Module Extraction for Acyclic Ontologies

William Gatens, Boris Konev, and Frank Wolter

The University of Liverpool, UK

Abstract. We present an implementation (AMEX) of a module extraction algorithm for acyclic description logic ontologies. The implementation uses a QBF solver (sKizzo) to check whether one ontology is a conservative extension of another ontology relativised to interpretations of cardinality one. We evaluate AMEX by applying it to NCI (the National Cancer Institute Thesaurus) and by comparing the extracted AMEX-modules with locality-based modules. We also present experiments for a hybrid approach in which AMEX and locality-based module extraction are applied iteratively to NCI.

1 Introduction

Module extraction is the task of computing, given an ontology and a signature Σ of interest, a subset (called module) of the ontology such that for certain applications that use the signature Σ only, the original ontology can be equivalently replaced by the module [14]. In most applications of module extraction it is desirable to compute a small (and, if possible, even minimal) module. In logic-based approaches to module extraction, the most robust and popular way to define modules is via model-theoretic Σ-inseparability, where two ontologies are called Σ-inseparable iff the Σ-reducts of their models coincide. Then, a Σ-module of an ontology is defined as a Σ-inseparable subset of the ontology [10, 7, 3, 8]. It is often helpful and necessary to refine this notion of Σ-module by considering self-contained Σ-modules (modules that are inseparable from the ontology not only w.r.t. Σ but also w.r.t. their own signature) and depleting modules (modules such that the remaining axioms in the ontology say nothing about Σ and the signature of the module, that is, these remaining axioms are inseparable from the empty ontology w.r.t. Σ and the signature of the module). Note that every depleting module is a self-contained module is a module. In all three cases it is often not possible to compute Σ-modules: by results in [8, 11], for acyclic $\mathcal{ALC}$-TBoxes and general $\mathcal{EL}$-TBoxes it is undecidable whether a given subset of a TBox is a (self-contained, depleting) Σ-module. The "maximal" description logics (DLs) for which efficient algorithms computing minimal self-contained and depleting Σ-modules have been developed are acyclic $\mathcal{EL}$ [8] and DL-Lite [9, 10, 6].[1] For this reason, for module extraction for ontologies given in expressive DLs or other expressive ontology languages one has to employ approximation algorithms: instead of computing a minimal (self-contained, depleting) Σ-module, one computes some (self-contained, depleting) Σ-module and

[1] For typical DL-Lite dialects, model-theoretic Σ-inseparability is decidable. Experimental evaluations of module extraction algorithms are, however, available only for language dependent notions of inseparability.

the main research problem is to minimise the size of the module (or, equivalently, to approximate minimal modules). Currently, the most popular and successful approximation algorithm is based on locality and computes so-called $\top\bot^*$-modules [4]. The size of $\top\bot^*$-modules and the performance of algorithms extracting $\top\bot^*$-modules has been analysed systematically and in great detail [4]. However, since no alternative logically sound and implemented module extraction algorithms are available for expressive DLs, it remained open how large and significant the difference between $\top\bot^*$-modules and minimal modules is and in how far it is possible to improve upon the approximation obtained by $\top\bot^*$-modules.[2]

The contribution of this paper is as follows.

1. We extend the module extraction algorithm introduced in [8] from acyclic $\mathcal{ALCI}$-TBoxes to acyclic $\mathcal{ALCQI}$-TBoxes with repeated concept inclusions and present a number of optimisations of the algorithm given in [8]. We note that our extraction algorithm is polynomial time except that it uses a QBF-solver as an oracle.
2. We describe our implementation, called AMEX, of this module extraction algorithm. AMEX is available from `http://www.csc.liv.ac.uk/~wgatens/software/amex.html`.
3. We evaluate its efficiency in experiments with NCI and compare the size of the computed AMEX-modules with the size of $\top\bot^*$-modules.
4. We introduce a hybrid approach to module extraction in which $\top\bot^*$-module extraction and AMEX-module extraction are applied iteratively. Unlike AMEX on its own, this hybrid approach is applicable to arbitrary description logic TBoxes. We demonstrate that on some inputs both AMEX and the hybrid approach lead to significant reductions in the size of modules.

2 Preliminaries

We use standard notation from logic and description logic (DL), details can be found in [1]. In a DL, concepts are constructed from countably infinite sets $\mathsf{N_C}$ of *concept names* and $\mathsf{N_R}$ of *role names* using the concept constructors defined by the DL. For example, *$\mathcal{ALCQI}$-concepts* are built according to the rule

$$C ::= A \quad | \quad \top \quad | \quad \neg C \quad | \quad \geq n\, r.C \quad | \quad \geq n\, r^{-}.C \quad | \quad C \sqcap D,$$

where $A \in \mathsf{N_C}$, n is a natural number, and $r \in \mathsf{N_R}$. As usual, we use the following abbreviations: $\bot$ denotes $\neg\top$, $\exists r.C$ denotes $\geq 1\ r.C$, $\forall r.C$ denotes $\neg\exists\ r.\neg C$, $C \sqcup D$ denotes $\neg(\neg C \sqcap \neg D)$, $\leq\ n\ r.C$ denotes $\neg(\geq\ (n+1)\ r.C)$, and $(=\ n\ r.C)$ for $((\geq\ n\ r.C) \sqcap (\leq\ n\ r.C))$.

A *general TBox* $\mathcal{T}$ is a finite set of *axioms*, where an axiom can be either a *concept inclusion (CI)* $C \sqsubseteq D$ or a *concept equality (CE)* $C \equiv D$, where C and D are concepts. A general TBox $\mathcal{T}$ is *acyclic* if all its axioms are of the form $A \sqsubseteq C$ or $A \equiv C$, where

[2] An implementation of semantic locality-based $\Delta\emptyset^*$-modules and a comparison between $\top\bot^*$ and $\Delta\emptyset^*$-modules have been presented in [4]; however, the authors found no significant difference between the two approaches. A promising approach to refine $\top\bot^*$-module extraction has recently been presented in [12], but an implemented system is not yet publicly available.

$A \in \mathsf{N_C}$, no concept name occurs more than once on the left-hand side and $A \not\prec_{\mathcal{T}}^{+} A$, for any $A \in \mathsf{N_C}$, where $\prec_{\mathcal{T}}^{+}$ is the transitive closure of the relation $\prec_{\mathcal{T}} \subseteq \mathsf{N_C} \times (\mathsf{N_C} \cup \mathsf{N_R})$ defined by setting $A \prec_{\mathcal{T}} X$ iff there exists an axiom of the form $A \sqsubseteq C$ or $A \equiv C$ in $\mathcal{T}$ with $X \in \mathsf{sig}(C)$.

