Effective computation of maximal sound approximations of Description Logic ontologies

Marco Console, Jose Mora, Riccardo Rosati, Valerio Santarelli, Domenico Fabio Savo

Dipartimento di Ingegneria Informatica Automatica e Gestionale Antonio Ruberti
Sapienza Università di Roma
lastname@dis.uniroma1.it

Abstract. We study the problem of approximating Description Logic (DL) ontologies specified in a source language $\mathcal{L}_S$ in terms of a less expressive target language $\mathcal{L}_T$. This problem is getting very relevant in practice: e.g., approximation is often needed in ontology-based data access systems, which are able to deal with ontology languages of a limited expressiveness. We first provide a general, parametric, and semantically well-founded definition of maximal sound approximation of a DL ontology. Then, we present an algorithm that is able to effectively compute two different notions of maximal sound approximation according to the above parametric semantics when the source ontology language is OWL 2 and the target ontology language is OWL 2 QL. Finally, we experiment the above algorithm by computing the two OWL 2 QL approximations of a large set of existing OWL 2 ontologies. The experimental results allow us both to evaluate the effectiveness of the proposed notions of approximation and to compare the two different notions of approximation in real cases.

1 Introduction

Description Logic (DL) ontologies are the core element of ontology-based data access (OBDA) [15], in which the ontology is utilized as a conceptual view, allowing user access to the underlying data sources. In OBDA, as well as in all the current applications of ontologies requiring automated reasoning, a trade-off between the expressiveness of the ontology specification language and the complexity of reasoning in such a language must be reached. More precisely, most of the current research and development in OBDA is focusing on languages for which reasoning, and in particular query answering, is not only tractable (in data complexity) but also *first-order* rewritable [2,5]. This imposes significant limitations on the set of constructs and axioms allowed in the ontology language.

The limited expressiveness of the current ontology languages adopted in OBDA provides a strong motivation for studying the approximation of ontologies formulated in (very) expressive languages with ontologies in low-complexity languages such as OWL 2 QL. Such a motivation is not only theoretical, but also practical, given the current availability of OBDA systems and the increasing interest in applying the OBDA approach in the real world [1,6,7,16]: for instance, ontology approximation is currently one of the main issues in the generation of ontologies for OBDA within the use cases of the Optique EU project.[1]

Several approaches have recently dealt with the problem of approximating Description Logic ontologies. These can roughly be partitioned in two types: *syntactic* and *semantic*. In the former, only the syntactic form of the axioms of the original ontology is considered, thus those axioms which do not comply with the syntax of the target ontology language are disregarded [17,18]. This approach generally can be performed quickly and through simple algorithms. However, it does not, in general, guarantee soundness, i.e., to infer only correct entailments, or completeness, i.e., all entailments of the original ontology that are also expressible in the target language are preserved [14]. In the latter, the object of the approximation are the entailments of the original ontology, and the goal is to preserve as much as possible of these entailments by means of an ontology in the target language, guaranteeing soundness of the result. On the other hand, this approach often necessitates to perform complex reasoning tasks over the ontology, possibly resulting significantly slower. For these reasons, the semantic approach to ontology approximation poses a more interesting but more complex challenge.

In this paper, we study the problem of approximating DL ontologies specified in a source language $\mathcal{L}_s$ in terms of a less expressive target language $\mathcal{L}_t$. We deal with this problem by first providing a general, parametric, and semantically well-founded definition of maximal sound approximation of a DL ontology. Our semantic definition captures and generalizes previous approaches to ontology approximation [4,8,11,14]. In particular, our approach builds on the preliminary work presented in [8], which proposed a similar, although non-parameterized, notion of maximal sound approximation.

Then, we present an algorithm that is able to effectively compute two different notions of maximal sound approximation according to the above parametric semantics, when the source ontology language is OWL 2 and the target ontology language is OWL 2 QL. In particular, we focus on the *local semantic approximation (LSA)* and the *global semantic approximation (GSA)* of a source ontology. These two notions of approximation correspond to the cases when the parameter of our semantics is set, respectively, to its minimum and to its maximum. Informally, the LSA of an ontology is obtained by considering (and reasoning over) one axiom α of the source ontology at a time, so this technique tries to approximate α independently of the rest of the source ontology. On the contrary, the GSA tries to approximate the source ontology by considering all its axioms (and reasoning over such axioms) at the same time. As a consequence, the GSA is potentially able to "approximate better" than the LSA, while the LSA appears in principle computationally less expensive than the GSA. Notably, in the case of OWL 2 QL, the GSA corresponds to the notion of approximation given in [14], which has been shown to be very well-suited for query answering purposes.

Finally, we experiment the above algorithm by computing the LSA and the GSA in OWL 2 QL of a large set of existing OWL 2 ontologies. The experimental results allow us both to evaluate the effectiveness of the proposed notions of approximation and to compare the two different notions of approximation in real cases. In particular, the main properties pointed out by our experimental results are the following:

[1] http://optique-project.eu

1. the computation of the LSA is usually less expensive than computing the GSA of a given source ontology;
2. in many cases, both the LSA and the GSA of an ontology are very good approximations of the ontology, in the sense that the approximated ontologies actually entail a large percentage of the axioms of the source ontology;
3. in many cases, the LSA and the GSA coincide. This and the previous property imply that the computationally less expensive LSA is usually already able to compute a high-quality sound approximation of the source ontology.

The paper is structured in the following way. First, in Section 2 we recall DL ontology languages, in particular OWL 2 and OWL 2 QL. Then, in Section 3 we present our formal, parameterized notion of semantic sound approximation of an ontology, and illustrate some general properties of such a notion. In Section 4 we present the techniques for computing the GSA and the LSA of OWL 2 ontologies in OWL 2 QL. Finally, we present an experimental evaluation of the above techniques in Section 5, and draw some conclusions in Section 6.

2 Preliminaries

Description Logics (DLs) [3] are logics that allow one to represent the domain of interest in terms of *concepts*, denoting sets of objects, *value-domains*, denoting sets of values, *attributes*, denoting binary relations between objects and values, and *roles* denoting binary relations over objects.

In this paper we consider the DL $\mathcal{SROIQ}$ [10], which is the logic underpinning OWL 2, as the "maximal" DL considered in this paper.

Let Σ be a signature of symbols for individual (object and value) constants and predicates, i.e., concepts, value-domains, attributes, and roles. Let Φ be the set of all $\mathcal{SROIQ}$ axioms over Σ.

An *ontology* over Σ is a finite subset of Φ.

A *DL language* over Σ (or simply *language*) $\mathcal{L}$ is a set of ontologies over Σ. We call *$\mathcal{L}$-ontology* any ontology $\mathcal{O}$ such that $\mathcal{O} \in \mathcal{L}$. Moreover, we denote by $\Phi_{\mathcal{L}}$ the set of axioms $\bigcup_{\mathcal{O}\in\mathcal{L}} \mathcal{O}$.

We call a language $\mathcal{L}$ *closed* if $\mathcal{L} = 2^{\Phi_{\mathcal{L}}}$. As we will see in the following, there exist both closed and non-closed DL languages among the standard ones.

