More is Sometimes Less: Succinctness in $\mathcal{EL}$

Nadeschda Nikitina[1] and Sven Schewe[2]

[1] University of Oxford, UK
[2] University of Liverpool, UK

Abstract. In logics, there are many ways to represent same facts. With respect to both reasoning and cognitive complexity, some representations are significantly less efficient than others. In this paper, we investigate different means of improving the succinctness of TBoxes expressed in the lightweight description logic $\mathcal{EL}$ that forms a basis of some large ontologies used in practice. As a measure of size, we consider the number of references to signature elements. We investigate the problem of finding minimal equivalent representations and show that this task is NP-complete.

A significant (up to triple-exponential) further improvement can be achieved by the introduction of auxiliary concept symbols. Thus, we additionally investigate the task of finding minimal representations for an ontology by extending its signature. Since arbitrary extension of the ontology with concept symbols can make the ontology unreadable, we only allow for auxiliary concepts acting as shortcuts for other concepts ($\mathcal{EL}$ concepts and disjunctions thereof) expressed by means of terms of the original ontology. We show that this task is also NP-complete if shortcuts represent only $\mathcal{EL}$ concepts, and between NP and Σ_2^P, otherwise.

1 Introduction

It is well-known that same facts can be represented in many ways, and that the size of these representations can vary significantly. Determining and increasing the degree of succinctness of a particular syntactic representation is an important, but also a very difficult task: for the average ontology, it is almost impossible to obtain the minimal representation without tool support. Thus, automated methods that help to assess the current succinctness of an ontology and generate suggestions on how to increase it would be highly valued by ontology engineers.

In description logics [1], only few results in this direction were obtained so far. Baader, Küster, and Molitor [2] investigate rewriting concepts using terminologies in the narrow sense (sets of equivalence axioms where each defined atomic concept has exactly one definition). The investigated problem is a special case of minimizing a knowledge base by computing a minimal equivalent knowledge base. Grimm et al. [3] propose an algorithm for eliminating semantically redundant axioms from ontologies. In the above approach, axioms are considered as atoms that cannot be split into parts or changed in any other way. Bienvenu [4] proposes a normal form called prime implicates normal form for $\mathcal{ALC}$ ontologies, which enables fast reasoning. However, as a side-effect of this transformation, a doubly-exponential blowup in concept size can occur.

In this paper, we investigate the succinctness for the lightweight description logic $\mathcal{EL}$ [5], which is the logical underpinning of one of the tractable sub-languages (the so-called *profiles* [6]) of the W3C-specified OWL Web Ontology Language [7].

First, we consider the problem of finding a minimal equivalent $\mathcal{EL}$ representation for a given ontology. We show that the related decision problem (is there an equivalent ontology of size $\leq k$?) is NP-complete.

Inspired by recent results on uniform interpolation in $\mathcal{EL}$ [8], we additionally consider an extended version of the problem. The above results imply that, even for the minimal equivalent representation of an ontology, an up to triple-exponentially more succinct representation can be obtained by extending its signature. Auxiliary concept symbols are therefore important contributors towards the succinctness of ontologies. It is easy to envision scenarios that demonstrate the usefulness of auxiliary concept symbols for improving succinctness. For instance, when a complex concept C is frequently used in the axioms of an ontology, the ontology will diminish in size when all occurrences of C are replaced by a fresh atomic concept A_C, and an axiom $A_C \equiv C$ is added to the ontology. However, an arbitrary extension of the ontology with concept symbols whose meaning is not obvious can certainly make the ontology unreadable. In order to preserve comprehensiveness, we only allow for auxiliary concepts acting as *shortcuts* – concepts that are defined using only terms of the original ontology. Presented with such a shortcut concept, an ontology engineer could find an appropriate comprehensive name for it. Otherwise, the ontology engineer has to guess the meaning of an auxiliary concept and the chance that he approves the extension suggested by the tool would be low.

We demonstrate that auxiliary concept symbols acting as shortcuts for $\mathcal{EL}$ concepts expressed only by means of original ontology terms can lead to an exponential improvement of succinctness and that the corresponding decision problem (is there such a representation of size $\leq k$?) is NP-complete.

Further, we show that, if we additionally allow for auxiliary concept symbols that act as shortcuts for disjunctions of $\mathcal{EL}$ concepts on the left-hand side of axioms (encodable in $\mathcal{EL}$ using several axioms), we can reduce the size of the representation by a further exponent, thereby obtaining doubly-exponentially more succinct representations.We show that the corresponding decision problem (is there such a representation of size $\leq k$?) is NP-hard and included in Σ_2^P.

The paper is organized as follows: In Section 2, we recall the necessary preliminaries on description logics. Section 3 demonstrates the potential of auxiliary concept symbols acting as shortcuts for achieving a higher succinctness. In the same section, we also introduce the basic definitions of the size of ontologies as well as the investigated notions of equivalents with and without signature extension. In Sections 4,5, we derive the complexity bounds for the corresponding decision problems. Finally, we conclude and outline future work in Section 6. Further details and proofs can be found in the extended version [9] of this paper.