The semantics of DLs is given by *interpretations* $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$, where the *domain* $\Delta^{\mathcal{I}}$ is a non-empty set and $\cdot^{\mathcal{I}}$ is an *interpretation function* that maps each $A \in \mathsf{N_C}$ to a subset $A^{\mathcal{I}}$ of $\Delta^{\mathcal{I}}$ and each $r \in \mathsf{N_R}$ to a binary relation $r^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$. The function $\cdot^{\mathcal{I}}$ is inductively expanded to complex concepts C in the standard way [1]. An interpretation $\mathcal{I}$ *satisfies* a CI $C \sqsubseteq D$ (written $\mathcal{I} \models C \sqsubseteq D$) if $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$, it *satisfies* a CE $C \equiv D$ (written $\mathcal{I} \models C \equiv D$) if $C^{\mathcal{I}} = D^{\mathcal{I}}$. $\mathcal{I}$ is a *model* of $\mathcal{T}$ if it satisfies all axioms in $\mathcal{T}$.

3 Module Extraction

In this section we define depleting modules and give an algorithm computing depleting modules of acyclic $\mathcal{ALCQI}$-TBoxes using a QBF solver. To cover the NCI Thesaurus, we also extend our extraction algorithm to TBoxes that are acyclic except that they contain repeated concept inclusions. The results presented in this section are extensions of the results presented in [8] for acyclic $\mathcal{ALCI}$-TBoxes to acyclic $\mathcal{ALCQI}$-TBoxes with repeated concept inclusions.

A *signature* Σ is a finite subset of $\mathsf{N_C} \cup \mathsf{N_R}$. The signature $\mathsf{sig}(C)$ ($\mathsf{sig}(\alpha)$, $\mathsf{sig}(\mathcal{T})$) of a concept C (axiom α, TBox $\mathcal{T}$, resp.) is the set of concept and role names that occur in C ($\alpha, \mathcal{T}$, resp.). If a $\mathsf{sig}(C) \subseteq \Sigma$ we call C a Σ-concept. The *Σ-reduct* $\mathcal{I}|_\Sigma$ of an interpretation $\mathcal{I}$ is obtained from $\mathcal{I}$ by setting $\Delta^{\mathcal{I}|_\Sigma} = \Delta^{\mathcal{I}}$, and $X^{\mathcal{I}|_\Sigma} = X^{\mathcal{I}}$ for all $X \in \Sigma$, and $X^{\mathcal{I}|_\Sigma} = \emptyset$ for all $X \notin \Sigma$. Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be TBoxes and Σ a signature. Then $\mathcal{T}_1$ and $\mathcal{T}_2$ are *Σ-inseparable*, in symbols $\mathcal{T}_1 \equiv_\Sigma \mathcal{T}_2$, if

$$\{\mathcal{I}|_\Sigma \mid \mathcal{I} \models \mathcal{T}_1\} \;=\; \{\mathcal{I}|_\Sigma \mid \mathcal{I} \models \mathcal{T}_2\}.$$

It is proved in [8] that TBoxes $\mathcal{T}_1$ and $\mathcal{T}_2$ are Σ-inseparable if, and only if, $\mathcal{T}_1 \models \varphi$ iff $\mathcal{T}_2 \models \varphi$ holds for any second-order sentence φ using symbols for Σ only. Thus, Σ-inseparable TBoxes cannot be distinguished by their second-order consequences formulated in Σ. We use Σ-inseparability to define modules.

Definition 1. *Let $\mathcal{M} \subseteq \mathcal{T}$ be TBoxes and Σ a signature. Then $\mathcal{M}$ is a* depleting Σ-module *of $\mathcal{T}$ if $\mathcal{T} \setminus \mathcal{M} \equiv_{\Sigma \cup \mathsf{sig}(\mathcal{M})} \emptyset$.*

Every depleting module $\mathcal{M}$ of $\mathcal{T}$ is inseparable from the $\mathcal{T}$ for its signature [8], that is, if $\mathcal{M}$ is a depleting Σ-module of $\mathcal{T}$ then $\mathcal{T} \equiv_{\Sigma \cup \mathsf{sig}(\mathcal{M})} \mathcal{M}$, and, in particular, $\mathcal{T} \equiv_\Sigma \mathcal{M}$. Thus, a TBox and its depleting Σ-module can be equivalently replaced by each other in applications which concern Σ only. Unfortunately, checking if a subset $\mathcal{M}$ of $\mathcal{T}$ is a depleting Σ-module of $\mathcal{T}$ for some given signature Σ is undecidable already for general TBoxes formulated in $\mathcal{EL}$ and for acyclic $\mathcal{ALC}$-TBoxes [8, 11].

We therefore consider syntactic restrictions that ensure that depleting modules become decidable. We say that an acyclic TBox $\mathcal{T}$ *has a direct Σ-dependency*, for some signature Σ, if there exists $\{A, X\} \subseteq \Sigma$ with $A \prec_{\mathcal{T}}^{+} X$; otherwise we say that $\mathcal{T}$ *has no direct Σ-dependencies*. Although one can construct TBoxes $\mathcal{T}$ and depleting Σ-modules $\mathcal{M}$ of $\mathcal{T}$ such that $\mathcal{T} \setminus \mathcal{M}$ contains direct $\Sigma \cup \mathsf{sig}(\mathcal{M})$-dependencies (see

[8]), for typical depleting Σ-modules $\mathcal{M}$, the set $\mathcal{T} \setminus \mathcal{M}$ should not contain direct $\Sigma \cup \mathsf{sig}(\mathcal{M})$-dependencies because such dependencies indicate a semantic link between two distinct symbols in $\Sigma \cup \mathsf{sig}(\mathcal{M})$. The main advantage of making the assumption that $\mathcal{T} \setminus \mathcal{M}$ has no direct $\Sigma \cup \mathsf{sig}(\mathcal{M})$-dependencies is that it becomes decidable whether $\mathcal{T} \setminus \mathcal{M} \equiv_{\Sigma \cup \mathsf{sig}(\mathcal{M})} \emptyset$ [8]. The following lemma directly implies this decidability result. For an acyclic TBox $\mathcal{T}$ and a signature Σ let

$$\mathsf{Lhs}_\Sigma(\mathcal{T}) = \{A \bowtie C \in \mathcal{T} \mid A \in \Sigma \text{ or } \exists X \in \Sigma\ (X \prec^+_{\mathcal{T}} A)\}.$$

The following is proved in [8] for acyclic $\mathcal{ALCI}$-TBoxes. The generalization to $\mathcal{ALCQI}$ is straightforward and omitted.