The semantics of an ontology is given in terms of first-order (FOL) interpretations (cf. [3]). We denote with $Mod(\mathcal{O})$ the set of models of $\mathcal{O}$, i.e., the set of FOL interpretations that satisfy all the axioms in $\mathcal{O}$ (we recall that every $\mathcal{SROIQ}$ axiom corresponds to a first-order sentence). As usual, an ontology $\mathcal{O}$ is said to be *satisfiable* if it admits at least one model, and $\mathcal{O}$ is said to *entail* a First-Order Logic (FOL) sentence α, denoted $\mathcal{O} \models \alpha$, if $\alpha^{\mathcal{I}} = true$ for all $\mathcal{I} \in Mod(\mathcal{O})$. Moreover, given two ontologies $\mathcal{O}$ and $\mathcal{O}'$, we say that $\mathcal{O}$ and $\mathcal{O}'$ are logically equivalent if $Mod(\mathcal{O}) = Mod(\mathcal{O}')$.

In this work we will mainly focus on two specific languages, which are OWL 2, the official ontology language of the World Wide Web Consortium (W3C) [9], and one of its profiles, OWL 2 QL [12]. Due to the limitation of space, here we do not provide a complete description of OWL 2, and refer the reader to the official W3C specification [13].

We now present the syntax of OWL 2 QL. We use the German notation for describing OWL 2 QL constructs and axioms, and refer the reader to [12] for the OWL functional style syntax.

Expressions in OWL 2 QL are formed according to the following syntax:

$$
\begin{array}{ll}
B \longrightarrow A \mid \exists Q \mid \delta_F(U) \mid \top_C \mid \bot_C & E \longrightarrow \rho(U) \\
C \longrightarrow B \mid \neg B \mid \exists Q.A & F \longrightarrow \top_D \mid T_1 \mid \cdots \mid T_n \\
Q \longrightarrow P \mid P^- \mid \top_P \mid \bot_P & V \longrightarrow U \mid \top_A \mid \bot_A \\
R \longrightarrow Q \mid \neg Q & W \longrightarrow V \mid \neg V
\end{array}
$$

where: A, P, and U are symbols denoting respectively an *atomic concept*, an *atomic role*, and an *atomic attribute*; P^- denotes the inverse of P; $\exists Q$, also called *unqualified existential role*, denotes the set of objects related to some object by the role Q; $\delta_F(U)$ denotes the *qualified domain* of U with respect to a value-domain F, i.e., the set of objects that U relates to some value in F; $\rho(U)$ denotes the *range* of U, i.e., the set of values related to objects by U; $T_1, \ldots, T_n$ denote n unbounded value-domains (i.e., datatypes); the concept $\exists Q.A$, or *qualified existential role*, denotes the *qualified domain* of Q with respect to A, i.e., the set of objects that Q relates to some instance of A. $\top_C$, $\top_P$, $\top_A$, and $\top_D$ denote, respectively, the universal concept, role, attribute, and value-domain, while $\bot_C$, $\bot_P$, and $\bot_A$ denote, respectively, the empty concept, role, and attribute.

An OWL 2 QL ontology $\mathcal{O}$ is a finite set of axioms of the form:

$$
\begin{array}{c}
B \sqsubseteq C \qquad Q \sqsubseteq R \qquad U \sqsubseteq V \qquad E \sqsubseteq F \qquad ref(P) \qquad irref(P) \\
A(a) \qquad P(a,b) \qquad U(a,v)
\end{array}
$$

From left to right, the first four above axioms denote subsumptions between concepts, roles, attributes, and value-domains, respectively. The fifth and sixth axioms denote reflexivity and irreflexivity on roles. The last three axioms denote membership of an individual to a concept, membership of a pair of individuals to a role, and membership of a pair constituted by an individual and a value to an attribute.

From the above definition, it immediately follows that OWL 2 QL is a closed language. On the other hand, we recall that OWL 2 is not a closed language. This is due to the fact that OWL 2 imposes syntactic restrictions that concern the simultaneous presence of multiple axioms in the ontology (for instance, there exist restrictions on the usage of role names appearing in role inclusions in the presence of the role chaining constructor).

3 Approximation

In this section, we illustrate our notion of approximation in a target language $\mathcal{L}_T$ of an ontology $\mathcal{O}_S$ in a language $\mathcal{L}_S$.

Typically, when discussing approximation, one of the desirable properties is that of soundness. Roughly speaking, when the object of approximation is a set of models, this property requires that the set of models of the approximation is a superset of those of the original ontology. Another coveted characteristic of the computed ontology is that it be the "best" approximation of the original ontology. In other words, the need of keeping a minimal distance between the original ontology and the ontology resulting from its approximation is commonly perceived.

On the basis of these observations, the following definition of approximation in a target language $\mathcal{L}_T$ of a satisfiable $\mathcal{L}_S$-ontology is very natural.

Definition 1. *Let $\mathcal{O}_S$ be a satisfiable $\mathcal{L}_S$-ontology, and let $\Sigma_{\mathcal{O}_S}$ be the set of predicate and constant symbols occurring in $\mathcal{O}_S$. An $\mathcal{L}_T$-ontology $\mathcal{O}_T$ over $\Sigma_{\mathcal{O}_S}$ is a* global semantic approximation *(GSA) in $\mathcal{L}_T$ of $\mathcal{O}_S$ if both the following statements hold:*

(i) $Mod(\mathcal{O}_S) \subseteq Mod(\mathcal{O}_T)$;
(ii) there is no $\mathcal{L}_T$-ontology $\mathcal{O}'$ over $\Sigma_{\mathcal{O}_S}$ such that $Mod(\mathcal{O}_S) \subseteq Mod(\mathcal{O}') \subset Mod(\mathcal{O}_T)$.

We denote with $globalApx(\mathcal{O}_S, \mathcal{L}_T)$ the set of all the GSAs in $\mathcal{L}_T$ of $\mathcal{O}_S$.

In the above definition, statement (i) imposes the soundness of the approximation, while statement (ii) imposes the condition of "closeness" in the choice of the approximation.

We observe that an $\mathcal{L}_T$-ontology which is the GSA in $\mathcal{L}_T$ of $\mathcal{O}_S$ may not exist. This is the case when, for each $\mathcal{L}_T$ ontology $\mathcal{O}'_T$ satisfying statement (i) of Definition 1, there always exists an $\mathcal{L}_T$-ontology $\mathcal{O}''_T$ which satisfies statement (i), but for which we have that $Mod(\mathcal{O}_S) \subseteq Mod(\mathcal{O}''_T) \subset Mod(\mathcal{O}'_T)$.

The following lemma provides a sufficient condition for the existence of the GSA in a language $\mathcal{L}_T$ of an ontology $\mathcal{O}_S$.

Lemma 1. *Given a language $\mathcal{L}_T$ and a finite signature Σ, if the set of non-equivalent axioms in $\Phi_{\mathcal{L}_T}$ that one can generate over Σ is finite, then for any $\mathcal{L}_S$-ontology $\mathcal{O}_S$ $globalApx(\mathcal{O}_S, \mathcal{L}_T) \neq \emptyset$.*

In cases where GSAs exist, i.e., $globalApx(\mathcal{O}_S, \mathcal{L}_T) \neq \emptyset$, given two ontologies $\mathcal{O}'$ and $\mathcal{O}''$ in $globalApx(\mathcal{O}_S, \mathcal{L}_T)$, they may be either logically equivalent or not. The condition of non-equivalence is due to the fact that the language in which the original ontology is approximated is not closed. We have the following lemma.