2 Preliminaries

We recall the basic notions in description logics [1] required in this paper. Let N_C and N_R be countably infinite and mutually disjoint sets of concept symbols and role symbols. An $\mathcal{EL}$ concept C is defined as

$$C ::= A|\top|C \sqcap C|\exists r.C,$$

where A and r range over N_C and N_R, respectively. In the following, we use symbols A, B to denote atomic concepts and C, D, E to denote arbitrary concepts. A *terminology* or *TBox* consists of *concept inclusion* axioms $C \sqsubseteq D$ and *concept equivalence* axioms $C \equiv D$ used as a shorthand for $C \sqsubseteq D$ and $D \sqsubseteq C$. The signature of an $\mathcal{EL}$ concept C or an axiom α, denoted by $\text{sig}(C)$ or $\text{sig}(\alpha)$, respectively, is the set of concept and role symbols occurring in it. To distinguish between the set of concept symbols and the set of role symbols, we use $\text{sig}_C(C)$ and $\text{sig}_R(C)$, respectively. The signature of a TBox $\mathcal{T}$, in symbols $\text{sig}(\mathcal{T})$ (correspondingly, $\text{sig}_C(\mathcal{T})$ and $\text{sig}_R(\mathcal{T})$), is defined analogously. Next, we recall the semantics of the above introduced DL constructs, which is defined by means of interpretations. An interpretation $\mathcal{I}$ is given by the domain $\Delta^{\mathcal{I}}$ and a function $\cdot^{\mathcal{I}}$ assigning each concept $A \in N_C$ a subset $A^{\mathcal{I}}$ of $\Delta^{\mathcal{I}}$ and each role $r \in N_R$ a subset $r^{\mathcal{I}}$ of $\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$. The interpretation of $\top$ is fixed to $\Delta^{\mathcal{I}}$. The interpretation of an arbitrary $\mathcal{EL}$ concept is defined inductively, i.e., $(C \sqcap D)^{\mathcal{I}} = C^{\mathcal{I}} \cap D^{\mathcal{I}}$ and $(\exists r.C)^{\mathcal{I}} = \{x \mid (x, y) \in r^{\mathcal{I}}, y \in C^{\mathcal{I}}\}$. An interpretation $\mathcal{I}$ satisfies an axiom $C \sqsubseteq D$ if $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$. $\mathcal{I}$ is a model of a TBox, if it satisfies all of its axioms. We say that a TBox $\mathcal{T}$ entails an axiom α (in symbols, $\mathcal{T} \models \alpha$), if α is satisfied by all models of $\mathcal{T}$. A TBox $\mathcal{T}$ entails another TBox $\mathcal{T}'$, in symbols $\mathcal{T} \models \mathcal{T}'$, if $\mathcal{T} \models \alpha$ for all $\alpha \in \mathcal{T}'$. $\mathcal{T} \equiv \mathcal{T}'$ is a shortcut for $\mathcal{T} \models \mathcal{T}'$ and $\mathcal{T}' \models \mathcal{T}$.

In addition to $\mathcal{EL}$, we will use disjunction on the left-hand side of axioms to obtain more succinct representations of $\mathcal{EL}$ TBoxes. Note that this extension is of a notational nature, i.e., does not give us the expressive power to represent more TBoxes than standard $\mathcal{EL}$. We define an $\mathcal{ELD}$ concept C as

$$C ::= A|\top|C \sqcap C|C \sqcup C|\exists r.C,$$

where A and r range over N_C and N_R, respectively. The interpretation of an arbitrary $\mathcal{ELD}$ concept is defined analogously to the interpretation of $\mathcal{EL}$ concepts with the extension $(C \sqcup D)^{\mathcal{I}} = C^{\mathcal{I}} \cup D^{\mathcal{I}}$. An $\mathcal{ELD}$ TBox consists of axioms that are either $\mathcal{EL}$ axioms or have the form $C \sqsubseteq D$, where C is an $\mathcal{ELD}$ concept and D is an $\mathcal{EL}$ concept. Note that equivalence axioms ($C \equiv D$) do not contain $\mathcal{ELD}$ concepts, since they are a shortcut for $C \sqsubseteq D$ and $D \sqsubseteq C$.

3 Achieving Succinctness in $\mathcal{EL}$

The size of a TBox is often measured by the number of axioms contained in it. This is, however, a very simplified view of the size in terms of both, cognitive complexity and reasoning. In this paper, we measure the size of a concept, an axiom, or a TBox by the number of references to signature elements.

Definition 1. *The* size *of an $\mathcal{EL}$ concept D is defined as follows:*

- *for $D \in sig(\mathcal{T})$, $\int(D) = 1$;*
- *for $D = \exists r.C$, $\int(D) = \int(C) + 1$ where $r \in sig_R(\mathcal{T})$ and C is an arbitrary concept;*
- *for $D = C_1 \sqcap C_2$, $\int(D) = \int(C_1) + \int(C_2)$ where C_1, C_2 are arbitrary concepts;*

The size *of an $\mathcal{EL}$ axiom or a TBox is accordingly defined as follows:*

- *$\int(C_1 \sqsubseteq C_2) = \int(C_1) + \int(C_2)$ for concepts C_1, C_2;*
- *$\int(C_1 \equiv C_2) = \int(C_1) + \int(C_2)$ for concepts C_1, C_2.*
- *$\int(\mathcal{T}) = \sum_{\alpha \in \mathcal{T}} \int(\alpha)$ for a TBox $\mathcal{T}$.*

In practice, the suitable means that can be used to obtain a compact representation can differ depending on the scenario. To address cases, in which a signature extension is not feasible, we first consider the problem of finding the minimal equivalent $\mathcal{EL}$ representation for a given TBox among representations that use the same signature. Popular examples for avoidable non-succinctness are axioms that follow from other axioms and sub-concepts that can be removed from axioms without losing any logical consequences. While non-succinctness is easy to detect in these simple cases, non-succinctness can occur in many other forms. The ontology $\mathcal{T} = \{C \sqsubseteq \exists r.C, \exists r.C \sqsubseteq \exists r.D, \exists r.D \sqsubseteq D\}$, for instance, does neither contain any axioms that are entailed by the remainder of the ontology, nor are there any sub-expressions that can be removed. However, there exists a smaller representation $\{C \sqsubseteq \exists r.C, C \sqsubseteq D, \exists r.D \sqsubseteq D\}$ of $\mathcal{T}$. The general version of the corresponding decision problem can be formulated as follows:

Definition 2 (P1). *Given an $\mathcal{EL}$ TBox $\mathcal{T}$ and a natural number k, is there an $\mathcal{EL}$ TBox $\mathcal{T}'$ with $\int(\mathcal{T}') \leq k$ such that $\mathcal{T}' \equiv \mathcal{T}$.*

We denote the set $\{\mathcal{T}' \mid \mathcal{T}' \equiv \mathcal{T}\}$ by $[\mathcal{T}]$. We will show that this decision problem, which does not involve any signature extensions, is already NP-complete.

