in $\texttt{con}(C)$ with $C_i \sqsubseteq G$ is either removed or replaced by a conjunction of existential restrictions $\exists r_i.F$ such that (by induction) $F \not\sqsubseteq G$. In addition, other existential restrictions are either removed or generalized. This clearly implies $C' \not\sqsubseteq D_{i_j}$ since $\exists r_i.G$ in $\texttt{con}(D_{i_j})$ is not covered by any element of $\texttt{con}(C')$. □

The next lemma states that every compliant generalization of C subsumes some element of $SCG(C, \mathcal{P})$.

Lemma 1. *Let C be an $\mathcal{EL}$-concept and $\mathcal{P} = \{D_1, \ldots, D_p\}$ a policy. If C'' is a $\mathcal{P}$-compliant generalization of C, then there is $C' \in SCG(C, \mathcal{P})$ such that $C' \sqsubseteq C''$.*

Proof. If C is compliant with $\mathcal{P}$, then we have $C \in SCG(C,\mathcal{P})$ and $C \sqsubseteq C''$ since C'' is a generalization of C. Thus, assume that C is not compliant with $\mathcal{P}$, and let $i_1, \dots, i_q$ be exactly the indices for which $C \sqsubseteq D_i$.

Now, let i_j be such an index. We have $C \sqsubseteq C'' \not\sqsubseteq D_{i_j}$ and $C \sqsubseteq D_{i_j}$. Since $C'' \not\sqsubseteq D_{i_j}$, there is an element $E_j \in \mathtt{con}(D_{i_j})$ that is not covered by any element of $\mathtt{con}(C'')$. Obviously, $H'' := \{E_1, \dots, E_q\}$ is a hitting set of $\mathtt{con}(D_{i_1}), \dots, \mathtt{con}(D_{i_q})$. Thus, there is a minimal hitting set H of $\mathtt{con}(D_{i_1}), \dots,$ $\mathtt{con}(D_{i_q})$ such that $H \subseteq H''$. Let C' be the element of $SCG(C,\mathcal{P})$ that was constructed using this hitting set H. We claim that $C' \sqsubseteq C''$. For this, it is sufficient to show that $\mathtt{con}(C')$ covers $\mathtt{con}(C'')$.

First, consider a concept name $A \in \mathtt{con}(C'')$. Since $C \sqsubseteq C''$, we also have $A \in \mathtt{con}(C)$. If $A \notin H''$, then $A \notin H$, and thus A is not removed in the construction of C'. Consequently, $A \in \mathtt{con}(C')$ covers $A \in \mathtt{con}(C'')$. If $A \in H''$, then A is not covered by any element of $\mathtt{con}(C'')$ according to our definition of H'', which contradicts our assumption that $A \in \mathtt{con}(C'')$.

Second, consider an existential restriction $\exists r_i.E \in \mathtt{con}(C'')$. Since $C \sqsubseteq C''$, there is an existential restriction $\exists r_i.C_i$ in $\mathtt{con}(C)$ such that $C_i \sqsubseteq E$. If this restriction is not removed or generalized when constructing C', then we are done since this restriction then belongs to $\mathtt{con}(C')$ and covers $\exists r_i.E$. Otherwise, $\mathcal{P}_i = \{G \mid \text{ there is } \exists r_i.G \in H \text{ such that } C_i \sqsubseteq G\}$ is non-empty.

If $\top \in \mathcal{P}_i$, then $\exists r_i.\top \in H \subseteq H''$. However, then $\exists r_i.E \in \mathtt{con}(C'')$ covers an element of H'', which is a contradiction.

Consequently, $\top \notin \mathcal{P}_i$, and thus $\exists r_i.C_i$ is replaced with $\bigsqcap_{F \in SCG(C_i,\mathcal{P}_i)} \exists r_i.F$ when constructing C' from C. According to our definition of H'' and the fact that $H \subseteq H''$, none of the existential restrictions $\exists r_i.G$ considered in the definition of $\mathcal{P}_i$ is covered by $\exists r_i.E \in \mathtt{con}(C'')$. This implies that E is a $\mathcal{P}_i$-compliant generalization of C_i. By induction (on the role depth) we can thus assume that there is an $F \in SCG(C_i, \mathcal{P}_i)$ such that $F \sqsubseteq E$. This shows that $\exists r_i.E \in \mathtt{con}(C'')$ is covered by $\exists r_i.F \in \mathtt{con}(C')$. □

As an easy consequence of this lemma, we obtain that all optimal compliant generalizations of C must belong to $SCG(C,\mathcal{P})$.