Lemma 2. *Let $\mathcal{L}_T$ be a closed language, and let $\mathcal{O}_S$ be an ontology. For each $\mathcal{O}'$ and $\mathcal{O}''$ belonging to $globalApx(\mathcal{O}_S, \mathcal{L}_T)$, we have that $\mathcal{O}'$ and $\mathcal{O}''$ are logically equivalent.*

Proof. Towards a contradiction, suppose that $Mod(\mathcal{O}') \neq Mod(\mathcal{O}'')$. From this, and from Definition 1 we have that $Mod(\mathcal{O}') \not\subset Mod(\mathcal{O}'')$ and $Mod(\mathcal{O}'') \not\subset Mod(\mathcal{O}')$. Hence, there exist axioms α and β in $\Phi_{\mathcal{L}_T}$ such that $\mathcal{O}' \models \alpha$ and $\mathcal{O}'' \not\models \alpha$, and $\mathcal{O}'' \models \beta$ and $\mathcal{O}' \not\models \beta$. Since both $\mathcal{O}'$ and $\mathcal{O}''$ are sound approximations of $\mathcal{O}_S$, $\mathcal{O}_S \models \{\alpha, \beta\}$. Because $\mathcal{L}_T$ is closed, the ontology $\mathcal{O}'_\beta = \mathcal{O}' \cup \{\beta\}$ is an $\mathcal{L}_T$-ontology. From the above considerations it directly follows that $Mod(\mathcal{O}_S) \subseteq Mod(\mathcal{O}'_\beta) \subset Mod(\mathcal{O}')$. This means that $\mathcal{O}'$ does not satisfy condition (ii) of Definition 1, and therefore $\mathcal{O}' \notin globalApx(\mathcal{O}_S, \mathcal{L}_T)$, which is a contradiction. The same conclusion can be reached analogously for $\mathcal{O}''$. ∎

In other words, if the target language is closed, Lemma 2 guarantees that, up to logical equivalence, the GSA is unique.

Definition 1 is non-constructive, in the sense that it does not provide any hint as to how to compute the approximation in $\mathcal{L}_T$ of an ontology $\mathcal{O}_S$. The following theorem suggests more constructive conditions, equivalent to those in Definition 1, but first we need to introduce the notion of *entailment set* [14] of a satisfiable ontology with respect to a language.

Definition 2. *Let $\Sigma_{\mathcal{O}}$ be the set of predicate and constant symbols occurring in $\mathcal{O}$, and let $\mathcal{L}'$ be a language. The* entailment set *of $\mathcal{O}$ with respect to $\mathcal{L}'$, denoted as $\mathsf{ES}(\mathcal{O}, \mathcal{L}')$, is the set of axioms from $\Phi_{\mathcal{L}'}$ that only contain predicates and constant symbols from $\Sigma_{\mathcal{O}}$ and that are entailed by $\mathcal{O}$.*

In other words, we say that an axiom α belongs to the entailment set of an ontology $\mathcal{O}$ with respect to a language $\mathcal{L}'$, if α is an axiom in $\Phi_{\mathcal{L}'}$ built over the signature of $\mathcal{O}$ and for each interpretation $\mathcal{I} \in Mod(\mathcal{O})$ we have that $\mathcal{I} \models \alpha$.

Clearly, given an ontology $\mathcal{O}$ and a language $\mathcal{L}'$, the entailment set of $\mathcal{O}$ with respect to $\mathcal{L}'$ is unique.

Theorem 1. *Let $\mathcal{O}_S$ be a satisfiable $\mathcal{L}_S$-ontology and let $\mathcal{O}_T$ be a satisfiable $\mathcal{L}_T$-ontology. We have that:*

(a) $Mod(\mathcal{O}_S) \subseteq Mod(\mathcal{O}_T)$ if and only if $\mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T) \subseteq \mathsf{ES}(\mathcal{O}_S, \mathcal{L}_T)$;
(b) there is no $\mathcal{L}_T$-ontology $\mathcal{O}'$ such that $Mod(\mathcal{O}_S) \subseteq Mod(\mathcal{O}') \subset Mod(\mathcal{O}_T)$ if and only if there is no $\mathcal{L}_T$-ontology $\mathcal{O}''$ such that $\mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T) \subset \mathsf{ES}(\mathcal{O}'', \mathcal{L}_T) \subseteq \mathsf{ES}(\mathcal{O}_S, \mathcal{L}_T)$.

Proof. We start by focusing on the first statement. ($\Leftarrow$) Suppose, by way of contradiction, that $\mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T) \subseteq \mathsf{ES}(\mathcal{O}_S, \mathcal{L}_T)$ and that $Mod(\mathcal{O}_S) \not\subseteq Mod(\mathcal{O}_T)$. This means that there exists at least one interpretation that is a model for $\mathcal{O}_S$ but not for $\mathcal{O}_T$. Therefore there exists an axiom $\alpha \in \mathcal{O}_T$ such that $\mathcal{O}_S \not\models \alpha$. Since $\mathcal{O}_T$ is an ontology in $\mathcal{L}_T$, then α is an axiom in $\Phi_{\mathcal{L}_T}$. It follows that $\alpha \in \mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T)$ and that $\alpha \notin \mathsf{ES}(\mathcal{O}_S, \mathcal{L}_T)$, which leads to a contradiction.

($\Rightarrow$) Towards a contradiction, suppose that $Mod(\mathcal{O}_S) \subseteq Mod(\mathcal{O}_T)$, but $\mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T) \not\subseteq \mathsf{ES}(\mathcal{O}_S, \mathcal{L}_T)$. This means that there exists at least one axiom $\alpha \in \mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T)$ such that $\alpha \notin \mathsf{ES}(\mathcal{O}_S, \mathcal{L}_T)$. It follows that $\mathcal{O}_T \models \alpha$ while $\mathcal{O}_S \not\models \alpha$, which immediately implies that $Mod(\mathcal{O}_S) \not\subseteq Mod(\mathcal{O}_T)$. Hence we have a contradiction.

Now we prove the second statement. ($\Leftarrow$) By contradiction, suppose that there exists an $\mathcal{L}_T$-ontology $\mathcal{O}'$ such that $Mod(\mathcal{O}_S) \subseteq Mod(\mathcal{O}') \subset Mod(\mathcal{O}_T)$, and that there does not exist any $\mathcal{L}_T$-ontology $\mathcal{O}''$ such that $\mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T) \subset \mathsf{ES}(\mathcal{O}'', \mathcal{L}_T) \subseteq \mathsf{ES}(\mathcal{O}_S, \mathcal{L}_T)$. From what shown before, we have that $Mod(\mathcal{O}_S) \subseteq Mod(\mathcal{O}') \subseteq Mod(\mathcal{O}_T)$ implies that $\mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T) \subseteq \mathsf{ES}(\mathcal{O}', \mathcal{L}_T) \subseteq \mathsf{ES}(\mathcal{O}_S, \mathcal{L}_T)$. Moreover, since both $\mathcal{O}'$ and $\mathcal{O}_T$ are $\mathcal{L}_T$ ontologies, $Mod(\mathcal{O}') \subset Mod(\mathcal{O}_T)$ implies that $\mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T) \neq \mathsf{ES}(\mathcal{O}', \mathcal{L}_T)$. Hence, $\mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T) \subset \mathsf{ES}(\mathcal{O}', \mathcal{L}_T) \subseteq \mathsf{ES}(\mathcal{O}_S, \mathcal{L}_T)$, which contradicts the hypothesis.