Extending the Signature

From the user's point of view as well as with respect to reasoning, it sometimes makes sense to introduce fresh concept symbols, for instance, used as shortcuts for complex concepts that occur frequently in the ontology. It can be a tedious task for an ontology engineer to do it in an advantageous way, since, as we will show later on, the corresponding decision problem is NP-hard. To account for scenarios, in which an introduction of auxiliary concept symbols is desirable, in addition to the decision problem introduced above we consider the problem of finding succinct representations containing shortcuts. We demonstrate by means of the following example the theoretical potential of such an extension of the signature with shortcuts: we show that it can lead to a doubly-exponentially more succinct representation of TBoxes.

Example 1. Let the sets $\mathcal{C}_i$ of concept descriptions be inductively defined by $\mathcal{C}_0 = \{A_1, A_2\}$, $\mathcal{C}_{i+1} = \{\exists r.C_1 \sqcap \exists s.C_2 \mid C_1, C_2 \in \mathcal{C}_i\}$. For a natural number n, consider the TBox $\mathcal{T}_n = \{C \sqsubseteq B \mid C \in \mathcal{C}_{n-1}\}$.

Intuitively, the sets $\mathcal{C}_i$ of concepts have the shape of binary trees with exponentially many leaves, each of which can be A_1 or A_2. Clearly, the concepts grow exponentially with i. Further, it holds that $|\mathcal{C}_{i+1}| = |\mathcal{C}_i|^2$ and consequently $|\mathcal{C}_i| = 2^{(2^i)}$. Thus, $\mathcal{T}_n$ contains doubly exponentially many axioms, each of which has exponential size. While there is no smaller equivalent representation of $\mathcal{T}_n$, this TBox can easily be represented in a more compact way using auxiliary concept symbols as shortcuts for complex $\mathcal{EL}$ or $\mathcal{ELD}$ concept expressions.

First, combining several axioms into a single axiom with a disjunction on the left-hand side would allow us to reduce the size of $\mathcal{T}_n$ from double-exponential to single-exponential: we can define $\mathcal{C}_0 = \{A_1 \sqcup A_2\}$ and thus express all elements of the set $\mathcal{C}_{n-1}$ by means of a single concept C_{n-1} that has the shape of a binary tree with the concept $A_1 \sqcup A_2$ as leaves. The corresponding $\mathcal{EL}$ TBox $\mathcal{T}'_n$ can be obtained by introducing the concept B_0 that represents the disjunction $A_1 \sqcup A_2$ by means of the axioms $A_1 \sqsubseteq B_0$ and $A_2 \sqsubseteq B_0$.

Second, by using fresh concept symbols as shortcuts for complex $\mathcal{EL}$ concepts, $\mathcal{T}'_n$ can be reduced by a further exponential as follows: we introduce concept symbols B_i with $i \in \{1, ..., n-1\}$ to represent each C_i and obtain the following TBox $\mathcal{T}''_n$:

$$A_1 \sqsubseteq B_0 \tag{1}$$

$$A_2 \sqsubseteq B_0 \tag{2}$$

$$B_{i+1} \equiv \exists r.B_i \sqcap \exists s.B_i \qquad i < n-1 \tag{3}$$

$$B_{n-1} \sqsubseteq B \tag{4}$$

As a result, the binary tree contracts into a chain of $n+3$ axioms α_j with $\int(\alpha_j) \leq 5$.

In general, an extension of the signature has to be meaning-preserving in the sense that the logical consequences expressed using only the originally given signature remain unchanged. Formally, the corresponding "equivalence" between TBoxes with different signatures is captured by the notion of *inseparability* as investigated by various authors [10–15] in different variations. We base this work on the deductive notion of inseparability for $\mathcal{EL}$. Two $\mathcal{EL}$ TBoxes, $\mathcal{T}_1$ and $\mathcal{T}_2$, are inseparable w.r.t. a signature Σ if they have the same $\mathcal{EL}$ consequences whose signature is a subset of Σ:

Definition 3. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two general $\mathcal{EL}$ TBoxes and Σ a signature. $\mathcal{T}_1$ and $\mathcal{T}_2$ are Σ-inseparable, in symbols $\mathcal{T}_1 \equiv^{\mathcal{EL}}_{\Sigma} \mathcal{T}_2$, if for all $\mathcal{EL}$ concepts C, D with sig$(C) \cup$ sig$(D) \subseteq \Sigma$ it holds that $\mathcal{T}_1 \models C \sqsubseteq D$, iff $\mathcal{T}_2 \models C \sqsubseteq D$.*

Thus, the formal requirement for any TBox $\mathcal{T}'$ obtained from $\mathcal{T}$ by means of a signature extension is that it remains Σ-inseparable from $\mathcal{T}$, where $\Sigma = \text{sig}(\mathcal{T})$. We take this into account in the subsequent definitions.

$\mathcal{EL}$-Shortcuts

We now consider the problem of finding small TBoxes that are Σ-inseparable from $\mathcal{T}$ (with $\Sigma = \text{sig}(\mathcal{T})$) and use explicitly defined $\mathcal{EL}$ shortcuts. From Example 1, we can

observe that a significantly higher effect can be achieved if shortcuts are introduced gradually such that previously introduced shortcuts can be used to define new ones. The definition below allows for a hierarchy of shortcuts. To ensure that shortcuts form a hierarchy, we impose an acyclicity condition on the syntactic references within the definitions of shortcuts.