Lemma 1. *Let $\mathcal{T}$ be an acyclic $\mathcal{ALCQI}$-TBox. If $\mathcal{T} \setminus \mathcal{M}$ has no direct $\Sigma \cup \mathsf{sig}(\mathcal{M})$-dependencies then the following conditions are equivalent for every $\mathcal{W} \subseteq \mathcal{T} \setminus \mathcal{M}$:*

(a) $\mathcal{W} \equiv_{\Sigma \cup \mathsf{sig}(\mathcal{M})} \emptyset$;
(b) for every $\mathcal{I}$ with $|\Delta^{\mathcal{I}}| = 1$ there exists a model $\mathcal{J}$ of $\mathsf{Lhs}_{\Sigma \cup \mathsf{sig}(\mathcal{M})}(\mathcal{W})$ such that $\mathcal{I}|_{\Sigma \cup \mathsf{sig}(\mathcal{M})} = \mathcal{J}|_{\Sigma \cup \mathsf{sig}(\mathcal{M})}$.

Since the condition (b) of Lemma 1 refers to interpretations with a singleton domain, it can be checked by reduction to validity of a quantified Boolean formula: take a propositional variable p_A for each name $A \in \Sigma \cup \mathsf{sig}(\mathcal{M})$ and a (distinct) propositional variable q_X for each symbol $X \in \mathsf{sig}(\mathcal{T}) \setminus (\Sigma \cup \mathsf{sig}(\mathcal{M}))$. Translate concepts D in the signature $\mathsf{sig}(\mathcal{T})$ into propositional formulas $D^\dagger$ by setting

$$\begin{aligned}
A^\dagger &= p_A && \text{for all } A \in \Sigma \cup \mathsf{sig}(\mathcal{M}) \\
A^\dagger &= q_A && \text{for all } A \in \mathsf{sig}(\mathcal{T}) \setminus (\Sigma \cup \mathsf{sig}(\mathcal{M})) \\
(D_1 \sqcap D_2)^\dagger &= D_1^\dagger \wedge D_2^\dagger \\
(\neg D)^\dagger &= \neg D^\dagger \\
(\geq 1\ r.D)^\dagger = (\geq 1\ r^-.D)^\dagger &= q_r \wedge D^\dagger && \text{for all } r \in \mathsf{sig}(\mathcal{T}) \\
(\geq n\ r.D)^\dagger = (\geq n\ r^-.D)^\dagger &= \bot && \text{for all } n > 1 \text{ and } r \in \mathsf{sig}(\mathcal{T})
\end{aligned}$$

Now let

$$\mathcal{T}^\dagger = \bigwedge_{C \sqsubseteq D \in \mathcal{T} \setminus \mathcal{M}} C^\dagger \to D^\dagger \wedge \bigwedge_{C \equiv D \in \mathcal{T} \setminus \mathcal{M}} C^\dagger \leftrightarrow D^\dagger$$

and let $\boldsymbol{p}$ denote the sequence of variables p_A, $A \in \Sigma \cup \mathsf{sig}(\mathcal{M})$, and $\boldsymbol{q}$ denote the sequence of variables q_X, $X \in \mathsf{sig}(\mathcal{T}) \setminus (\Sigma \cup \mathsf{sig}(\mathcal{M}))$. One can show that condition (b) of Lemma 1 holds if, and only if, the QBF $\varphi_{\mathcal{T}} := \forall \boldsymbol{p} \exists \boldsymbol{q} \mathcal{T}^\dagger$ is valid. Thus, for TBoxes with no direct $\Sigma \cup \mathsf{sig}(\mathcal{M})$-dependencies the separability check can be implemented using a QBF solver.

Lemma 1 can be used directly for a naïve module extraction algorithm which goes through all subsets of $\mathcal{T}$ to identify a smallest possible $\mathcal{M}$ such that $\mathcal{T} \setminus \mathcal{M}$ has no direct $\Sigma \cup \mathsf{sig}(\mathcal{M})$-dependencies and $\mathcal{T} \setminus \mathcal{M} \equiv_{\Sigma \cup \mathsf{sig}(\mathcal{M})} \emptyset$. Instead, we consider a refined goal-oriented approched based on the notion of a *separability causing axiom*. Let $\mathcal{M} \subseteq \mathcal{T}$ and a signature Σ be such that $\mathcal{T} \setminus \mathcal{M}$ has no direct $\Sigma \cup \mathsf{sig}(\mathcal{M})$-dependencies. We call an axiom $A \bowtie C \in \mathcal{T} \setminus \mathcal{M}$, where $\bowtie \in \{\sqsubseteq, \equiv\}$, separability causing if there exists a $\mathcal{W} \subseteq \mathcal{T} \setminus \mathcal{M}$ such that

$$A \bowtie C \in \mathcal{W}; \qquad (\mathcal{W} \setminus \{A \bowtie C\}) \equiv_{\Sigma \cup \mathsf{sig}(\mathcal{M})} \emptyset; \qquad \mathcal{W} \not\equiv_{\Sigma \cup \mathsf{sig}(\mathcal{M})} \emptyset.$$

Input: Acyclic $\mathcal{ALCQI}$ TBox $\mathcal{T}$, Signature Σ
Apply Rules 1 and 2 exhaustively, preferring Rule 1.
Output: (Minimal) Module $\mathcal{M}$ s.t $\mathcal{T} \setminus \mathcal{M} \equiv_{\Sigma \cup \mathsf{sig}(\mathcal{M})} \emptyset$ and $\mathcal{T} \setminus \mathcal{M}$ has no direct $\Sigma \cup \mathsf{sig}(\mathcal{M})$ dependencies.