($\Rightarrow$) Suppose, by way of contradiction, that there exists an $\mathcal{L}_T$-ontology $\mathcal{O}''$ such that $\mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T) \subset \mathsf{ES}(\mathcal{O}'', \mathcal{L}_T) \subseteq \mathsf{ES}(\mathcal{O}_S, \mathcal{L}_T)$ and there is no $\mathcal{L}_T$-ontology $\mathcal{O}'$

such that $Mod(\mathcal{O}_S) \subseteq Mod(\mathcal{O}') \subset Mod(\mathcal{O}_T)$. From property (a) we have that $Mod(\mathcal{O}_S) \subseteq Mod(\mathcal{O}'') \subseteq Mod(\mathcal{O}_T)$. Since both $\mathcal{O}''$ and $\mathcal{O}_T$ are $\mathcal{L}_T$ ontologies, then $\mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T) \subset \mathsf{ES}(\mathcal{O}'', \mathcal{L}_T)$ implies that $Mod(\mathcal{O}'') \neq Mod(\mathcal{O}_T)$, which directly leads to a contradiction. ∎

From Theorem 1 it follows that every ontology $\mathcal{O}_T$ which is a GSA in $\mathcal{L}_T$ of an ontology $\mathcal{O}_S$ is also an approximation in $\mathcal{L}_T$ of $\mathcal{O}_S$ according to [8], and, as we shall show in the following section, for some languages, this corresponds to the approximation in [14].

As discussed in [8], the computation of a GSA can be a very challenging task even when approximating into tractable fragments of OWL 2 [12]. This means that even though a GSA is one that best preserves the semantics of the original ontology, it currently suffers from a significant practical setback: the outcome of the computation of the approximation is tightly linked to the capabilities of the currently available reasoners for $\mathcal{L}_S$-ontologies. This may lead, in practice, to the impossibility of computing GSAs of very large or complex ontologies when the source language is very expressive.

We observe that the critical point behind these practical difficulties in computing a GSA of an ontology is that, in current implementations, any reasoner for $\mathcal{L}_S$ must reason over the ontology as a whole. From this observation, the idea for a new notion of approximation, in which we do not reason over the entire ontology but only over portions of it, arises. At the basis of this new notion, which we call *k-approximation*, is the idea of obtaining an approximation of the original ontology by computing the global semantic approximation of each set of k axioms of the original ontology in isolation. Below we give a formal definition of the k-approximation.

In what follows, given an ontology $\mathcal{O}$ and a positive integer k such that $k \leq |\mathcal{O}|$, we denote with $subset_k(\mathcal{O})$ the set of all the sets of cardinality k of axioms of $\mathcal{O}$.

Definition 3. *Let $\mathcal{O}_S$ be a satisfiable $\mathcal{L}_S$-ontology and let $\Sigma_{\mathcal{O}_S}$ be the set of predicate and constant symbols occurring in $\mathcal{O}_S$. Let $\mathcal{U}_k = \{\mathcal{O}_i^j \mid \mathcal{O}_i^j \in globalApx(\mathcal{O}_i, \mathcal{L}_T)$, such that $\mathcal{O}_i \in subset_k(\mathcal{O}_S)\}$. An $\mathcal{L}_T$-ontology $\mathcal{O}_T$ over $\Sigma_{\mathcal{O}_S}$ is a* k-approximation *in $\mathcal{L}_T$ of $\mathcal{O}_S$ if both the following statements hold:*

- $\bigcap_{\mathcal{O}_i^j \in \mathcal{U}_k} Mod(\mathcal{O}_i^j) \subseteq Mod(\mathcal{O}_T)$;
- *there is no $\mathcal{L}_T$-ontology $\mathcal{O}'$ over $\Sigma_{\mathcal{O}_S}$ such that* $\bigcap_{\mathcal{O}_i^j \in \mathcal{U}_k} Mod(\mathcal{O}_i^j) \subseteq Mod(\mathcal{O}') \subset Mod(\mathcal{O}_T)$.

The following theorem follows from Theorem 1 and provides a constructive condition for the k-approximation.

Theorem 2. *Let $\mathcal{O}_S$ be a satisfiable $\mathcal{L}_S$-ontology and let $\Sigma_{\mathcal{O}_S}$ be the set of predicate and constant symbols occurring in $\mathcal{O}_S$. An $\mathcal{L}_T$-ontology $\mathcal{O}_T$ over $\Sigma_{\mathcal{O}_S}$ is a k-approximation in $\mathcal{L}_T$ of $\mathcal{O}_S$ if and only if:*

(i) $\mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T) \subseteq \mathsf{ES}(\bigcup_{\mathcal{O}_i \in subset_k(\mathcal{O}_S)} \mathsf{ES}(\mathcal{O}_i, \mathcal{L}_T), \mathcal{L}_T)$;
(ii) *there is no $\mathcal{L}_T$-ontology $\mathcal{O}'$ over $\Sigma_{\mathcal{O}_S}$ such that* $\mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T) \subset \mathsf{ES}(\mathcal{O}', \mathcal{L}_T) \subseteq \mathsf{ES}(\bigcup_{\mathcal{O}_i \in subset_k(\mathcal{O}_S)} \mathsf{ES}(\mathcal{O}_i, \mathcal{L}_T), \mathcal{L}_T)$.

Proof. $(sketch)$ The proof can be easily adapted from the proof of Theorem 1 by observing that in order to prove the theorem one has to show that:
(a) $\bigcap_{\mathcal{O}_i^j \in \mathcal{U}_k} Mod(\mathcal{O}_i^j) \subseteq Mod(\mathcal{O}_T)$ if and only if $\mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T) \subseteq \mathsf{ES}(\bigcup_{\mathcal{O}_i \in subset_k(\mathcal{O}_S)} \mathsf{ES}(\mathcal{O}_i, \mathcal{L}_T), \mathcal{L}_T)$;
(b) and that there is no $\mathcal{L}_T$-ontology $\mathcal{O}'$ over $\Sigma_{\mathcal{O}_S}$ such that $\bigcap_{\mathcal{O}_i^j \in \mathcal{U}_k} Mod(\mathcal{O}_i^j) \subseteq Mod(\mathcal{O}') \subset Mod(\mathcal{O}_T)$ if and only if there is no $\mathcal{L}_T$-ontology $\mathcal{O}''$ over $\Sigma_{\mathcal{O}_S}$ such that $\mathsf{ES}(\mathcal{O}_T, \mathcal{L}_T) \subset \mathsf{ES}(\mathcal{O}'', \mathcal{L}_T) \subseteq \mathsf{ES}(\bigcup_{\mathcal{O}_i \in subset_k(\mathcal{O}_S)} \mathsf{ES}(\mathcal{O}_i, \mathcal{L}_T), \mathcal{L}_T)$. ∎

We note that in the case in which $k = |\mathcal{O}_S|$, the k-approximation actually coincides with the GSA. At the other end of the spectrum, we have the case in which $k = 1$. Here we are treating each axiom α in the original ontology in isolation, i.e., we are considering ontologies formed by a single axiom α. We refer to this approximation as *local semantic approximation* (LSA).