Definition 4 ($\mathcal{EL}$-Shortcuts). *Let $\mathcal{T}$ be an $\mathcal{EL}$ TBox with $sig(\mathcal{T}) = \Sigma$. Then an $\mathcal{EL}$ TBox $\mathcal{T}'$ is an* equivalent with $\mathcal{EL}$-shortcuts*, in symbols $\mathcal{T}' \in [\mathcal{T}]_{\mathcal{EL}}$, iff*

1. $\mathcal{T}' \equiv_{\Sigma}^{\mathcal{EL}} \mathcal{T}$;
2. $sig_R(\mathcal{T}') = sig_R(\mathcal{T})$;
3. *for all $A_i \in \{A_1, \ldots, A_n\} = sig_C(\mathcal{T}') \setminus sig_C(\mathcal{T})$ there exists exactly one concept C_i (called* definition *of A_i) such that $A_i \equiv C_i \in \mathcal{T}'$;*
4. *for all $i \in \{1, \ldots, n\}$ it holds that $sig(C_i) \subseteq sig(\mathcal{T}) \cup \{A_j \mid j < i\}$.*

The introduction of $\mathcal{EL}$-shortcuts corresponds to the second transformation of the TBox given in Example 1. The corresponding decision problem can be stated as follows:

Definition 5 (P2). *Given an $\mathcal{EL}$ TBox $\mathcal{T}$ and a natural number k, is there an $\mathcal{EL}$ TBox $\mathcal{T}'$ with $\int(\mathcal{T}') \leq k$ such that $\mathcal{T}' \in [\mathcal{T}]_{\mathcal{EL}}$.*

It can be shown that the equivalence relation between $\mathcal{T}$ and its equivalent given in Definition 4 is stronger than deductive inseparability. It is called *emulation* and is defined as follows:

Definition 6. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two $\mathcal{EL}$ TBoxes. $\mathcal{T}_2$* emulates *$\mathcal{T}_1$, in symbols $\mathcal{T}_2 \models^{em} \mathcal{T}_1$, iff $\mathcal{T}_2 \models \mathcal{T}_1$ and every model of $\mathcal{T}_1$ can be extended into a model of $\mathcal{T}_2$.*

Clearly, $\mathcal{T}_2 \models^{em} \mathcal{T}_1$ implies $\mathcal{T}_2 \equiv_{\Sigma}^{\mathcal{EL}} \mathcal{T}_1$ with $\Sigma = \text{sig}(\mathcal{T}_1)$. The following lemma establishes the role of $\mathcal{EL}$-shortcuts within TBoxes:

Lemma 1. *Let $\mathcal{T}, \mathcal{T}'$ be two $\mathcal{EL}$ TBoxes such that $\mathcal{T}' \in [\mathcal{T}]_{\mathcal{EL}}$ and $\{A_1, \ldots, A_n\} = sig_C(\mathcal{T}') \setminus sig_C(\mathcal{T})$. Further, let C_i be the corresponding definition of A_i. Then for the TBox $\mathcal{T}_{\texttt{ext}} = \mathcal{T} \cup \{A_i \equiv C_i \mid i \in \{1, \ldots, n\}\}$ it holds that $\mathcal{T}_{\texttt{ext}} \models^{em} \mathcal{T}$.*

Proof Sketch. Clearly, the interpretation of each $A_i \notin \Sigma$ is completely determined by the interpretations of symbols in Σ (due to acyclicity condition on the syntactic references within the definitions of the shortcuts). Thus, we can extend each model of $\mathcal{T}$ by assigning $A_i^{\mathcal{I}} = C_i^{\mathcal{I}}$ and obtain a model of $\mathcal{T}_{\texttt{ext}}$. Additionally, $\mathcal{T}_{\texttt{ext}} \models \mathcal{T}$, since $\mathcal{T} \subseteq \mathcal{T}_{\texttt{ext}}$. □

⊔-Shortcuts

The second important contribution of additional vocabulary elements to succinctness of $\mathcal{EL}$ TBoxes is their ability to act as a replacement for disjunction on the left-hand side of axioms. We can obtain a corresponding $\mathcal{EL}$ TBox $\mathcal{T}'$ from an $\mathcal{ELD}$ TBox $\mathcal{T}$ by replacing each disjunction $C_1 \sqcup ... \sqcup C_n$ occurring in $\mathcal{T}$ by a fresh concept symbol A and extending $\mathcal{T}$ with axioms $C_1 \sqsubseteq A, ..., C_n \sqsubseteq A$, called *definitions* of A. We denote such an $\mathcal{EL}$ representation of $\mathcal{T}$ by $T_{\mathcal{EL}}(\mathcal{T})$.

Definition 7 ($\sqcup$-Shortcuts). *Let $\mathcal{T}$ be an $\mathcal{EL}$ TBox. Then an $\mathcal{EL}$ TBox $\mathcal{T}'$ is an* equivalent with $\sqcup$-shortcuts*, in symbols $\mathcal{T}' \in [\mathcal{T}]_{\sqcup}$, iff there is an $\mathcal{ELD}$ TBox $\mathcal{T}''$ such that $\mathcal{T}'' \equiv \mathcal{T}, sig(\mathcal{T}'') \subseteq sig(\mathcal{T})$ and $T_{\mathcal{EL}}(\mathcal{T}'') = \mathcal{T}'$.*

Introduction of $\sqcup$-shortcuts corresponds to the first transformation in Example 1. The corresponding decision problem is as follows:

Definition 8 (P3). *Given an $\mathcal{EL}$ TBox $\mathcal{T}$ and a natural number k, is there an $\mathcal{EL}$ TBox $\mathcal{T}'$ with $\int(\mathcal{T}') \leq k$ such that $\mathcal{T}' \in [\mathcal{T}]_{\sqcup}$.*

$\mathcal{ELD}$-Shortcuts

If we simultaneously allow for both types of shortcuts (note that these roles can never be played by a single concept at the same time!), we obtain the following definition of equivalents:

Definition 9 ($\mathcal{ELD}$-Shortcuts). *Let $\mathcal{T}$ be an $\mathcal{EL}$ TBox. Then an $\mathcal{EL}$ TBox $\mathcal{T}'$ is an* equivalent with $\mathcal{ELD}$-shortcuts*, in symbols $\mathcal{T}' \in [\mathcal{T}]_{\mathcal{ELD}}$, iff there is an $\mathcal{ELD}$ TBox $\mathcal{T}''$ such that Conditions 1-4 of Definition 4 are true for $\mathcal{T}''$ and $\mathcal{T}' = T_{\mathcal{EL}}(\mathcal{T}'')$.*

The corresponding decision problem is stated as follows:

Definition 10 (P4). *Given an $\mathcal{EL}$ TBox $\mathcal{T}$ and a natural number k, is there an $\mathcal{EL}$ TBox $\mathcal{T}'$ with $\int(\mathcal{T}') \leq k$ such that $\mathcal{T}' \in [\mathcal{T}]_{\mathcal{ELD}}$.*

The following inclusion relations between the above introduced notions hold:

$$[\mathcal{T}] \subseteq [\mathcal{T}]_{\mathcal{EL}} \subseteq [\mathcal{T}]_{\mathcal{ELD}}$$

$$[\mathcal{T}] \subseteq [\mathcal{T}]_{\sqcup} \subseteq [\mathcal{T}]_{\mathcal{ELD}}$$

In the following, we show that problems **P1-P2** are NP-complete, while the two problems involving $\sqcup$- and $\mathcal{ELD}$-shortcuts (**P3-P4**) are between NP and Σ_2^P.

4 Inclusion in NP resp. Σ_2^P

In this section, we investigate the upper complexity bound for the problems **P1 -P4** and show that **P1 -P2** are in NP and **P3 -P4** in Σ_2^P. In case of **P1**, showing the upper bound is simple:

Theorem 1. **P1** *is in* NP.

Proof. We ask the non-deterministic algorithm to guess such an equivalent TBox $\mathcal{T}' \equiv \mathcal{T}$ of size $\leq k$. Then, we check $\mathcal{T}' \equiv \mathcal{T}$ in PTIME [5]. □

The inclusion of **P2** in NP (and of **P3** and **P4** in Σ_2^P) is less straightforward, since deciding inseparability of $\mathcal{EL}$ TBoxes is known to be EXPTIME-complete and emulation is even undecidable [14]. For the inclusion of **P3** and **P4** in Σ_2^P, we make use of the following simple lemma:

Lemma 2. *Let $\mathcal{T}$ be an $\mathcal{ELD}$ TBox. $T_{\mathcal{EL}}(\mathcal{T}) \models^{em} \mathcal{T}$.*

Proof Sketch. We can transform each model of $\mathcal{T}$ into a model of $\mathcal{T}' = T_{\mathcal{EL}}(\mathcal{T})$ by successively adding $A^{\mathcal{I}} = \bigcup_{i=1}^{n} C_i^{\mathcal{I}}$ for each concept A that is introduced to replace the disjunction $\bigsqcup_{i=1}^{n} C_i$. Additionally, it can be show that $T_{\mathcal{EL}}(\mathcal{T}) \models \mathcal{T}$ holds, since $\{\bigsqcup_{i=1}^{n} C_i \sqsubseteq A\} \equiv \{C_i \sqsubseteq A \mid i \in \{1, \ldots, n\}\}$ and "$\sqsubseteq$" is transitive. □

Theorem 2. **P2** *is in* NP *and* **P3**, **P4** *are in* Σ_2^P.

Proof. Let $\mathcal{T}'$ the corresponding equivalent of an $\mathcal{EL}$ TBox $\mathcal{T}$ returned by the non-deterministic algorithm of size $\leq k$. Now we consider how to verify that $\mathcal{T}'$ indeed fulfills the requirements stated in Definitions 4,7,9.

For **P2**, we have to verify Conditions 2-4 of Definition 4, which clearly can be done in polynomial time. In order to verify Condition 1 ($\mathcal{T}' \equiv_{\Sigma}^{\mathcal{EL}} \mathcal{T}$), it is sufficient to insert the shortcut definitions into $\mathcal{T}$ and then test the equivalence of this extended TBox $\mathcal{T}_{\mathtt{ext}}$ and $\mathcal{T}'$ for the following reasons: By Lemma 1, $\mathcal{T}_{\mathtt{ext}} \models^{em} \mathcal{T}$. Due to transitivity of $\equiv_{\Sigma}^{\mathcal{EL}}$, $\mathcal{T}_{\mathtt{ext}} \equiv \mathcal{T}'$ implies $\mathcal{T}' \equiv_{\Sigma}^{\mathcal{EL}} \mathcal{T}$. It remains to show that $\mathcal{T}_{\mathtt{ext}} \equiv \mathcal{T}'$ only if $\mathcal{T}' \equiv_{\Sigma}^{\mathcal{EL}} \mathcal{T}$. Let us assume for contradiction that there exists an inclusion axiom $C \sqsubseteq D \in \mathcal{T}'$ such that $\mathcal{T}_{\mathtt{ext}} \not\models C \sqsubseteq D$. Then we can obtain concepts C', D' with $\mathrm{sig}(C') \cup \mathrm{sig}(D') \subseteq \Sigma$ by recursively replacing shortcuts by their definitions such that $\mathcal{T}_{\mathtt{ext}} \not\models C' \sqsubseteq D'$ and $\mathcal{T}_{\mathtt{ext}} \models C' \sqsubseteq D'$. With, $\mathcal{T} \equiv_{\Sigma}^{\mathcal{EL}} \mathcal{T}_{\mathtt{ext}}$ we can conclude $\mathcal{T} \not\equiv_{\Sigma}^{\mathcal{EL}} \mathcal{T}'$. ↯