(R1) If an axiom $A \bowtie C \in \mathcal{T} \setminus \mathcal{M}$ is such that $A \in \Sigma \cup \mathsf{sig}(\mathcal{M}))$ and $A \prec^{+}_{\mathcal{T} \setminus \mathcal{M}} X$, for some $X \in (\Sigma \cup \mathsf{sig}(\mathcal{M}))$, then set $\mathcal{M} := \mathcal{M} \cup \{A \bowtie C\}$

(R2) If an axiom $A \bowtie C \in \mathcal{T} \setminus \mathcal{M}$ is a *separability causing axiom* then set $\mathcal{M} := \mathcal{M} \cup \{A \bowtie C\}$

Fig. 1. Module extraction in $\mathcal{ALCQI}$

Input: TBox $\mathcal{T}$, subset $\mathcal{M} \in \mathcal{T}$ and signature Σ such that
$\mathcal{T} \setminus \mathcal{M}$ contains no direct $\Sigma \cup \mathsf{sig}(\mathcal{M})$-dependencies and $\mathcal{T} \setminus \mathcal{M} \not\equiv_{\Sigma \cup \mathsf{sig}(\mathcal{M})} \emptyset$
Output: Separability causing axiom α

$\mathcal{W} = \mathsf{lastAdded} = \mathit{topHalf}(\mathsf{Lhs}_{\Sigma \cup \mathsf{sig}(\mathcal{M})}(\mathcal{T} \setminus \mathcal{M}))$
$\mathsf{lastRemoved} = \mathit{bottomHalf}(\mathsf{Lhs}_{\Sigma \cup \mathsf{sig}(\mathcal{M})}(\mathcal{T} \setminus \mathcal{M}))$
while $\mathsf{lastAdded} \neq \emptyset$ **do**
    if $\mathcal{W} \equiv_{\Sigma \cup \mathsf{sig}(\mathcal{M})} \emptyset$ **then**
        $\mathsf{lastAdded} = \mathit{topHalf}(\mathsf{lastRemoved})$
        $\mathcal{W} = \mathcal{W} \cup \mathsf{lastAdded}$
        $\mathsf{lastRemoved} = \mathsf{lastRemoved} \setminus \mathsf{lastAdded}$
    else
        $\mathsf{lastRemoved} = \mathit{bottomHalf}(\mathsf{lastAdded})$
        $\mathcal{W} = \mathcal{W} \setminus \mathsf{lastRemoved}$
        $\mathsf{lastAdded} = \mathsf{lastAdded} \setminus \mathsf{lastRemoved}$
return the last axiom of $\mathcal{W}$

Fig. 2. Finding separability causing axiom

Clearly, if $\mathcal{T} \setminus \mathcal{M} \not\equiv_{\Sigma \cup \mathsf{sig}(\mathcal{M})} \emptyset$ then $\mathcal{T} \setminus \mathcal{M}$ contains a separability causing axiom.

The algorithm computing a depleting Σ-module of acyclic $\mathcal{ALCQI}$-TBoxes is now given in Figure 1. In the algorithm, the extraction of depleting Σ-modules is broken into the rules **R1** and **R2**. The rule **R1** checks for direct $\Sigma \cup \mathsf{sig}(\mathcal{M})$-dependencies. The rule **R2** implements an inseparability check. Notice that **R2** only applies when **R1** is not applicable, that is only if $\mathcal{T} \setminus \mathcal{M}$ contains no direct $\Sigma \cup \mathsf{sig}(\mathcal{M})$-dependencies. Notice that applications of the **R1** rule can lead to axioms unnecessarily being included into the module; but such is the price we pay for regaining the decidability of the inseparability check.

To reduce the number of calls to the QBF solver, rule **R2** is implemented as binary search. We first consider $\mathcal{T} \setminus \mathcal{M}$ itself as $\mathcal{W}$. If $\mathcal{T} \setminus \mathcal{M} \equiv_{\Sigma \cup \mathsf{sig}(\mathcal{M})} \emptyset$ then $\mathcal{T} \setminus \mathcal{M}$ contains no separability causing axioms. Otherwise, we consider $\mathcal{W}$ to be equal to the top half of $\mathcal{T} \setminus \mathcal{M}$ (we treat $\mathcal{T} \setminus \mathcal{M}$ as an ordered set). We then check if $\mathcal{W} \equiv_{\Sigma \cup \mathsf{sig}(\mathcal{M})} \emptyset$ and, if this is the case, we grow $\mathcal{W}$ from the bottom and if not, we half it again as shown in

Figure 2. In the worst case we perform $\log_2(|\mathcal{T} \setminus \mathcal{M}|)$ inseparability checks to locate a separability causing axiom.

To summarise, the module extraction algorithm in Figure 1 runs in polynomial time with each call to the QBF solver being treated as a constant time oracle call. Note that QBF solvers have been used before in module extraction [9, 10], but the task solved by the solver here is completely different from its task in [9, 10].

It should be clear that if neither **R1** nor **R2** is applicable then $\mathcal{T} \setminus \mathcal{M} \equiv_{\Sigma \cup \mathsf{sig}(\mathcal{M})} \emptyset$ and so the output of the algorithm in Figure 1 is a depleting Σ-module. By a straightforward generalisation of the results of [8] to $\mathcal{ALCQI}$ one can actually show that the module computed in Figure 1 is uniquely determined:

Theorem 1. *Given an acyclic $\mathcal{ALCQI}$ TBox $\mathcal{T}$ and signature Σ the algorithm in Figure 1 computes the unique minimal depleting Σ-module s.t. $\mathcal{T} \setminus \mathcal{M}$ contains no direct $\Sigma \cup \mathsf{sig}(\mathcal{M})$-dependencies.*

Note that the minimality condition in the theorem means that for any $\mathcal{M}' \subseteq \mathcal{T}$ such that $\mathcal{T} \setminus \mathcal{M}'$ has no direct $\Sigma \cup \mathsf{sig}(\mathcal{M}')$-dependencies and $\mathcal{T} \setminus \mathcal{M}' \equiv_{\Sigma \cup \mathsf{sig}(\mathcal{M}')} \emptyset$ we have $\mathcal{M} \subseteq \mathcal{M}'$. It is, however, still possible that there exists a $\mathcal{M}'' \subseteq \mathcal{T}$ with $\mathcal{T} \setminus \mathcal{M}'' \equiv_{\Sigma \cup \mathsf{sig}(\mathcal{M}'')} \emptyset$, $\mathcal{M} \not\subseteq \mathcal{M}''$ and such that $\mathcal{T} \setminus \mathcal{M}''$ has some direct $\Sigma \cup \mathsf{sig}(\mathcal{M}'')$-dependencies.