We conclude this section with an example highlighting the difference between the GSA and the LSA.

Example 1. Consider the following OWL 2 ontology $\mathcal{O}$.

$$\begin{array}{rl}\mathcal{O} = \{ & A \sqsubseteq B \sqcup C \quad B \sqsubseteq D \quad A \sqsubseteq \exists R.D \\ & B \sqcap C \sqsubseteq F \quad C \sqsubseteq D \quad \exists R.D \sqsubseteq E \,\}.\end{array}$$

The following ontology is a GSA in OWL 2 QL of $\mathcal{O}$.

$$\begin{array}{rl}\mathcal{O}_{GSA} = \{ & A \sqsubseteq D \quad B \sqsubseteq D \quad A \sqsubseteq \exists R \quad A \sqsubseteq \exists R.D \\ & A \sqsubseteq E \quad C \sqsubseteq D \quad D \sqsubseteq F \;\}.\end{array}$$

Indeed, it is possible to show that, according to Theorem 1, each axiom entailed by $\mathcal{O}_{GSA}$ is also entailed by $\mathcal{O}$, and that it is impossible to build an OWL 2 QL ontology $\mathcal{O}'$ such that $\mathsf{ES}(\mathcal{O}_{GSA}, OWL\,2\,QL) \subset \mathsf{ES}(\mathcal{O}', OWL\,2\,QL) \subseteq \mathsf{ES}(\mathcal{O}, OWL\,2\,QL)$.

Computing the LSA in OWL 2 QL of $\mathcal{O}$, i.e., its 1-approximation in OWL 2 QL, we obtain the following ontology.

$$\begin{array}{rl}\mathcal{O}_{LSA} = \{ & B \sqsubseteq D \quad A \sqsubseteq \exists R \\ & C \sqsubseteq D \quad A \sqsubseteq \exists R.D \,\}.\end{array}$$

It is easy to see that $Mod(\mathcal{O}) \subset Mod(\mathcal{O}_{GSA}) \subset Mod(\mathcal{O}_{LSA})$, which means that the ontology $\mathcal{O}_{GSA}$ approximates $\mathcal{O}$ better than $\mathcal{O}_{LSA}$. This expected result is a consequence of the fact that reasoning over each single axiom in $\mathcal{O}$ in isolation does not allow for the extraction all the OWL 2 QL consequences of $\mathcal{O}$.

Moreover, from Lemma 2, it follows that every $\mathcal{O}' \in globalApx(\mathcal{O}_S, OWL\,2\,QL)$ is logically equivalent to $\mathcal{O}_{GSA}$. □

4 Approximation in OWL 2 QL

In this section we deal with the problem of approximating ontologies in OWL 2 with ontologies in OWL 2 QL.

Based on the characteristics of the OWL 2 QL language, we can give the following theorem.

Algorithm 1: $computeKApx(\mathcal{O}, k)$

Input: a satisfiable OWL 2 ontology $\mathcal{O}$, a positive integer k such that $k \leq |\mathcal{O}|$
Output: an OWL 2 QL ontology $\mathcal{O}_{Apx}$
begin
  $\mathcal{O}_{Apx} \leftarrow \emptyset$;
  foreach ontology $\mathcal{O}_i \in subset_k(\mathcal{O}_S)$
    $\mathcal{O}_{Apx} \leftarrow \mathcal{O}_{Apx} \cup \mathsf{ES}(\mathcal{O}_i, OWL\,2\,QL)$;
  return $\mathcal{O}_{Apx}$;
end

Theorem 3. *Let $\mathcal{O}_S$ be a satisfiable OWL 2 ontology. Then the OWL 2 QL ontology $\bigcup_{\mathcal{O}_i \in subset_k(\mathcal{O}_S)} \mathsf{ES}(\mathcal{O}_i, OWL\,2\,QL)$ is the k-approximation in OWL 2 QL of $\mathcal{O}_S$.*

Proof. $(sketch)$ To prove the claim, we observe that Lemma 1 holds for OWL 2 QL ontologies, and this guarantees that for every OWL 2 ontology $\mathcal{O}_S$, there exists at least one OWL 2 QL ontology which is its GSA, i.e., $globalApx(\mathcal{O}_S, OWL\ 2\ QL) \neq \emptyset$. Moreover, we have that since OWL 2 QL is closed, for Lemma 2, all ontologies in $\mathsf{ES}(\mathcal{O}_S, OWL\ 2\ QL)$ are pairwise logically equivalent. Another consequence of the fact that OWL 2 QL is closed is that, whichever language the original ontology $\mathcal{O}_S$ is expressed in, $\mathsf{ES}(\mathcal{O}_S, OWL\,2\,QL)$ is an OWL 2 QL ontology. Furthermore, given a set of OWL 2 QL ontologies, the union of these ontologies is still an OWL 2 QL ontology. From these observations, it is easy to see that, given an OWL 2 ontology $\mathcal{O}_S$ and an integer $k \leq |\mathcal{O}_S|$, the set $\bigcup_{\mathcal{O}_i \in subset_k(\mathcal{O}_S)} \mathsf{ES}(\mathcal{O}_i, OWL\,2\,QL)$ satisfies conditions (i) and (ii) of Theorem 2. Hence, we have the claim. ■

Notably, we observe that for $k = |\mathcal{O}_S|$ the k-approximation $\mathcal{O}_T$ in OWL 2 QL of $\mathcal{O}_S$ is unique and coincides with its entailment set in OWL 2 QL. This means that $\mathcal{O}_T$ is also the approximation in OWL 2 QL of $\mathcal{O}_S$ according to the notion of approximation presented in [14]. Therefore, all the properties that hold for the semantics in [14] also hold for the GSA. In particular, the evaluation of a conjunctive query q without non-distinguished variables over $\mathcal{O}_S$ coincides with the evaluation of q over $\mathcal{O}_T$ (Theorem 5 in [14]).

From Theorem 3, one can easily come up with Algorithm 1 for computing the k-approximation of an $\mathcal{L}_S$-ontology $\mathcal{O}_S$ in OWL 2 QL. The algorithm first computes every subset with size k of the original ontology $\mathcal{O}_S$. Then, it computes the ontology which is the result of the k-approximation in OWL 2 QL of the ontology in input as the union of the entailment sets with respect to OWL 2 QL of each such subset. A naive algorithm for computing the entailment set with respect to OWL 2 QL can be easily obtained from the one given in [14] for *DL-Lite* languages. We can summarize it as follows. Let $\mathcal{O}$ be an ontology and let $\Sigma_{\mathcal{O}}$ be the set of predicate and constant symbols occurring in $\mathcal{O}$. The algorithm first computes the set Γ of axioms in $\Phi_{OWL\ 2\ QL}$ which can be built over $\Sigma_{\mathcal{O}}$, and then, for each axiom $\alpha \in \Gamma$ such that $\mathcal{O} \models \alpha$, adds α to the set $\mathsf{ES}(\mathcal{O}, OWL\,2\,QL)$. In practice, to check if $\mathcal{O} \models \alpha$ one can use an OWL 2 reasoner.