For **P3**, we need to show that there exists an $\mathcal{ELD}$ TBox $\mathcal{T}''$ such that $\mathrm{sig}(\mathcal{T}'') \subseteq \mathrm{sig}(\mathcal{T}), \mathcal{T}' = T_{\mathcal{EL}}(\mathcal{T}'')$ and $\mathcal{T}'' \equiv \mathcal{T}$. In case there exists such $\mathcal{T}''$, we can obtain it from $\mathcal{T}'$ by replacing the introduced concept symbols by the corresponding disjunctions of definitions. As $\mathcal{T}'$ and $\mathcal{T}''$ are Σ-inseparable with $\Sigma = \mathrm{sig}(\mathcal{T}'')$ by Lemma 2, it suffices to show $\mathcal{T}' \models \mathcal{T}$ and $\mathcal{T} \models \mathcal{T}''$. The first is standard reasoning in $\mathcal{EL}$ and can clearly be performed in polynomial time. The refutation of the latter, i.e., showing $\mathcal{T} \not\models \mathcal{T}''$, can be done in NP: if $\mathcal{T} \not\models \mathcal{T}''$, then, for some concretization $C \sqsubseteq D$ of some axiom of $\mathcal{T}'$ (where a concretization is simply the replacement of each disjunction by one of its disjuncts) $\mathcal{T} \not\models C \sqsubseteq D$ holds. A non-deterministic machine can simply guess the axiom and its concretization. Consequently, testing $\mathcal{T} \models \mathcal{T}''$ is in CoNP and **P3** thus in Σ_2^P. (Note that it suffices to call the oracle once at the end.)

For **P4**, we can simply combine these tests. □

Clearly, the results of this section also apply to tractable extensions of $\mathcal{EL}$.

5 NP-Hardness of P1- P4

In this section, we show the NP-hardness of problems **P1** through **P4** by a reduction from the set cover problem, which is one of the standard NP-complete problems. For a given set $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$ with carrier set $S = \bigcup_{i=1}^{n} S_i$, a *cover* $\mathcal{C} \subseteq \mathcal{S}$ is a subset of $\mathcal{S}$, such that the union of the sets in $\mathcal{C}$ covers S, i.e., $S = \bigcup_{C \in \mathcal{C}} C$.

The *set cover problem* is the problem to determine, for a given set $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$ and a given integer k, if there is a cover $\mathcal{C}$ of $\mathcal{S}$ with at most $k \geq |\mathcal{C}|$ elements.

We will use a restricted version of the set cover problem, which we call the *dense set cover problem* (DSCP). In the dense set cover problem, we require that

– neither the carrier set S nor the empty set is in $\mathcal{S}$,
– all singleton subsets (sets with exactly one element) of S are in $\mathcal{S}$, and
– if a non-singleton set S is in $\mathcal{S}$, so is some subset $S' \subseteq S$, which contains only one element less than S ($|S \setminus S'| = 1$).

Lemma 3. *The dense set cover problem is NP-complete.*

Proof. Inclusion in NP is inherited from the set cover problem, of which it is a special instance.

We now reduce solving the set cover problem to solving the dense set cover problem. We start with a set cover problem for a given $\mathcal{S}$ and k, and first check if the carrier set S is contained in $\mathcal{S}$ (if so, the problem is solved). If it is not the case, we identify the size l of the largest set in $\mathcal{S}$, initialise $\mathcal{S}'$ to $\mathcal{S}$ and extend $\mathcal{S}'$ using the following algorithm:

– while $l > 1$ do
  • for all $S \in \mathcal{S}'$, choose an $s \in S$ and join $\mathcal{S}'$ with $S \setminus \{s\}$
  • decrement l by one.

After this, we join $\mathcal{S}$ with $\{\{s\} \mid s \in S\}$, and remove the empty set from $\mathcal{S}$ if applicable. Note that $\mathcal{S}'$ can easily be constructed in polynomial time. Now we show that there is a cover $\mathcal{C}$ of size $\leq k$ of $\mathcal{S}$ exactly if there is a cover $\mathcal{C}'$ of size $\leq k$ of $\mathcal{S}'$. W.l.o.g., we can assume that $\emptyset \notin \mathcal{C}$, since we always obtain a cover from any cover $\mathcal{C}$ by removing $\emptyset$ from it. Since $\mathcal{S} \subseteq \mathcal{S}' \cup \{\emptyset\}$, any cover of $\mathcal{S}$ is a cover of $\mathcal{S}'$. Let $\mathcal{C}'$ be a cover of size $\leq k$ of $\mathcal{S}'$. We can construct a cover $\mathcal{C}$ of $\mathcal{S}$ by replacing each $S' \in \mathcal{C}'$ by the corresponding superset $S \in \mathcal{S}$. □

Given the above NP-completeness result, we show that the size of minimal equivalents specified in **P1** through **P4** is a linear function of the size of the minimal cover. To this end, we use the lemma below to obtain a lower bound on the size of equivalents. Intuitively, it states that for each entailed non-trivial equivalence $C \equiv A$, the TBox must contain at least one axiom that is at least as large as $C' \equiv A$ for some C' with $\mathcal{T} \models C \equiv C'$:

Lemma 4. *Let $\mathcal{T}$ be an $\mathcal{EL}$ TBox, $A \in sig(\mathcal{T})$ and C, D $\mathcal{EL}$ concepts such that $\mathcal{T} \models C \equiv A$, $\mathcal{T} \models A \sqsubseteq D$ (the latter is required for induction). Then, one of the following is true:*

1. *A is a conjunct of C (including the case $C = A$);*
2. *there exists an $\mathcal{EL}$ concept C' such that $\mathcal{T} \models C \equiv C'$ and $C' \bowtie A \in \mathcal{T}$ or $C' \bowtie A \sqcap D' \in \mathcal{T}$ for some $\bowtie \in \{\equiv, \sqsubseteq\}$ and some concept D'.*

Proof Sketch. For the full version of the proof, see extended version of the paper. We use the sound and complete proof system for general subsumption in $\mathcal{EL}$ terminologies introduced in [8] and prove the lemma by induction on the depth of the derivation of $C \sqsubseteq A \sqcap D$. We assume that the proof has minimal depth and consider the possible rules that could have been applied last to derive $C \sqsubseteq A \sqcap D$. In each case the lemma holds. □

The encoding of the dense set cover problem as **P1**-**P4** is as follows.