Example 1. We apply the algorithm in Figure 1 to the following acyclic TBox $\mathcal{T}$ inspired by the NCI Thesaurus (we have simplified some axioms and abbreviated 'kidney' with K, 'ureter' with U and 'tract' with T)

$$\mathsf{Renal\_Pelvis\_and\_U} \sqsubseteq \exists \mathsf{partOf.K\_and\_U} \quad (1)$$
$$\mathsf{K\_and\_U\_Neoplasm} \equiv \mathsf{U\_T\_Neoplasm} \sqcap (\forall \mathsf{hasSite.K\_and\_U}) \quad (2)$$
$$\mathsf{Malignt\_U\_T\_Neoplasm} \equiv \mathsf{U\_T\_Neoplasm} \sqcap (\forall \mathsf{hasAbnCell.Malignt\_Cell}) \quad (3)$$
$$\mathsf{Benign\_U\_T\_Neoplasm} \equiv \mathsf{U\_T\_Neoplasm} \sqcap (\forall \mathsf{excludesAbnCell.Malignt\_Cell}) \quad (4)$$

and $\Sigma = \{\mathsf{Malignt\_U\_T\_Neoplasm}, \mathsf{K\_and\_U\_Neoplasm}, \mathsf{Renal\_Pelvis\_and\_U}\}$. It can be seen that **R1** is not applicable. To see why $\mathsf{Lhs}_\Sigma(\mathcal{T}) \not\equiv_\Sigma \emptyset$ consider an interpretation $\mathcal{I}$ with $\Delta^\mathcal{I} = \{d\}$ such that $\mathsf{Renal\_Pelvis\_and\_U}^\mathcal{I} = \mathsf{Malignt\_U\_T\_Neoplasm}^\mathcal{I} = \{d\}$ and $\mathsf{K\_and\_U\_Neoplasm}^\mathcal{I} = \emptyset$. It can be readily checked for any $\mathcal{J}$ with $\mathcal{J}|_\Sigma = \mathcal{I}|_\Sigma$ that $\mathcal{J} \not\models \mathcal{T}$. This check can be delegated to a QBF solver as explained above.

The algorithm in Figure 2 splits $\mathsf{Lhs}_\Sigma(\mathcal{T})$ into two parts, $\mathsf{lastAdded} = \{(1), (2)\}$ and $\mathsf{lastRemoved} = \{(3)\}$. For $\mathcal{W} = \mathsf{lastAdded}$ it can be checked that $\mathcal{W} \equiv_\Sigma \emptyset$. Then the algorithm grows $\mathcal{W}$ with (the upper part of) lastRemoved. The same argument as above shows that for $\mathcal{W} = \{(1), (2), (3)\}$ we have $\mathcal{W} \not\equiv_\Sigma \emptyset$ and so the algorithm identifies (3) as a separability causing axiom. After applying the rule **R2**, $\Sigma \cup \mathsf{sig}(\mathcal{M}) = \{\mathsf{Malignt\_U\_T\_Neoplasm}, \mathsf{K\_and\_U\_Neoplasm}, \mathsf{Renal\_Pelvis\_and\_U}, \mathsf{U\_T\_Neoplasm}, \mathsf{hasAbnCell}\}$ and then the rule **R1** adds axioms (1) and (2) to $\mathcal{M}$.

It can be seen that neither **R1** nor **R2** applies to $\mathcal{T} \setminus \mathcal{M} = \{(4)\}$ and the computation concludes with $\mathcal{M} = \{(1), (2), (3)\}$. Notice that although $\{(4)\} \equiv_{\Sigma \cup \mathsf{sig}(\mathcal{M})} \emptyset$, axiom (4) is neither Δ- nor $\emptyset$-local for $\Sigma \cup \mathsf{sig}(\mathcal{M})$ and so the $\top\bot^*$-module of $\mathcal{T}$ w.r.t. Σ coincides with $\mathcal{T}$ (see below and [3] for definitions).

It is often the case (e.g., for the NCI Thesaurus) that a real-world ontology satisfies all conditions for acyclic TBoxes with the exception that it contains multiple concept inclusions of the form $A \sqsubseteq C_1, \ldots, A \sqsubseteq C_n$. We call such TBoxes *acyclic with repeated*

concept inclusions. Clearly, one can convert such a TBox into an equivalent acyclic TBox by replacing all repeated concept inclusions of the form $A \sqsubseteq C_1$,..., $A \sqsubseteq C_n$ with $A \sqsubseteq C_1 \sqcap \ldots \sqcap C_n$. However, such an explicit conversion is an unattractive solution for module extraction because if such an axiom is added to a Σ-module the signature of the module now contains every symbol in the definition of every repeated name increasing the size of the resulting module considerably. The approach we take to handle acyclic TBoxes with repeated concept inclusions is to introduce fresh concept names for different repeated occurrences of a concept name in the left-hand side of concept inclusions, extract modules from the resulting acyclic TBox and then substitute away the added names as follows.

Theorem 2. *Let $\mathcal{T}$ be an acyclic TBox with repeated concept inclusions and Σ a signature. Let $\mathcal{T}'$ consist of all $A \bowtie C \in \mathcal{T}$ which are not repeated in $\mathcal{T}$ and all $A'_1 \sqsubseteq C_1, \ldots, A'_n \sqsubseteq C_n, A \sqsubseteq A'_1 \sqcap \ldots \sqcap A'_n$, where $A \sqsubseteq C_1$,..., $A \sqsubseteq C_n$ are all concept inclusions in $\mathcal{T}$ with A on the left hand side, $n > 1$, and $A'_1, \ldots, A'_n$ are fresh concept names.*

Let $\mathcal{M}'$ be a depleting Σ-module of $\mathcal{T}'$ and let $\mathcal{M}$ be obtained from $\mathcal{M}'$ by dropping the added axioms of the form $A \sqsubseteq A'_1 \sqcap \ldots \sqcap A'_n$ and by replacing every occurrence of the introduced symbols $A'_1, \ldots, A'_n$ with A. Then $\mathcal{M}$ is a depleting Σ-module of $\mathcal{T}$.