Proposition 4. *Let C be an $\mathcal{EL}$-concept and $\mathcal{P} = \{D_1, \dots, D_p\}$ a policy. If C'' is an optimal $\mathcal{P}$-compliant generalization of C, then $C'' \in SCG(C,\mathcal{P})$ (up to equivalence of concepts).*

Proof. Let C'' be an optimal $\mathcal{P}$-compliant generalization of C. By Lemma 1, there is an element $C' \in SCG(C,\mathcal{P})$ such that $C' \sqsubseteq C''$. In addition, by Proposition 3, C' is a $\mathcal{P}$-compliant generalization of C. Thus, optimality of C'' implies $C'' \equiv C'$. □

We are now ready to formulate and prove the main result of this section.

Theorem 1. *Let C be an $\mathcal{EL}$-concept and $\mathcal{P} = \{D_1, \dots, D_p\}$ a policy. Then the set of all optimal $\mathcal{P}$-compliant generalizations of C can be computed in time exponential in the size of C and $D_1, \dots, D_p$.*

Proof. It is sufficient to show that the set $SCG(C, \mathcal{P})$ can be computed in exponential time. In fact, given $SCG(C, \mathcal{P})$, we can compute the set of all optimal $\mathcal{P}$-compliant generalizations of C by removing elements that are not minimal w.r.t. subsumption, which requires at most exponentially many subsumption tests. Each subsumption test takes at most exponential time since subsumption in $\mathcal{EL}$ is in P, and the elements of $SCG(C, \mathcal{P})$ have at most exponential size, as shown below.

We show by induction on the role depth that $SCG(C, \mathcal{P})$ consists of at most exponentially many elements of at most exponential size. The at most exponential cardinality of $SCG(C, \mathcal{P})$ is an immediate consequence of the fact that there are at most exponentially many hitting sets of $\mathtt{con}(D_{i_1}), \ldots, \mathtt{con}(D_{i_q})$, and each yields exactly one element of $SCG(C, \mathcal{P})$ (see Definition 4). Regarding the size of these elements, note that we may assume by induction that an existential restriction may be replaced by a conjunction of at most exponentially many existential restrictions, where each is of at most exponential size. The overall size of the concept description obtained this way is thus also of at most exponential size. Given this, it is easy to see that the computation of these elements also takes at most exponential time. □

The following example shows that the exponential upper bounds can indeed by reached.

Example 2. Let $C = P_1 \sqcap Q_1 \sqcap \ldots \sqcap P_n \sqcap Q_n$ and $\mathcal{P} = \{P_i \sqcap Q_i \mid 1 \leq i \leq n\}$. Then $SCG(C, \mathcal{P})$ contains 2^n elements since the sets $\{P_1, Q_1\}, \ldots, \{P_n, Q_n\}$ obviously have exponentially many hitting sets. To be more precise,

$$SCG(C, \mathcal{P}) = \{X_1 \sqcap \ldots \sqcap X_n \mid X_i \in \{P_i, Q_i\} \text{ for } i = 1, \ldots, n\}.$$

This example can easily be modified to enforce an element of exponential size. Consider $\widehat{C} = \exists r.C$ and $\widehat{\mathcal{P}} = \{\exists r.(P_i \sqcap Q_i) \mid 1 \leq i \leq n\}$. Then $SCG(\widehat{C}, \widehat{\mathcal{P}}) = \{\bigsqcap_{F \in SCG(C, \mathcal{P})} \exists r.F\}$. We leave it to the reader to further modify the example in order to obtain exponentially many elements of exponential size.

4 Characterizing safety

Before we can characterize safety, we need to remove redundant elements from $\mathcal{P}$. We say that $D_i \in \mathcal{P}$ is *redundant* if there is a different element $D_j \in \mathcal{P}$ such that $D_i \sqsubseteq D_j$. The following lemma is easy to prove.

Lemma 2. *Let $\mathcal{P}$ be a policy and assume that $D_i \in \mathcal{P}$ is redundant. Then the following holds for all $\mathcal{EL}$ concepts C, C':*

- *C' is compliant with $\mathcal{P}$ iff C' is compliant with $\mathcal{P} \setminus \{D_i\}$;*
- *C is safe for $\mathcal{P}$ iff C is safe for $\mathcal{P} \setminus \{D_i\}$.*

This lemma shows that we can assume without loss of generality that our policies do not contain redundant concepts. However, elements of D_i of $\mathcal{P}$ may also contain redundant atoms. In fact, if E, F are different atoms in $\texttt{con}(D_i)$ such that $E \sqsubseteq F$, then the concept obtained from D_i by removing F from its top-level conjunction is equivalent to D_i. By iteratively removing such redundant atoms from the top-level conjunction of D_i we obtain (in polynomial time) a concept D_i' equivalent to D_i such that the elements of $\texttt{con}(D_i')$ are incomparable w.r.t. subsumption. We call a policy *redundancy-free* if it does not contain redundant elements and every element is *normalized* in this sense.