Consider an instance of the dense set cover problem with the carrier set $A = \{B_1, \ldots, B_n\}$, the set $\mathcal{S} = \{A_1, \ldots, A_m, \{B_1\}, \ldots, \{B_n\}\}$ of subsets that can be used to form a cover. By interpreting the set and element names as atomic concepts, we can construct $\mathcal{T}_{\mathcal{S}\mathtt{base}}$ as follows:

$$\mathcal{T}_{\mathcal{S}\mathtt{base}} = \{A'' \equiv A' \sqcap B \mid A'', A' \in \mathcal{S}, B \in A, A'' = A' \cup \{B\}, A'' \neq A'\}.$$

Observe that the size of $\mathcal{T}_{\mathcal{S}\mathtt{base}}$ is at least $3m$. Clearly, $\mathcal{T}_{\mathcal{S}\mathtt{base}} \models A_i \equiv \bigsqcap_{B \in A_i} B$. Let $\mathcal{T}_{\mathcal{S}} = \mathcal{T}_{\mathcal{S}\mathtt{base}} \cup \{A \equiv \bigsqcap_{B \in A} B\}$. We establish the connection between the size of $\mathcal{T}_{\mathcal{S}}$ equivalents and the size of the cover of $\mathcal{S}$ as follows:

Lemma 5. *$\mathcal{T}_{\mathcal{S}}$ has an equivalent (as specified in* **P1**-**P4***) of size $\int(\mathcal{T}_{\mathcal{S}\mathtt{base}}) + k + 1$ if, and only if, $\mathcal{S}$ has a cover of size k.*

Proof. For the if-direction, assume that $\mathcal{S}$ has a cover of size k. We construct $\mathcal{T}'_{\mathcal{S}}$ of size $\int(\mathcal{T}_{\mathcal{S}\mathtt{base}}) + k + 1$ as follows: $\mathcal{T}'_{\mathcal{S}} = \mathcal{T}_{\mathcal{S}\mathtt{base}} \cup \{A \equiv \bigsqcap_{A' \in \mathcal{C}} A'\}$. Clearly, $\mathcal{T}'_{\mathcal{S}} \equiv \mathcal{T}_{\mathcal{S}}$. Note that $\mathcal{T}'_{\mathcal{S}} \in [\mathcal{T}_{\mathcal{S}}]$ and, therefore, also $\mathcal{T}'_{\mathcal{S}} \in [\mathcal{T}_{\mathcal{S}}]_{\sqcup}, [\mathcal{T}_{\mathcal{S}}]_{\mathcal{EL}}, [\mathcal{T}_{\mathcal{S}}]_{\mathcal{ELD}}$.

For the only-if-direction, we assume that k is minimal and argue that no equivalent $\mathcal{T}' \in [\mathcal{T}_{\mathcal{S}}]_{\mathcal{ELD}}$ of size $\leq \int(\mathcal{T}_{\mathcal{S}\mathtt{base}}) + k$ can exist. Assume that $\mathcal{T}$ is a minimal TBox with $\mathcal{T} \in [\mathcal{T}_{\mathcal{S}}]_{\mathcal{ELD}}$. With the observation, that the $m + n$ atomic concepts that represent elements of $\mathcal{S}$ are pairwise not equivalent with each other or the concept A that represents the carrier set, we can conclude that no two atomic concepts are equivalent. From Lemma 4 it follows that, for each A_i with $i \in \{1, \ldots, m\}$, there is an axiom $C_i \equiv C'_i \in \mathcal{T}$ or $C_i \sqsubseteq C'_i \in \mathcal{T}$ such that $\mathcal{T} \models C_i \equiv A_i$ and A_i is a conjunct of C'_i or $A_i = C'_i$. Since there are no equivalent atomic concepts and $C_i \neq A_i$ due to the minimality of $\mathcal{T}$, the size of each such axiom is at least 3 and none of these axioms coincide. We will later make use of two obvious properties (*) of these axioms:

1. since $\mathcal{T}_{\mathcal{S}} \not\models A_i \sqsubseteq A$, A cannot occur as a conjunct of C_i or as a conjunct of C'_i;
2. these axioms cannot be (parts of) the definitions of atomic concepts representing disjunctions (as A_i is a conjunct of C'_i) or shortcuts ($\mathcal{T} \models C_i \equiv A_i$).

Finally, we estimate the size of the remaining axioms and show that their cumulative size is $> k$. It also follows from Lemma 4 that there exists an axiom $C \equiv C' \in \mathcal{T}$ or $C \sqsubseteq C' \in \mathcal{T}$ such that $\mathcal{T} \models C \equiv A$ and A is a conjunct of C' or $A = C'$. It holds that $\mathcal{T} \models C \equiv \bigsqcap_{B \in A} B$. We also know that for no proper subset $S' \subsetneq A$ it holds that $\mathcal{T} \models \bigsqcap_{B \in S'} B \sqsubseteq C$.