4 Experiments and Evaluation

We implemented the algorithm presented in Figure 1 and the refinement for acyclic TBoxes with repeated concept inclusions in the AMEX system which is written in Java aided by the OWL-API library [5] for ontology manipulation. The inseparability check was implemented using the reduction to the validity of Quantified Boolean Formulae (QBF) and uses the QBF solver sKizzo [2].

To evaluate the efficiency of AMEX and the size of the modules computed by AMEX we compare it to $\top\bot^*$ locality-based module extraction [3, 13] as implemented in the OWL-API library version 3.2.4.1806 (called STAR-modules for ease of pronunciation).

To evaluate the performance of both approaches we consider random and axiom signatures. To generate a random signature size n given a TBox $\mathcal{T}$ we take the set of all concepts in $\mathcal{T}$, i.e. $\mathsf{sig}(\mathcal{T}) \cap \mathsf{N_C}$ and select at random n symbols from this set. For each concept signature size we also include a percentage of role names randomly selected from $\mathsf{sig}(\mathcal{T})$, varying between 0% which equates to just using a concept signature to 100% which would be equal to $\Sigma \cup (\mathsf{sig}(\mathcal{T}) \cap \mathsf{N_R})$. For experiments on axiom signatures, for a given number m, we select at random m axioms from $\mathcal{T}$ and then extract a module for each of the signatures of selected axioms.

In our experiments we used the NCI Thesaurus version 08.09d taken from the Bioportal [15] repository. This version of NCI contains 116 515 logical axioms among which 87 934 are concept inclusions of the form $A \sqsubseteq C$ and 10 366 are concept equations of the form $A \equiv C$. In what follows, $\mathsf{NCI}^\star(\sqsubseteq)$ denotes the TBox consisting of all such inclusions, $\mathsf{NCI}^\star(\equiv)$ denotes the TBox consisting of all such equations, and $\mathsf{NCI}^\star$ denotes the union of both. All three TBoxes are acyclic (with repeated concept inclusions), so AMEX can be applied to them. $\mathsf{NCI}^\star$ together with the rest of the ontology

Role%	0%			25%			50%			75%			100%		
$\|\Sigma\|$	Star	AMEX	% Diff	Star	AMEX	% Diff	Star	AMEX	% Diff	Star	AMEX	% Diff	Star	AMEX	% Diff
NCI*															
100	3835.7	676.6	467%	3848.6	943.7	308%	3891.7	984.0	295%	3929.4	1014.7	287%	3929.8	1016.5	287%
250	5310.2	1725.9	208%	5365.6	1795.2	199%	5463.1	1871.5	192%	5506.3	1919.3	187%	5505.4	1918.0	187%
500	6985.9	2735.9	155%	7109.6	2844.9	150%	7165.5	2930.3	145%	7252.8	3002.1	142%	7245.9	2990.1	142%
750	8223.3	3572.7	130%	8355.2	3698.8	126%	8464.4	3806.1	122%	8538.5	3878.7	120%	8526.1	3872.0	120%
1000	9276.7	4333.6	114%	9397.2	4458.4	111%	9492.8	4573.9	108%	9564.9	4627.1	107%	9565.3	4642.7	106%
NCI* ($\sqsubseteq$)															
100	55.47	65.04	-15%	232.76	281.90	-17%	286.13	318.81	-10%	312.59	333.65	-6%	339.83	351.70	-3%
250	328.28	390.81	-16%	559.56	657.37	-15%	651.05	718.62	-9%	712.87	759.06	-6%	765.30	796.47	-4%
500	852.89	1007.34	-15%	1046.44	1190.43	-12%	1193.75	1301.77	-8%	1278.48	1355.05	-6%	1378.30	1435.99	-4%
750	1325.96	1541.33	-14%	1517.68	1692.20	-10%	1675.37	1808.43	-7%	1802.61	1905.29	-5%	1921.13	1993.60	-4%
1000	1786.342	2039.67	-12%	1973.82	2174.04	-9%	2157.33	2314.21	-7%	2299.00	2416.32	-5%	2440.76	2527.34	-3%
NCI* ($\equiv$)															
100	2784.33	316.31	780%	2792.99	319.73	774%	2785.51	318.49	775%	2770.32	318.61	770%	2779.03	318.43	773%
250	3982.18	622.74	539%	3988.51	626.22	537%	3984.76	624.03	539%	3989.88	624.62	539%	3982.62	625.91	536%
500	4975.97	1001.20	397%	4988.08	1003.83	397%	4984.67	1002.06	397%	4983.22	1004.13	396%	4988.71	1004.00	397%
750	5529.94	1309.98	322%	5540.59	1315.34	321%	5533.68	1309.22	323%	5532.04	1310.77	322%	5531.00	1311.72	322%
1000	5899.871	1577.42	274%	5897.36	1576.94	274%	5891.82	1576.71	274%	5894.13	1574.57	274%	5900.37	1578.06	274%

Fig. 3. Random signature comparison

(18 215 axioms) is called NCI and contains, in addition, role inclusions, domain and range restrictions, disjointness axioms, data properties, and 17 763 ABox assertions.

The majority of NCI* (all but 4 588 axioms) are $\mathcal{EL}$-inclusions. The non-$\mathcal{EL}$ inclusions contain 7 806 occurrences of value restrictions. The signature of NCI* contains 68 862 concept and 88 role names.

Experiments with NCI* and its Fragments The results given in Table 5 show the average sizes (over 1 000 random signatures for each signature size and role percentage combination) of the modules computed by the two approaches for random signatures. It can be seen that

- in NCI$^\star(\equiv)$, AMEX-modules are significantly smaller than STAR-modules (between 270% and 780%);
- in NCI$^\star(\sqsubseteq)$, STAR-modules are, on average, slightly smaller than AMEX modules;
- in NCI*, AMEX-modules are still significantly smaller than STAR-modules, but less so than in NCI$^\star(\equiv)$.

The huge difference between modules in NCI$^\star(\sqsubseteq)$ and NCI$^\star(\equiv)$ can be explained as follows: it is shown in [8] that for acyclic $\mathcal{EL}$-TBoxes without concept equations, AMEX-modules and STAR-modules coincide. This is not the case for acyclic $\mathcal{ALCQI}$-TBoxes (there can be axioms in STAR-modules that are not AMEX-modules and vice versa), but since the vast majority of axioms in NCI$^\star(\sqsubseteq)$ are $\mathcal{EL}$-inclusions one should not expect any significant difference between the two types of modules. Thus, it is exactly those acyclic TBoxes that contain many concept equations for which AMEX-modules are significantly smaller than STAR-modules (see Example 1 for an illustration).