Proposition 5. *Let $\mathcal{P} = \{D_1, \ldots, D_p\}$ be a redundancy-free policy. The $\mathcal{EL}$ concept C' is safe for $\mathcal{P}$ iff there is no pair of atoms (E, F) such that $E \in$ $\texttt{con}(C')$, $F \in \texttt{con}(D_1) \cup \ldots \cup \texttt{con}(D_p)$, and $E \sqsubseteq F$.*

Proof. First, assume that C' is not safe for $\mathcal{P}$, i.e., there is an $\mathcal{EL}$ concept C'' that is compliant with $\mathcal{P}$, but for which $C' \sqcap C''$ is not compliant with $\mathcal{P}$. The latter implies that there is $D_i \in \mathcal{P}$ such that $C' \sqcap C'' \sqsubseteq D_i$, which is equivalent to saying that $\texttt{con}(C') \cup \texttt{con}(C'')$ covers $\texttt{con}(D_i)$. On the other hand, we know that $\texttt{con}(C'')$ does not cover $\texttt{con}(D_i)$ since C'' is compliant with $\mathcal{P}$. Thus, there is an element $F \in \texttt{con}(D_i)$ that is covered by an element E of $\texttt{con}(C')$. This yields (E, F) such that $E \in \texttt{con}(C')$, $F \in \texttt{con}(D_1) \cup \ldots \cup \texttt{con}(D_p)$, and $E \sqsubseteq F$.

Conversely, assume that there is a pair of atoms (E, F) such that $E \in$ $\texttt{con}(C')$, $F \in \texttt{con}(D_i)$, and $E \sqsubseteq F$. Let C'' be the concept obtained from D_i by removing F from the top-level conjunction of D_i. Then we clearly have $D_i \sqsubseteq C''$. In addition, since D_i is normalized, we also have $C'' \not\sqsubseteq D_i$. Consider $D_j \in \mathcal{P}$ different from D_i, and assume that $C'' \sqsubseteq D_j$. But then $D_i \sqsubseteq C'' \sqsubseteq D_j$ contradicts our assumption that $\mathcal{P}$ does not contain redundant elements. Thus, we have shown that C'' is compliant with $\mathcal{P}$. In addition, $\texttt{con}(C') \cup \texttt{con}(C'')$ covers $\texttt{con}(D_i)$. In fact, the elements of $\texttt{con}(D_i) \setminus \{F\}$ belong to $\texttt{con}(C'')$, and thus cover themselves. In addition, F is covered by $E \in \texttt{con}(C')$. Thus $C' \sqcap C'' \sqsubseteq D_i$, which shows that C' is not safe for $\mathcal{P}$. □

Clearly, the necessary and sufficient condition for safety stated in this proposition can be decided in polynomial time. If needed, the policy can first be made redundancy-free, which can also be done in polynomial time.

Corollary 1. *Safety of an $\mathcal{EL}$ concept for an $\mathcal{EL}$ policy is in P.*

We now consider the problem of computing optimal $\mathcal{P}$-safe generalizations of a given $\mathcal{EL}$ concept C. First note that, up to equivalence, there can be only one optimal $\mathcal{P}$-safe generalization of C. This is an immediate consequence of the fact that the conjunction of safe concepts is again safe, which in turn is an easy consequence of Proposition 5.

Lemma 3. *Let C_1', C_2' be two $\mathcal{EL}$ concepts that are $\mathcal{P}$-safe generalizations of C, where $\mathcal{P}$ is redundancy-free. Then $C_1' \sqcap C_2'$ is also a $\mathcal{P}$-safe generalization of C.*

Thus there cannot be non-equivalent optimal $\mathcal{P}$-safe generalizations of a given $\mathcal{EL}$ concept C since their conjunction would then be more specific, contradicting their optimality. This property is independent of whether the policy is redundancy-free or not since turning a policy into one that is redundancy-free preserves the set of concepts that are compliant with (safe for) the policy.