If C does not contain any shortcuts or disjunction replacements, then we have found a cover of S and the size of the axiom must be $\geq k+1$. Assume that it contains auxiliary shortcut and disjunction concepts and let C' be the concept obtained by replacing all these concepts recursively in C until $\mathrm{sig}(C') \subseteq \mathrm{sig}(\mathcal{T}_{\mathcal{S}})$. It is clear that the cumulative size of the corresponding definitions for these auxiliary concept symbols cannot be smaller than the size of C', which does not contain any concept symbols twice. Since $\mathcal{T} \models C' \equiv C$, we have once more found a cover of S and the size of this axiom plus the size of definition axioms must be $\geq k + 1$. From the two properties (*) of the axioms definition A_i we can conclude that none of these axioms can coincide. Thus, the overall size of $\mathcal{T}$ must be $\geq \int(\mathcal{T}_{\mathcal{S}\mathtt{base}}) + k + 1$. □

Theorem 3. P1 *through* **P4** *are NP-hard.*

Proof. The theorem is an immediate consequence of Lemma 5. It establishes that all four problems can be used to solve the dense set cover problem, which is NP-complete according to Lemma 3. □

Thus, we establish completeness of the first two problems:

Theorem 4. P1 *and* **P2** *are NP-complete.* □

6 Summary and Outlook

In this paper, we have considered the problem of finding minimal equivalent representations for ontologies expressed in the lightweight description logic $\mathcal{EL}$ that forms a basis of some large ontologies used in practice. We have shown that the task of finding such a representation (or rather: its related decision problem) is NP-complete.

In addition to studying the problem of computing minimal equivalent TBoxes, we investigated the task of finding minimal representations for ontologies under signature extension. We considered scenarios, where auxiliary concepts are allowed to be used as shortcuts for complex $\mathcal{EL}$ concepts. We showed that this task is also NP-complete. For the corresponding decision problem with auxiliary concepts acting as shortcuts for a disjunction of $\mathcal{EL}$ concepts, we have established NP-hardness and inclusion in Σ_2^P. The same bounds hold for the combination of the two ways of extending the signature.

There are various natural extensions of this work. The results obtained within this paper can easily be transferred to the context of ontology reuse, where a sub-signature becomes obsolete in a new context and a compact representation of the facts about the remaining terms is sought-after. Recent results on ontology reuse show that neither uniform interpolation nor standard module extraction guarantee the optimality of the extracted ontology [16].

Further, a question that naturally arises is that of tight complexity bounds when shortcuts for disjunctions are allowed for. Another target would be the complexity of identifying minimal TBoxes by the means of an arbitrary inseparable TBox, where we waive the requirement of explicitly defining the meaning of new concepts. An EXPTIME upper bound for this problem is implied from the fact that the set of candidate TBoxes is exponential, and so is the general test for inseparability in $\mathcal{EL}$.

Minimizing representations is, of course, an interesting problem for all logics, and similar questions can (and should) be asked for more expressive ontology languages.

While the concern of this paper is the complexity of the above problems, a natural follow-up task would be to develop efficient algorithms and tools that support ontology engineers in the development of succinct representations of their ontologies. Natural targets would be good heuristics and efficient approximations. For the latter, our proofs contain the bad news that there is no linear approximation scheme, as the set cover problem has no logarithmic approximations unless P equals NP.

Finally, from practical point of view, it would be very interesting to investigate the potential improvement of succinctness in existing medical ontologies. Such a case study can be carried out after the corresponding tool support becomes available.

Acknowledgments This work is supported by the EPSRC grant EP/H046623/1 'Synthesis and Verification in Markov Game Structures' and the University of Oxford.

References

1. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.: The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press (2003)
2. Baader, F., Küsters, R., Molitor, R.: Rewriting concepts using terminologies. In: Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR 2000). (2000) 297–308
3. Grimm, S., Wissmann, J.: Elimination of redundancy in ontologies. In: Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011). (2011) 260–274
4. Bienvenu, M.: Prime implicates and prime implicants: From propositional to modal logic. Journal of Artificial Intelligence Research (JAIR) **36** (2009) 71–128
5. Baader, F., Brandt, S., Lutz, C.: Pushing the $\mathcal{EL}$ envelope. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005). (2005) 364–369
6. Motik, B., Cuenca Grau, B., Horrocks, I., Wu, Z., Fokoue, A., Lutz, C., eds.: OWL 2 Web Ontology Language: Profiles. W3C Recommendation (27 October 2009) Available at http://www.w3.org/TR/owl2-profiles/.
7. OWL Working Group, W.: OWL 2 Web Ontology Language: Document Overview. W3C Recommendation (27 October 2009) Available at http://www.w3.org/TR/owl2-overview/.
8. Nikitina, N., Rudolph, S.: ExpExpExplosion: Uniform interpolation in general EL terminologies. In: Proceedings of the 20th European Conference on Artificial Intelligence (ECAI 2012). (2012) 618–623
9. Nikitina, N., Schewe, S.: More is Sometimes Less: Succinctness in $\mathcal{EL}$. Techreport, Department of Computer Science, University of Oxford, Oxford (Mai 2013)
10. Ghilardi, S., Lutz, C., Wolter, F.: Did I Damage my Ontology? A Case for Conservative Extensions in Description Logics. In: Proceedings of the 10th International Conference on the Principles of Knowledge Representation and Reasoning (KR 2006). (2006) 187–197
11. Lutz, C., Walther, D., Wolter, F.: Conservative extensions in expressive description logics. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007). (2007) 453–458
12. Konev, B., Lutz, C., Walther, D., Wolter, F.: Semantic modularity and module extraction in description logics. In: Proceedings of the 18th European Conference on Artificial Intelligence (ECAI 2008). (2008) 55–59
13. Konev, B., Lutz, C., Walther, D., Wolter, F.: Formal properties of modularisation. In Stuckenschmidt, H., Parent, C., Spaccapietra, S., eds.: Modular Ontologies. Springer-Verlag (2009) 25–66
14. Lutz, C., Wolter, F.: Deciding inseparability and conservative extensions in the description logic $\mathcal{EL}$. Journal of Symbolic Computation **45**(2) (2010) pp.194–228
15. Kontchakov, R., Wolter, F., Zakharyaschev, M.: Can you tell the difference between dl-lite ontologies? In: Proceedings of the 11th International Conference on Principles of Knowledge Representation and Reasoning (KR 2008). (2008) 285–295
16. Nikitina, N., Glimm, B.: Hitting the sweetspot: Economic rewriting of knowledge bases. In: Proceedings of the 11th International Semantic Web Conference (ISWC 2012). (2012) 394–409