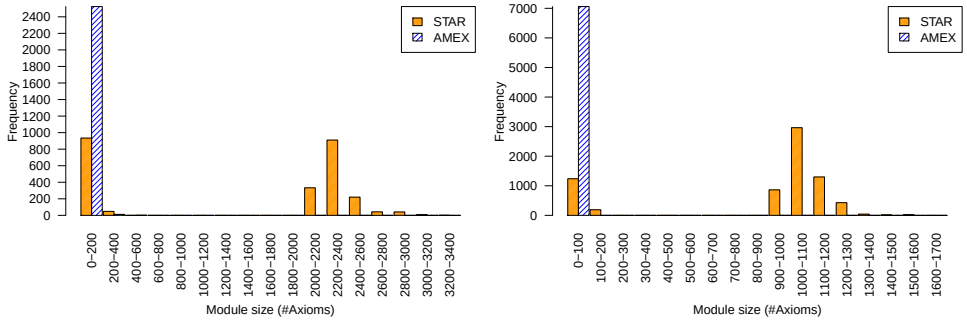

Fig. 4. Frequency of module sizes for NCI* (left) and NCI$^\star(\equiv)$ (right).

Figure 4 summarises our experimental results for modules extracted for axiom signatures. The figure shows the frequency of AMEX and STAR-modules of a given size within NCI* and NCI$^\star(\equiv)$ for the cases when the modules differ – which is in 13% and 68% of cases, respectively. For NCI$^\star(\equiv)$ in the cases in which we find a difference the STAR module is always larger than the corresponding AMEX module with an average difference of 865.6 axioms. For NCI* in a few (87 cases) the STAR modules are smaller than the corresponding AMEX ones by an average difference of 6.9 axioms whereas in the rest of the cases the STAR modules are much larger with an average difference of 1427 axioms. We do not show the results for NCI$^\star(\sqsubseteq)$ since, as explained above for the experiments with random signatures, there is essentially no difference between AMEX and STAR-modules.

These experiments were carried out on a PC with an Intel i5 CPU @ 3.30GHz with 2GB of Java heap space available to the program. For NCI* the average time taken per extraction was just under 3s and the maximum time taken was 15s. Interestingly, in almost all experiments the QBF solver was called just once. Thus, in most cases the modules were computed purely syntactically and the QBF solver simply provided an assurance that the extracted axioms indeed constituted a depleting module. Only in 3% of all experiments the QBF solver identified separability causing axioms. The maximal number of separability causing axioms recorded in any single extraction was 4 and the maximal number of QBF solver calls themselves was 73.

Experiments with full NCI Although AMEX-modules are significantly smaller than STAR-modules for acyclic TBoxes containing many concept equations, the applications of AMEX alone are very limited since most ontologies contain additional axioms such as disjointness axioms, role inclusions, and domain and range restrictions. To tackle this problem we first observe that, in principle, AMEX can be applied to any general TBoxes: given such a TBox $\mathcal{T}$, one can split $\mathcal{T}$ into two parts $\mathcal{T}_1$ and $\mathcal{T}_2$, where $\mathcal{T}_1$ is an acyclic $\mathcal{ALCQI}$-TBox (and as large as possible) and $\mathcal{T}_2 := \mathcal{T} \setminus \mathcal{T}_1$. Then for any signature Σ it follows from the robustness properties [7] of the inseparability relation $\equiv_\Sigma$ that if $\mathcal{M}$ is a depleting $\Sigma \cup \mathsf{sig}(\mathcal{T}_2)$-module of $\mathcal{T}_1$ (note that $\mathcal{M}$ can be

computed by AMEX), then $\mathcal{M} \cup \mathcal{T}_2$ is a depleting Σ-module of $\mathcal{T}$ as well. Such a direct application of AMEX to general TBoxes is unlikely to compute small modules when $\mathcal{T}_2$ is large. However, our first experimental results suggest that this approach is beneficial when iterated with STAR-module extraction. The following result provides the theoretical underpinning for our experiments.

Theorem 3. *Let $\mathcal{M} \subseteq \mathcal{M}' \subseteq \mathcal{T}$ be TBoxes and Σ a signature such that $\mathcal{M}'$ is a depleting Σ-module of $\mathcal{T}$ and $\mathcal{M}$ is a depleting Σ-module of $\mathcal{M}'$. Then $\mathcal{M}$ is a depleting Σ-module of $\mathcal{T}$.*

Since both AMEX and STAR compute depleting Σ-modules, given a signature Σ and ontology $\mathcal{T}$ one can extract an AMEX module from the STAR module (and vice versa) and have the guarantee the resulting module is still a depleting Σ-module of $\mathcal{T}$. In this way, one can repeatedly extract from the output of one extraction approach again a module using the other approach until the sequence of modules becomes stable.

The following experiments are based on a naïve implementation of this hybrid approach and extract modules from the full version of NCI. Again we consider random concept signatures with varying amount of role names. The experiments shown in Figure 5 are based on 200 signatures for each concept signature size/role percentage combination and compare the average size of modules extracted using the hybrid approach and using STAR extraction only.

Role%	0%			25%			50%			75%			100%		
$\lvert\Sigma\rvert$	Star	Iterated	% Diff	Star	Iterated	% Diff	Star	Iterated	% Diff	Star	Iterated	% Diff	Star	Iterated	% Diff
100	5385.7	1949.5	176%	9569.8	6177.7	55%	13733.8	10339.0	33%	19486.4	16089.1	21%	23196.6	19810.2	17%
250	7298.6	3268.7	123%	11959.8	7963.9	50%	16072.1	12069.6	33%	20974.9	16978.8	24%	25141.0	21134.7	19%
500	9445.0	4827.6	96%	13165.1	8533.4	54%	16406.7	11767.0	39%	23046.8	18418.3	25%	27331.2	22691.9	20%
750	11070.2	6058.6	74%	15268.3	10235.9	49%	19696.2	14683.3	34%	23705.7	18689.6	27%	28917.3	23903.4	21%
1000	12370.7	7108.5	74%	16434.7	11174.0	47%	21978.6	16737.6	31%	25529.0	20286.5	26%	30218.5	24965.4	21%

Fig. 5. Iterative module extraction from NCI

For all signatures we found a reduction in the size of the module when iterated with the STAR module on its own being between 17% and 176% larger than the hybrid module.