Since each invocation of the OWL 2 reasoner is N2EXPTIME, the computation of the entailment set can be very costly [4].

A more efficient technique for its computation is given in [8], where the idea is to limit the number of invocations to the OWL 2 reasoner by exploiting the knowledge acquired through a preliminary exploration of the ontology. To understand the basic idea behind this technique, consider, for example, an ontology $\mathcal{O}$ that entails the inclusions $A_1 \sqsubseteq A_2$ and $P_1 \sqsubseteq P_2$, where A_1 and A_2 are concepts and P_1 and P_2 are roles. Exploiting these inclusions we can deduce the hierarchical structure of the general concepts that can be built on these four predicates. For instance, we know that $\exists P_2.A_2 \sqsubseteq \exists P_2$, that $\exists P_2.A_1 \sqsubseteq \exists P_2.A_2$, that $\exists P_1.A_1 \sqsubseteq \exists P_2.A_1$, and so on. To obtain the entailed inclusion axioms, we begin by invoking the OWL 2 reasoner, asking for the children of the general concepts which are in the highest position in the hierarchy. So we first compute the subsumees of $\exists P_2$ through the OWL 2 reasoner. If there are none, we avoid invoking the reasoner asking for the subsumees of $\exists P_2.A_2$ and so on. Regarding the entailed disjointness axioms, we follow the same approach but starting from the lowest positions in the hierarchy.

The following theorem establishes correctness and termination of algorithm $computeKApx$.

Theorem 4. *Let $\mathcal{O}_S$ be a satisfiable OWL 2 ontology. $computeKApx(\mathcal{O}_S, k)$ terminates and computes the k-approximation in OWL 2 QL of $\mathcal{O}_S$.*

Proof. $(sketch)$ Termination of $computeKApx(\mathcal{O}_S, k)$ directly follows from the fact that $\mathcal{O}_S$ is a finite set of axioms and that, for each $\mathcal{O}_i \in subset_k(\mathcal{O}_S)$, $\mathsf{ES}(\mathcal{O}_i, OWL\ 2\ QL)$ can be computed in finite time. The correctness of the algorithm directly follows from Theorem 3.

5 Experiments

In this section we present the experimental tests that we have performed for the approximation of a suite of OWL 2 ontologies into OWL 2 QL through the two notions of approximation we have introduced earlier.

We notice that by choosing a value for k different from $|\mathcal{O}_S|$, the computation of the entailment set becomes easier. However, observing Algorithm 1, the number of times that this step must be repeated can grow very quickly. In fact, the number of sets of axioms in $subset_k(\mathcal{O}_S)$ is equal to the binomial coefficient of $|\mathcal{O}_S|$ over k, and therefore for large ontologies this number can easily become enormous, and this can be in practice a critical obstacle in the computation of the k-approximation.

For this reason, in these experiments we have focused on comparing the GSA (k-approximation with $k = |\mathcal{O}_S|$) to the LSA (k-approximation with $k = 1$), and we reserve the study of efficient techniques for k-approximation with $1 < k < |\mathcal{O}_S|$ for future works. Furthermore, to provide a standard baseline against which to compare the results of the GSA and LSA, we have compared both our approaches with a syntactic sound approximation approach, consisting in first normalizing the axioms in the ontology and then eliminating the ones that are not syntactically compliant with OWL 2 QL. We will refer to this approach as “SYNT”.

Ontology	Expressiveness	Axioms	Concepts	Roles	Attributes	OWL2 QL Axioms
Homology	$\mathcal{ALC}$	83	66	0	0	83
Cabro	$\mathcal{ALCHIQ}$	100	59	13	0	99
Basic vertebrate anatomy	$\mathcal{SHIF}$	108	43	14	0	101
Fungal anatomy	$\mathcal{ALEI}+$	115	90	5	0	113
Pmr	$\mathcal{ALU}$	163	137	16	0	159
Ma	$\mathcal{ALE}+$	168	166	8	0	167
General formal Ontology	$\mathcal{SHIQ}$	212	45	41	0	167
Cog analysis	$\mathcal{SHIF}(\mathcal{D})$	224	92	37	9	213
Time event	$\mathcal{ALCROIQ}(\mathcal{D})$	229	104	28	7	170
Spatial	$\mathcal{ALEHI}+$	246	136	49	0	155
Translational medicine	$\mathcal{ALCRIF}(\mathcal{D})$	314	225	18	6	298
Biopax	$\mathcal{SHIN}(\mathcal{D})$	391	69	55	41	240
Vertebrate skeletal anatomy	$\mathcal{ALER}+$	455	314	26	0	434
Image	$\mathcal{S}$	548	624	2	0	524
Protein	$\mathcal{ALCF}(\mathcal{D})$	691	45	50	133	490
Pizza	$\mathcal{SHOIN}$	712	100	8	0	660
Ontology of physics for biology	$\mathcal{ALCHIQ}(\mathcal{D})$	954	679	33	3	847
Plant trait	$\mathcal{ALE}+$	1463	1317	4	0	1461
Dolce	$\mathcal{SHOIN}(\mathcal{D})$	1667	209	313	4	1445
Ont. of athletic events	$\mathcal{ALEH}$	1737	1441	15	1	1722
Neomark	$\mathcal{ALCHQ}(\mathcal{D})$	1755	55	105	488	842
Pato	$\mathcal{SH}$	1979	2378	36	0	1779
Protein Modification	$\mathcal{ALE}+$	1986	1338	8	0	1982
Po anatomy	$\mathcal{ALE}+$	2128	1294	11	0	2064
Lipid	$\mathcal{ALCHIN}$	2375	716	46	0	2076
Plant	$\mathcal{S}$	2615	1644	16	0	2534
Mosquito anatomy	$\mathcal{ALE}+$	2733	1864	5	0	2732
Idomal namespace	$\mathcal{ALER}+$	3467	2597	24	0	3462
Cognitive atlas	$\mathcal{ALC}$	4100	1701	6	0	3999
Genomic	$\mathcal{ALCQ}$	4322	2265	2	0	3224
Mosquito insecticide resistance	$\mathcal{ALE}+$	4413	4362	21	0	4409
Galen-A	$\mathcal{ALEHIF}+$	4979	2748	413	0	3506
Ni gene	$\mathcal{SH}$	8835	4835	8	0	8834
Fyp	$\mathcal{SH}$	15105	4751	69	0	12924
Fly anatomy	$\mathcal{SH}$	20356	8064	72	0	20353
EL-Galen	$\mathcal{ALEH}+$	36547	23136	950	0	25138
Galen full	$\mathcal{ALEHIF}+$	37696	23141	950	0	25613
Pr reasoned	$\mathcal{S}$	46929	35470	14	0	40031
Snomed fragment for FMA	$\mathcal{ALER}$	73543	52635	52	0	35004
Gene	$\mathcal{SH}$	73942	39160	12	0	73940
FMA OBO	$\mathcal{ALE}+$	119560	75139	2	0	119558

Table 1: Characteristics of the ontologies used in the GSA and LSA tests.

The suite of ontologies used during testing contains 41 ontologies and was assembled from the Bioportal ontology repository[2]. The ontologies that compose this suite were selected to test the scalability of our approaches both to larger ontologies and to ontologies formulated in more expressive languages. In Table 1 we present the most relevant metrics of these ontologies.

