Artificial Intelligence 203 (2013) 66–103

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

# Model-theoretic inseparability and modularity of description logic ontologies

Boris Konev [a,*], Carsten Lutz [b], Dirk Walther [c], Frank Wolter [a]

[a] *University of Liverpool, Liverpool, UK*
[b] *Universität Bremen, Bremen, Germany*
[c] *Universidad Politécnica de Madrid, Madrid, Spain*

A R T I C L E   I N F O

*Article history:*
Received 16 October 2012
Received in revised form 25 July 2013
Accepted 26 July 2013
Available online 6 August 2013

*Keywords:*
Description logic
Modularisation
Computational complexity
Reasoning

A B S T R A C T

The aim of this paper is to introduce and study model-theoretic notions of modularity in description logic and related reasoning problems. Our approach is based on a generalisation of logical equivalence that is called model-theoretic inseparability. Two TBoxes are inseparable w.r.t. a vocabulary $\Sigma$ if they cannot be distinguished by the $\Sigma$-reducts of their models and thus can equivalently be replaced by one another in any application where only vocabulary items from $\Sigma$ are relevant. We study in-depth the complexity of deciding inseparability for the description logics $\mathcal{EL}$ and $\mathcal{ALC}$ and their extensions with inverse roles. We then discuss notions of modules of a TBox based on model-theoretic inseparability and develop algorithms for extracting minimal modules from acyclic TBoxes. Finally, we provide an experimental evaluation of our module extraction algorithm based on the large-scale medical TBox Snomed ct.



## 1. Introduction

The main use of ontologies in computer and information science is to formalise the vocabulary of an application domain; i.e., to fix the vocabulary as a logical signature and to provide a logical theory that defines the semantics of the terms and relations in the vocabulary. The wide adoption of the W3C-endorsed ontology language OWL and its profiles [1,2], the success of logic-based reasoning support for concept classification and instance retrieval [3–6], and the availability of ontology editors and management systems such as Protégé and SWOOP [7,8] has led to the development and use of large-scale and complex ontologies that capture the vocabulary and knowledge of a wide diversity of domains. Especially in the Life Sciences and other knowledge intensive domains, many such ontologies have been created. Important examples are the National Cancer Institute's (NCI) thesaurus and ontology, the gene ontology (GO), and Snomed ct, the Systematized Nomenclature of Medicine, Clinical Terms, which comprises about three hundred thousand vocabulary items and is used in the healthcare systems of more than twenty countries [9–11].

Engineering and maintaining professional ontologies such as the ones mentioned above is a complex and challenging task, and it has to be carried out with great care for the resulting ontology to be of high quality. Ontology design may involve a group of ontology engineers and domain experts that co-operate in order to design the ontology, update it to reflect changes/developments in the domain, and integrate it with other ontologies so as to cover larger domains. In such an environment, automated tool support for comparing, merging, updating, modularising, and reusing ontologies is of critical importance.

* Corresponding author.
*E-mail addresses:* konev@liverpool.ac.uk (B. Konev), clu@uni-bremen.de (C. Lutz), dirk.walther@upm.es (D. Walther), wolter@liverpool.ac.uk (F. Wolter).

0004-3702/$ – see front matter 
http://dx.doi.org/10.1016/j.artint.2013.07.004

The aim of this paper is to propose and investigate a model-theoretic notion of inseparability between ontologies that can serve as a logical underpinning for many of these operations. From an application perspective, two ontologies are inseparable if they can be equivalently replaced by one another in any context. In logic, this can be formalised by the notion of logical equivalence, according to which two ontologies are inseparable if they have the same models. However, logical equivalence as such is clearly not sufficiently flexible to serve as a logical underpinning of modularity. For example, a module of an ontology is typically not logically equivalent to the ontology itself, and an updated ontology is typically not logically equivalent to the original ontology. By parameterising logical equivalence with a signature $\Sigma$ of vocabulary items of interest, we obtain a notion of inseparability that has exactly the flexibility and properties required. We say that two ontologies are $\Sigma$*-inseparable* if the $\Sigma$-reducts of their models coincide, where the $\Sigma$-reduct of a structure is simply the restriction of that structure to the symbols in $\Sigma$. Then, a module can be defined as a subset of an ontology that is self-contained in the sense that it is $\Sigma$-inseparable from the ontology, where $\Sigma$ is the set of vocabulary items that occur in the module. Similarly, an update can be categorised as not harmful for a signature $\Sigma$ if the updated ontology is $\Sigma$-inseparable from the original ontology. Many other relevant notions such as controlled import of ontologies and safety of an ontology for a signature as studied in [12] can be formalised as well using $\Sigma$-inseparability, see [12,13].