In Figure 6, we show the results of our experiments for axioms signatures. They are based on 20 000 randomly selected axioms from the full NCI Thesaurus. 13% of such signatures showed a difference from the STAR module. The frequency of module sizes for the cases when the modules differ is given in Figure 6. The average difference in size, for the cases when there is a difference, is 295.2 axioms.

All individual extractions using the hybrid approach saw exactly 2 alternations of the STAR module extraction whereas the AMEX extraction varied between 1 and 2 times. The cases in which the AMEX extraction alternated just once happened much more often as the signature sizes grew and the difference between the respective module sizes became smaller.

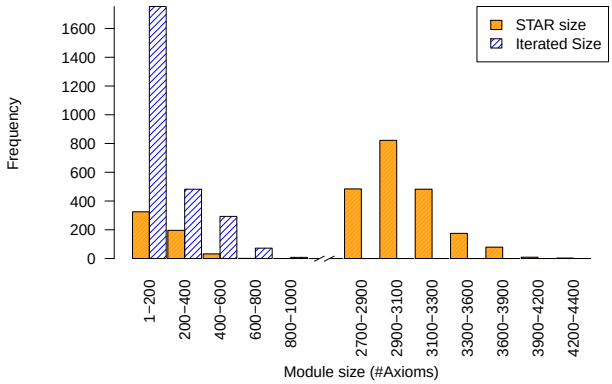

Fig. 6. Frequency of module sizes

Following both AMEX and STAR module extraction works very efficiently, especially for small input ontologies, the additional time taken to extract the hybrid module compared to the STAR extraction alone was at most only 2.2 seconds.

5 Conclusion

We have presented a new system, AMEX, for depleting module extraction from acyclic $\mathcal{ALCQI}$-TBoxes. Using the NCI Thesaurus, we have compared the size of AMEX-modules with the size of $\top\bot^*$-modules computed by the OWL-API library implementation (referred as STAR-modules) and we have presented a hybrid approach in which STAR and AMEX-module extraction are used iteratively. The results show that for TBoxes with many axioms of the form $A \equiv C$, AMEX-modules can be significantly smaller than STAR-modules and that an iterative approach can lead to significantly smaller modules than 'pure' STAR-modules. In contrast to [4], where a large number of ontologies are used to compare STAR-modules and MEX-modules we consider NCI only. The reason is that the majority of ontologies considered in [4] contain no (or only a very small set) of axioms of the form $A \equiv C$ that form an acyclic subset of the ontology. For such ontologies it follows both from theoretical results in [8] and experimental results in [4] that there is no significant difference between AMEX and STAR-modules. Instead, we focus on a high quality ontology with a reasonable number of concept equations and where theory predicts that minimal depleting modules can be much smaller than STAR-modules. Many research questions remain to be explored. In particular, to apply AMEX to a larger class of ontologies in an iterative approach, one has to generalise the notion of acyclic TBoxes in such a way that the underpinning methodology of AMEX can still be generalised.

References

1. F. Baader, D. Calvanes, D. McGuiness, D. Nardi, and P. Patel-Schneider. *The Description Logic Handbook: Theory, implementation and applications*. Cambridge University Press, Cambridge, UK, 2003.
2. M. Benedetti. sKizzo: a QBF decision procedure based on propositional skolemization and symbolic reasoning. Technical Report 04-11-03, ITC-irst, 2004.
3. B. Cuenca Grau, I. Horrocks, Y. Kazakov, and U. Sattler. Modular reuse of ontologies: theory and practice. *Journal of Artificial Intelligence Research (JAIR)*, 31:273–318, 2008.
4. C. Del Vescovo, P. Klinov, B. Parsia, U. Sattler, T. Schneider, and D. Tsarkov. Empirical study of logic-based modules: Cheap is cheerful. Technical report, University of Manchester, 2013.
5. M. Horridge and S. Bechhofer. The OWL API: A Java API for OWL ontologies. *Semantic Web*, 2(1):11–21, 2011.
6. B. Konev, R. Kontchakov, M. Ludwig, T. Schneider, F. Wolter, and M. Zakharyaschev. Conjunctive query inseparability of OWL 2 QL TBoxes. In *Proceedings of the 25th Conference on Artificial Intelligence, AAAI 2011*, pages 221–226, Menlo Park, CA, USA, 2011. AAAI Press.
7. B. Konev, C. Lutz, D. Walther, and F. Wolter. Formal properties of modularisation. In *Modular Ontologies: Concepts, Theories and Techniques for Knowledge Modularization*, volume 5445 of *Lecture Notes in Computer Science*, pages 25–66. Springer, Berlin, Heidelberg, 2009.
8. B. Konev, C. Lutz, D. Walther, and F. Wolter. Model-theoretic inseparability and modularity of description logic ontologies. *Artificial Intelligence*, 203:66–103, 2013.
9. R. Kontchakov, L. Pulina, U. Sattler, T. Schneider, P. Selmer, F. Wolter, and M. Zakharyaschev. Minimal module extraction from DL-Lite ontologies using QBF solvers. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence, IJCAI 2009*, pages 836–841, Menlo Park, CA, USA, 2009. AAAI Press.
10. R. Kontchakov, F. Wolter, and M. Zakharyaschev. Logic-based ontology comparison and module extraction, with an application to DL-Lite. *Artificial Intelligence*, 174(15):1093–1141, 2010.
11. C. Lutz and F. Wolter. Deciding inseparability and conservative extensions in the description logic $\mathcal{EL}$. *Journal of Symbolic Computing*, 45(2):194–228, 2010.
12. R. Nortjé, K. Britz, and T. Meyer. Module-theoretic properties of reachability modules for sriq. In *Proceedings of the 26th international workshop on description logic, DL 2013*, CEUR Workshop Proceedings. CEUR-WS.org, 2013.
13. U. Sattler, T. Schneider, and M. Zakharyaschev. Which kind of module should I extract? In *Proceedings of the 22nd International Workshop on Description Logics, DL 2009*, volume 477 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.
14. H. Stuckenschmidt, C. Parent, and S. Spaccapietra, editors. *Modular Ontologies: Concepts, Theories and Techniques for Knowledge Modularization*, volume 5445 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, 2009.
15. P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39(Web-Server-Issue):541–545, 2011.