All tests were performed on a DELL Latitude E6320 notebook with Intel Core i7-2640M 2.8Ghz CPU and 4GB of RAM, running Microsoft Windows 7 Premium operating system, and Java 1.6 with 2GB of heap space. Timeout was set at eight hours, and execution was aborted if the maximum available memory was exhausted. The tool used in the experiments and the suite of ontologies are available at http://diag.uniroma1.it/~mora/ontology_approximation/iswc2014/.

As mentioned in Section 4, the computation of the entailment set involves the use of an external OWL 2 reasoner. Therefore, the performance and the results of the computed approximations are greatly effected by the choice of the reasoner. For our tests, we have used the Pellet[3] OWL 2 reasoner (version v.2.3.0.6).

In Table 2 we present the results of the evaluation. An analysis of these results leads to the following observations.

Firstly, we were able to compute the GSA for 26 out of the 41 tested ontologies. For the remaining fifteen, this was not possible, either due to the size of the ontology, in terms of the number of its axioms, e.g., the FMA 2.0 or Gene ontologies, which have more than seventy thousand and one hundred thousand axioms, respectively, or due to its high expressivity, e.g., the Dolce ontology or the General formal ontology. The LSA approach is instead always feasible, it is quicker than the GSA approach for all but one of the tested ontologies, and it is overall very fast: no ontology took more than 250 seconds to approximate with the LSA.

Secondly, it is interesting to observe the comparison between the quality of the approximation that one can obtain through the LSA with respect to that obtained through the GSA. This relationship answers the question of whether the ontology obtained through the LSA (the "LSA ontology") is able to capture a significant portion of the one obtained through the GSA (the "GSA ontology"). Our tests in fact confirm that this is the case: out of the 26 ontologies for which we were able to compute the GSA, in only four cases the LSA ontology entails less than 60 percent of the axioms of the GSA ontology, while in twenty cases it entails more than 90 percent of them. The average percentage of axioms in the original ontologies entailed by the GSA ontologies is roughly 80 percent, and of the axioms of the GSA ontologies entailed by the LSA ontologies is roughly 87 percent.

Furthermore, the LSA provides a good approximation even for those ontologies for which the GSA is not computable. In fact, Table 3 shows the percentage of axioms of the original ontology that are entailed by the LSA ontology. Out of the twelve ontologies for which we were able to obtain this value (the remaining three ontologies caused an "out of memory" error), only in three cases it was less than 60 percent, while in four cases it was higher than 80 percent. These results are particularly interesting with respect to those ontologies for which the GSA approach is not feasible due to their

[2] http://bioportal.bioontology.org/
[3] http://clarkparsia.com/pellet/

Ontology	GSA axioms	GSA entails original (%)	LSA axioms	LSA entails GSA (%)	SYNT axioms	SYNT entails GSA (%)	SYNT entails LSA (%)	GSA time (s)	LSA time (s)
Homology	83	100	83	100	83	100	100	1	4
Cabro	233	96	121	100	100	100	100	4	2
Basic vertebrate anatomy	192	93	141	97	71	56	67	3	3
Fungal anatomy	318	98	140	69	113	69	100	2	2
Pmr	162	97	159	98	159	98	100	2	2
Ma	411	99	240	95	167	96	100	4	4
General formal ontology	–	–	286	–	177	–	100	–	6
Cog analysis	104407	75	474	46	215	1	82	36	7
Time event	93769	71	662	99	196	1	58	45	11
Spatial	510	63	371	86	155	42	52	9	4
Translational medicine	4089	86	505	99	275	30	64	19	7
Biopax	2182057	–	3217	–	251	–	81	–	11
Vertebrate skeletal anatomy	9488	95	581	92	434	57	99	27	5
Image	1016	95	596	98	571	98	100	178	5
Protein	–	–	10789	–	475	–	88	–	20
Pizza	2587	91	755	92	678	92	99	7	4
Ont. of physics for biology	1789821	–	1505	–	1241	–	100	–	7
Plant trait	2370	99	1496	99	1461	100	100	10	9
Dolce	–	–	2959	–	1555	–	100	–	8
Ontology of athletic events	5073	99	2392	99	1731	92	100	42	9
Neomark	–	–	39807	–	1723	–	63	–	50
Pato	4066	89	2209	100	1976	78	99	99	18
Protein Modification	2195	99	2001	100	1982	100	100	12	19
Po anatomy	11486	96	2783	77	2078	78	100	455	18
Lipid	14659	87	3165	97	2759	89	97	47	10
Plant	18375	96	3512	80	2574	81	100	929	15
Mosquito anatomy	21303	99	4277	43	2732	44	100	214	16
Idomal namespace	67417	99	4259	98	3461	59	100	496	16
Cognitive atlas	7449	97	5324	100	1364	26	30	162	17
Genomic	–	–	86735	–	85037	–	98	–	54
Mosquito insecticide res.	6794	99	4502	100	4409	100	100	86	14
Galen-A	–	–	8568	–	4971	–	90	–	26
Ni gene	46148	99	10415	90	8834	91	100	472	32
Fyp	–	–	19675	–	11800	–	82	–	43
Fly anatomy	460849	99	28436	67	20346	67	100	25499	45
EL-Galen	–	–	70272	–	43804	–	89	–	59
Galen full	–	–	72172	–	44279	–	89	–	61
Pr reasoned	–	–	56085	–	47662	–	100	–	93
Snomed fragment for FMA	–	–	140629	–	101860	–	76	–	250
Gene	–	–	86292	–	73940	–	100	–	178
FMA OBO	–	–	143306	–	119558	–	100	–	113

Table 2: Results of the GSA, LSA, and SYNT. The values represent, from left to right, the number of axioms in the ontology obtained through the GSA, the percentage of axioms of the original ontology that are entailed by the ontology obtained through the GSA, the number of axioms in the ontology obtained through the LSA, the percentage of axioms of the ontology obtained through the GSA that are entailed by the LSA, the number of axioms in the ontology obtained by the SYNT, the percentage of axioms of the ontology obtained through the GSA that are entailed by the ontology obtained through the SYNT, the percentage of axioms of the ontology obtained through the LSA that are entailed by the ontology obtained through the SYNT, and finally the GSA time and the LSA time (both in seconds).

Ontology	Original axioms	LSA axioms	LSA entails original (%)	LSA time (s)
General formal ontology	212	264	67	6
Biopax	391	3204	53	11
Protein	691	10720	47	20
Ontology of physics for biology	954	1074	75	7
Dolce	1667	2914	78	8
Neomark	1755	38966	46	50
Genomic	4322	9844	65	54
Galen-A	4979	8568	70	26
Fyp	15105	19672	85	43
EL-Galen	36547	70272	–	59
Galen full	37696	72172	–	61
Pr reasoned	46929	55391	83	93
SNOMED fragment for FMA	73543	140629	–	250
Gene	73942	86289	99	178
FMA OBO	119560	143306	99	113

Table 3: LSA results for ontologies for which the GSA is not computable.

complexity, as is the case for example for the Dolce ontology, for Galen-A, and for the Ontology of physics for biology. Indeed, even though these ontologies are expressed in highly expressive DL languages, the structure of the axioms that compose them is such that reasoning on each of them in isolation does not lead to much worse approximation results than reasoning on the ontology as a whole: for the nine smallest ontologies in Table 3, for which the GSA fails not because of the size of the ontology, the average percentage is 68.6.