Proposition 6. *If C'_1, C'_2 are optimal $\mathcal{P}$-safe generalizations of the $\mathcal{EL}$ concept C, then $C'_1 \equiv C'_2$.*

The following theorem shows how an optimal safe generalization of C can be constructed.

Theorem 2. *Let C be an $\mathcal{EL}$ concept and $\mathcal{P} = \{D_1, \ldots, D_p\}$ a redundancy-free policy. We construct the concept C' from C by removing or modifying atoms in the top-level conjunction of C in the following way:*

- *For every concept name $A \in \mathtt{con}(C)$, remove A from the top-level conjunction of C if $A \in \mathtt{con}(D_1) \cup \ldots \cup \mathtt{con}(D_p)$;*
- *For every existential restriction $\exists r_i.C_i \in \mathtt{con}(C)$, consider the set of concepts*

  $$\mathcal{P}_i := \{G \mid \text{ there is } \exists r_i.G \in \mathtt{con}(D_1) \cup \ldots \cup \mathtt{con}(D_p) \text{ such that } C_i \sqsubseteq G\}.$$

  - *If $\mathcal{P}_i = \emptyset$ then leave $\exists r_i.C_i$ as it is.*
  - *If $\top \in \mathcal{P}_i$, then remove $\exists r_i.C_i$.*
  - *Otherwise, replace $\exists r_i.C_i$ with $\bigsqcap_{F \in OCG(C_i, \mathcal{P}_i)} \exists r_i.F$, where $OCG(C_i, \mathcal{P}_i)$ is the set of all optimal $\mathcal{P}_i$-compliant generalizations of C_i.*

Then C' is an optimal $\mathcal{P}$-safe generalization of C.

Proof. Obviously $C \sqsubseteq C'$ since, when constructing C' from C, atoms from the top-level conjunction of C are left unchanged, are removed, or are replaced by a conjunction of more general atoms.

To show that C' is safe for $\mathcal{P}$, we must show that the condition of Proposition 5 holds. Thus assume that it is violated, i.e., there is a pair of atoms (E, F) such that $E \in \mathtt{con}(C')$, $F \in \mathtt{con}(D_1) \cup \ldots \cup \mathtt{con}(D_p)$, and $E \sqsubseteq F$.

- First, we consider the case where $E = A$ is a concept name. Then $E \sqsubseteq F$ implies that $F = A$, and thus A is a concept name occurring in $\mathtt{con}(D_1) \cup \ldots \cup \mathtt{con}(D_p)$. However, all such concept names have been removed from the top-level conjunction of C when constructing C'. This contradicts our assumption that $E = A$ belongs to $\mathtt{con}(C')$.
- Second, assume that E is an existential restriction $E = \exists r_i.E'$. Then F is of the form $F = \exists r_i.G'$ and $E' \sqsubseteq G'$. In addition, there is an existential restriction $\exists r_i.C_i \in \mathtt{con}(C)$ from which $E = \exists r_i.E'$ was derived. By construction, $C_i \sqsubseteq E'$. In the construction of C', we consider the set $\mathcal{P}_i := \{G \mid \text{ there is } \exists r_i.G \in \mathtt{con}(D_1) \cup \ldots \cup \mathtt{con}(D_p) \text{ such that } C_i \sqsubseteq G\}$. Since $C_i \sqsubseteq E' \sqsubseteq G'$, this set is non-empty, and since $\exists r_i.E'$ is derived from $\exists r_i.C_i$, it does not contain $\top$. Consequently, we have $E' \in OCG(C_i, \mathcal{P}_i)$. However, $G' \in \mathcal{P}_i$ then implies that $E' \not\sqsubseteq G'$, which yields the desired contradiction.

It remains to show that C' is optimal. Thus assume that C'' is a $\mathcal{P}$-safe generalization of C. It is sufficient to show that $C' \sqsubseteq C''$, i.e., that $\mathtt{con}(C')$ covers $\mathtt{con}(C'')$.

- Assume that $A \in \texttt{con}(C'')$ is a concept name. Then $C \sqsubseteq C''$ implies that $A \in \texttt{con}(C)$. In addition, since C'' is safe for $\mathcal{P}$, Proposition 5 implies that $A \notin \texttt{con}(D_1) \cup \ldots \cup \texttt{con}(D_p)$. Thus, A is not removed in the construction of C', which yields $A \in \texttt{con}(C')$.
- Second, consider an existential restriction $\exists r_i.E \in \texttt{con}(C'')$. Since $C \sqsubseteq C''$, there is an existential restriction $\exists r_i.C_i$ in $\texttt{con}(C)$ such that $C_i \sqsubseteq E$. If this restriction is not removed or generalized when constructing C', then we are done since this restriction then belongs to $\texttt{con}(C')$ and covers $\exists r_i.E$. Otherwise, $\mathcal{P}_i = \{G \mid \text{ there is } \exists r_i.G \in \texttt{con}(D_1) \cup \ldots \cup \texttt{con}(D_p) \text{ such that } C_i \sqsubseteq G\}$ is non-empty.