An important feature of model-theoretic inseparability is its language independence. Alternative notions of inseparability are based on deduction and use logical consequence within a given language to define inseparability of TBoxes. For example, two $\mathcal{ALC}$-TBoxes are called deductively $\Sigma$-inseparable in $\mathcal{ALC}$ iff they entail the same $\mathcal{ALC}$-concept inclusions in $\Sigma$. Distinct description logics typically define distinct versions of deductive inseparability [14] which is clearly undesirable if one moves from one language to another. In contrast, model-theoretic inseparability implies inseparability w.r.t. any standard description logic and even w.r.t. second-order logic.

The first main contribution of this paper is a systematic analysis of the complexity of deciding whether two ontologies are $\Sigma$-inseparable. We focus on ontologies that are formulated as a general or acyclic TBox in the description logic $\mathcal{EL}$ that underpins the OWL2 EL profile, the paradigmatic expressive description logic $\mathcal{ALC}$, and their extensions with inverse roles $\mathcal{ELI}$ and $\mathcal{ALCI}$. Our analysis starts with two fundamental undecidability results: $\Sigma$-inseparability is undecidable for $\mathcal{ALC}$-TBoxes even if one TBox is acyclic and the other is empty; for $\mathcal{EL}$, the undecidability result is slightly weaker, applying to the case where one TBox is empty and the other is a general TBox. It is due to such undecidability results that automated tool support for modular ontology design and maintenance is currently not based on model-theoretic inseparability, but either on deductive versions of inseparability (see, for example, [14]) or stronger inseparability notions based on locality [12]. However, the second part of our complexity analysis reveals that there are natural conditions on the TBox and/or signature that lead to a dramatic drop in complexity.

The first such condition is the restriction of the signature to a concept signature, i.e., a signature that comprises only concept names, but no role names. In this case, deciding $\Sigma$-inseparability becomes coNExp$^{\text{NP}}$-complete for general $\mathcal{EL}$- and $\mathcal{ALC}$-TBoxes. The proof of this result is of particular interest since it reveals a close connection between $\Sigma$-inseparability on the one hand and satisfiability in non-monotonic description logics based on circumscription on the other hand. By combining concept signatures with the additional condition that one TBox is empty, the complexity goes down to $\Pi_2^p$ for $\mathcal{ALC}$ and PTime for $\mathcal{EL}$. Finally, $\Sigma$-inseparability of acyclic $\mathcal{EL}$-TBoxes from the empty TBox also turns out to be in PTime, even for signatures with role names. While these cases appear to be rather restricted at first glance, they actually play a central role in our algorithms for module checking and module extraction, discussed below. All mentioned results also hold for the extensions of $\mathcal{EL}$ and $\mathcal{ALC}$ with inverse roles. Note that, in various relevant cases, model-theoretic inseparability thus turns out to be strictly less complex than standard reasoning tasks such as subsumption, which is PSpace-complete for acyclic TBoxes formulated in $\mathcal{ALC}$ and in $\mathcal{ELI}$ [15]. It is also interesting that the difference in complexity of subsumption between $\mathcal{EL}$ and $\mathcal{ELI}$ (PTime vs. ExpTime [15]) is not reflected in the complexity of inseparability.

The PTime and $\Pi_2^p$-complexity results indicate that model-theoretic $\Sigma$-inseparability can be useful not only as a theoretical tool to define an "ideal" form of modularity, but also for practical purposes. The second aim of this paper is to apply $\Sigma$-inseparability to define several notions of a module in a TBox, and to develop algorithms for module checking and for extracting minimal modules. The former task is, given a TBox and a subset of the TBox, to decide whether the subset is a module; the latter task is, given a TBox and a signature $\Sigma$ of interest, to extract an as small as possible module for the signature $\Sigma$. We give a polynomial time algorithm for module checking in acyclic $\mathcal{ELI}$-TBoxes and a $\Pi_2^p$-module checking algorithm for acyclic $\mathcal{ALCI}$-TBoxes under a natural additional syntactic condition. Note that acyclic TBoxes are used in relevant practical applications, including Snomed ct, several versions of NCI, and GO. For module extraction, we consider two approaches. First, we show that a generic module extraction algorithm can be applied to acyclic $\mathcal{ELI}$- and $\mathcal{ALCI}$-TBoxes in a black box manner by using our module checking algorithm as an oracle. Second, we pursue a white box approach in which we directly use the module checking algorithm for acyclic $\mathcal{ELI}$-TBoxes to obtain a more direct module extraction algorithm. Finally, we introduce the module extraction software MEX that implements the white box approach and carry out a case study by extracting minimal modules from the Snomed ct ontology. The study shows that our algorithms scale effortlessly to ontologies of very large size and very often extract modules that are significantly smaller than those produced by the standard $\top\bot^*$-module extraction algorithm [16,17] and all other existing approaches.

The paper is organised as follows. After a section introducing basic definitions and terminology we define model-theoretic inseparability in Section 3. In this section, we also introduce and investigate basic properties and applications of model-theoretic inseparability. In Sections 4 and 5, we investigate the computational complexity of deciding $\Sigma$-inseparability and then, in Sections 6 and 7 we introduce and investigate module checking and module extraction based on $\Sigma$-inseparability.