Finally, both the GSA and LSA compare favorably against the syntactic sound approximation approach. In fact, the average percentage of axioms in the LSA and GSA ontologies that are entailed by the ontologies obtained through the SYNT approach are respectively roughly 90 percent and 72 percent. While the latter result is to be expected, the former is quite significant, even more so when one considers that the LSA is very fast. Indeed, a “gain” of 10 percent of axiom entailments by the LSA with respect to the SYNT in the case of large ontologies such as Galen and Snomed translates to tens of thousands of preserved axioms in very little computation time.

In conclusion, the results gathered from these tests corroborate the usefulness of both the global semantic approximation and the local semantic approximation approaches. The former provides a maximal sound approximation in the target language of the original approach, and is in practice computable in a reasonable amount of time for the majority of the tested ontologies. The latter instead represents a less optimal but still very effective solution for those ontologies for which the GSA approach goes beyond the capabilities of the currently-available ontology reasoners.

6 Conclusions

In this paper we have addressed the problem of ontology approximation in Description Logics and OWL, presenting (i) a parameterized semantics for computing sound approximations of ontologies, (ii) algorithms for the computation of approximations (the GSA and the LSA) of OWL 2 ontologies in OWL 2 QL, and (iii) an extensive experimental evaluation of the above techniques, which empirically proves the validity of our approach.

The present work can be extended in several ways. First, while we have focused on sound approximations, it would be interesting to also consider complete approximations of ontologies. Also, we would like to study the development of techniques for k-approximations different from GSA and LSA, i.e., for k such that $1 < k < |\mathcal{O}_S|$, as well as to analyze the possibility of integrating ontology module extraction techniques in our approach. Then, this work has not addressed the case when there are differences in the semantic assumptions between the source and the target ontology languages. For instance, differently from OWL 2 and its profiles, some DLs (e.g., $\textit{DL-Lite}_A$ [15]) adopt the Unique Name Assumption (UNA). This makes our approach not directly applicable, for instance, if we consider OWL 2 as the source language and $\textit{DL-Lite}_A$ as the target language. The reason is that the UNA implies some axioms (inequalities between individuals) that can be expressed in OWL 2 but cannot be expressed in $\textit{DL-Lite}_A$. We aim at extending our approach to deal with the presence of such semantic discrepancies in the ontology languages. Finally, we are very interested in generalizing our approach to a full-fledged ontology-based data access scenario [15], in which data sources are connected through declarative mappings to the ontology. In that context, it might be interesting to use both the ontology and the mappings in the target OBDA specification to approximate a given ontology in the source OBDA specification.

Acknowledgments. This research has been partially supported by the EU under FP7 project Optique (grant n. FP7-318338).

References

1. Natalia Antonioli, Francesco Castanò, Cristina Civili, Spartaco Coletta, Stefano Grossi, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, Domenico Fabio Savo, and Emanuela Virardi. Ontology-based data access: the experience at the Italian Department of Treasury. In *Proc. of the Industrial Track of CAiSE 2013*, volume 1017 of *CEUR*, `ceur-ws.org`, pages 9–16, 2013.
2. Alessandro Artale, Diego Calvanese, Roman Kontchakov, and Michael Zakharyaschev. The *DL-Lite* family and relations. *J. of Artificial Intelligence Research*, 36:1–69, 2009.
3. Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2003.
4. Elena Botoeva, Diego Calvanese, and Mariano Rodriguez-Muro. Expressive approximations in DL-Lite ontologies. *Proc. of AIMSA 2010*, pages 21–31, 2010.
5. Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family. *J. of Automated Reasoning*, 39(3):385–429, 2007.

6. Diego Calvanese, Martin Giese, Peter Haase, Ian Horrocks, Thomas Hubauer, Yannis E. Ioannidis, Ernesto Jiménez-Ruiz, Evgeny Kharlamov, Herald Kllapi, Johan W. Klüwer, Manolis Koubarakis, Steffen Lamparter, Ralf Möller, Christian Neuenstadt, T. Nordtveit, Özgür L. Özçep, Mariano Rodriguez-Muro, Mikhail Roshchin, Domenico Fabio Savo, Michael Schmidt, Ahmet Soylu, Arild Waaler, and Dmitriy Zheleznyakov. Optique: OBDA Solution for Big Data. In *ESWC (Satellite Events)*, volume 7955 of *Lecture Notes in Computer Science*, pages 293–295. Springer, 2013.
7. Cristina Civili, Marco Console, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Lorenzo Lepore, Riccardo Mancini, Antonella Poggi, Riccardo Rosati, Marco Ruzzi, Valerio Santarelli, and Domenico Fabio Savo. MASTRO STUDIO: Managing ontology-based data access applications. *PVLDB*, 6:1314–1317, 2013.
8. Marco Console, Valerio Santarelli, and Domenico Fabio Savo. Efficient approximation in DL-Lite of OWL 2 ontologies. In *Proc. of DL 2013*, volume 1014 of *CEUR Workshop Proceedings*, pages 132–143, 2013.
9. Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F. Patel-Schneider, and Sebastian Rudolph. *OWL 2 Web Ontology Language: Primer*. W3C Recommendation, 2012. Available at http://www.w3.org/TR/2012/REC-owl2-primer-20121211/.
10. Ian Horrocks, Oliver Kutz, and Ulrike Sattler. The even more irresistible $\mathcal{SROIQ}$. In *Proc. of KR 2006*, pages 57–67, 2006.
11. Carsten Lutz, Inanç Seylan, and Frank Wolter. An Automata-Theoretic Approach to Uniform Interpolation and Approximation in the Description Logic EL. In *Proc. of KR 2012*. AAAI Press, 2012.
12. Boris Motik, Bernardo Cuenca Grau, Ian Horrocks, Zhe Wu, Achille Fokoue, and Carsten Lutz. *Owl 2 web ontology language: Profiles (2nd edition)*. W3C Recommendation, 2012. Available at http://www.w3.org/TR/owl2-profiles/.
13. Boris Motik, Bijan Parsia, and Peter F. Patel-Schneider. *OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax*. W3C Recommendation, 2012. Available at http://www.w3.org/TR/2012/REC-owl2-syntax-20121211/.
14. Jeff Z. Pan and Edward Thomas. Approximating OWL-DL ontologies. In *Proc. of AAAI 2007*, pages 1434–1439, 2007.
15. Antonella Poggi, Domenico Lembo, Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, and Riccardo Rosati. Linking data to ontologies. *J. on Data Semantics*, X:133–173, 2008.
16. Mariano Rodriguez-Muro, Roman Kontchakov, and Michael Zakharyaschev. Ontology-based data access: Ontop of databases. In *Proc. of ISWC 2013*, volume 8218 of *LNCS*, pages 558–573. Springer, 2013.
17. Tuvshintur Tserendorj, Sebastian Rudolph, Markus Krötzsch, and Pascal Hitzler. Approximate OWL-reasoning with Screech. In *Proc. of RR 2008*, pages 165–180. Springer, 2008.
18. Holger Wache, Perry Groot, and Heiner Stuckenschmidt. Scalable instance retrieval for the semantic web by approximation. In *Proc. of WISE 2005*, pages 245–254. Springer, 2005.