  If $\top \in \mathcal{P}_i$, then $\exists r_i.\top \in \texttt{con}(D_1) \cup \ldots \cup \texttt{con}(D_p)$. However, then $\exists r_i.E \in \texttt{con}(C'')$ covers an element of $\texttt{con}(D_1) \cup \ldots \cup \texttt{con}(D_p)$, which is a contradiction to our assumption that C'' is safe for $\mathcal{P}$.

  Consequently, $\top \notin \mathcal{P}_i$, and thus $\exists r_i.C_i$ is replaced with $\bigsqcap_{F \in OCG(C_i,\mathcal{P}_i)} \exists r_i.F$ when constructing C' from C. Since C'' is safe for $\mathcal{P}$, none of the existential restrictions $\exists r_i.G$ considered in the definition of $\mathcal{P}_i$ is covered by $\exists r_i.E \in \texttt{con}(C'')$. This implies that E is a $\mathcal{P}_i$-compliant generalization of C_i. Consequently, there is an $F \in OCG(C_i, \mathcal{P}_i)$ such that $F \sqsubseteq E$. This shows that $\exists r_i.E \in \texttt{con}(C'')$ is covered by $\exists r_i.F \in \texttt{con}(C')$. □

Since, by Theorem 1, $OCG(C_i, \mathcal{P}_i)$ can be computed in exponential time, the construction described in Theorem 2 can also be performed in exponential time.

Corollary 2. *Let C be an $\mathcal{EL}$ concept and $\mathcal{P} = \{D_1, \ldots, D_p\}$ a redundancy-free policy. Then an optimal $\mathcal{P}$-safe generalization of C can be computed in exponential time.*

Example 2 can easily be modified to provide an example that shows that this exponential bound can actually not be improved since there are cases where the safe generalization is of exponential size.

5 Conclusion

We have introduced the notions of compliance with and safety for a policy in the simple setting where both the knowledge about individuals and the policy are given by $\mathcal{EL}$ concepts. In this setting, we were able to characterize compliant (safe) generalization of a given concept w.r.t. a policy, and have used these characterizations to obtain algorithms for computing these generalizations. These algorithms need exponential time, which is optimal since the generalizations may be of exponential size.

In the future, we intend to extend this work in two directions. On the one hand, we will consider $\mathcal{EL}$ concepts w.r.t. a background ontology. On the other hand, we will consider a setting where the ABox contains not just concept assertions, but also role assertions. In the latter case, one can use not just generalization of concepts, but also renaming of individuals as operations for achieving compliance (safety). Finally, of course, these two extensions should be combined.

References

1. F. Baader, S. Brandt, and C. Lutz. Pushing the $\mathcal{EL}$ envelope. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence IJCAI-05*, Edinburgh, UK, 2005. Morgan-Kaufmann Publishers.
2. F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications.* Cambridge University Press, New York, NY, USA, 2003.
3. F. Baader, R. Küsters, and R. Molitor. Computing least common subsumers in description logics with existential restrictions. In *Proc. of the 16th Int. Joint Conf. on Artificial Intelligence (IJCAI'99)*, pages 96–101, 1999.
4. F. Baader and B. Morawska. Unification in the description logic EL. *Logical Methods in Computer Science*, 6(3), 2010.
5. S. Brandt. Polynomial time reasoning in a description logic with existential restrictions, GCI axioms, and—what else? In R. L. de Mántaras and L. Saitta, editors, *Proc. of the 16th Eur. Conf. on Artificial Intelligence (ECAI 2004)*, pages 298–302, 2004.
6. B. Cuenca Grau and E. V. Kostylev. Logical foundations of privacy-preserving publishing of linked data. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 943–949, 2016.
7. B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4):14:1–14:53, 2010.
8. I. Horrocks, L. Li, D. Turi, and S. Bechhofer. The instance store: DL reasoning with large numbers of individuals. In *Proceedings of the 2004 International Workshop on Description Logics (DL2004), Whistler, British Columbia, Canada, June 6-8, 2004*, 2004.