Proceedings of the Twelfth International Conference on the Principles of Knowledge Representation and Reasoning (KR 2010)

The Modular Structure of an Ontology: An Empirical Study*

Bijan Parsia and **Thomas Schneider**
School of Computer Science, University of Manchester, UK
{bparsia,schneider}@cs.man.ac.uk

1. Introduction

Why modularize an ontology? In software engineering, modularly structured systems are desirable. Given a well-designed modular program, it is generally easier to process, modify, and analyze it and to reuse parts by exploiting the modular structure. As a result, support for modules (or components, classes, objects, packages, aspects) is a commonplace feature in programming languages.

Ontologies are computational artefacts and, like programs, have to be designed, modified etc. and can get large and complex. Therefore, research into modularity for ontologies has been an active area for ontology engineering. Recently, much effort has gone into developing *logically sensible* modules: modules offering strong logical guarantees for intuitive modular properties. One such guarantee is *coverage*. It means that the module captures all the ontology's knowledge about a given set of terms (signature). It is provided by modules based on conservative extensions and by efficient approximations, e.g., locality-based modules.

The task of extracting one module given a signature, GetOne, is well understood and starting to be deployed in standard ontology development environments, such as Protégé 4. The extraction of locality-based modules has been effectively used in the field for ontology reuse (Jimeno et al. 2008) and a subservice for incremental reasoning (Cuenca Grau, Halaschek-Wiener, and Kazakov 2007).

Here, we are interested in the modular structure of the ontology as a whole, determined by the set of *all* modules, or at least a subset. We call the task of a-posteriori determining the modular structure GetAll. While GetOne is well-understood and often computationally cheap, GetAll has hardly been examined for module notions with logical guarantees, the work described in (Cuenca Grau et al. 2006) being a promising exception. GetOne also requires the user to know in advance the set of terms to input to the extractor: we call this a *seed* signature for the module. One module can have several seed signatures. Since there are non-obvious relations between the final signature of a module and its seed signature, users are often unsure how to generate a request and confused by the results. The modular structure of the ontology determined by GetAll could guide their extraction choices. Supported by the experience described in (Cuenca Grau et al. 2006), we believe that, by revealing the modular structure of an ontology, we can obtain information about its topicality, connectedness, structure, superfluous parts, or agreement between actual and intended modeling.

In the worst case, the number of *all* modules of an ontology is exponential in the minimum of the number of terms and the number of axioms in the ontology. Thus, it is possible that all real ontologies have too many modules to extract all of them, even if an optimized extraction methodology were at hand. Even with only polynomially many modules, there may be too many for direct user inspection. Then, some other form of analysis would have to be designed.

In this paper, we report on experiments to obtain or estimate this number and to evaluate the modular structure of an ontology where we succeeded to compute it.

Related work. One solution to GetAll are partitions related to $\mathcal{E}$-connections (Cuenca Grau et al. 2006; Cuenca Grau, Parsia, and Sirin 2006). The resulting modules are disjoint, and the technique, when it suceeds, divides an ontology into three kinds of modules: (A) those which import vocabulary from others, (B) those whose vocabulary is imported, and (C) isolated parts. In experiments and user experience, the extracted parts were often few and corresponded usefully to user understanding. For instance, the tutorial ontology Koala, consisting of 42 logical axioms, is partitioned into one A-module about animals and three B-modules about genders, degrees and habitats. It was also shown in (Cuenca Grau et al. 2006) that certain combinations of these parts provide coverage. For Koala and other, well structured ontologies, such a combination would still be the whole ontology. Furthermore, robustness properties (e.g., under vocabulary extension) of these parts are not as well-understood as for locality-based modules.

Other approaches to GetAll, and most approaches to GetOne, either do not provide coverage or are restricted to fragments of OWL 2. See also our more detailed discussion of related work in (Del Vescovo et al. 2010b).

A-priori approaches require the ontology developer to specify modules syntactically in advance. Here we may still want to understand the modular structure of these parts. Furthermore, it is not always clear whether the imposed structure is correct: decisions about modular structure have to be

*Supported by the UK EPSRC grant EP/E065155/1.

taken early in the modeling which may enshrine misunderstandings. Examples were reported in (Cuenca Grau et al. 2006), where user attempts to capture the modular structure of their ontology by separating the axioms into separate files were totally at odds with the analyzed structure.

Overview. We report on experiments where we extracted *all* modules from real ontologies as a first solution candidate for GetAll. We use modules based on syntactic locality (Cuenca Grau et al. 2008), which provide coverage and other useful properties of modules (Sattler, Schneider, and Zakharyaschev 2009). At this stage, we are mainly interested in module *numbers*, to find out whether the suspected combinatorial explosion occurs. We also sampled subsets of the ontologies and fully modularized them, measuring the relation between module number and subontology size for each ontology. We have also tried filtering modules in different ways.

An extended version of this paper and additional material about the experiments, such as spreadsheets and charts, are available online (Del Vescovo et al. 2010b; 2010a).

2. Preliminaries

We are assuming that the reader is familiar with OWL and the underlying description logics (DLs) (Horrocks, Patel-Schneider, and van Harmelen 2003; Horrocks, Kutz, and Sattler 2006). We consider an ontology to be a finite set of concept or role inclusion axioms, disregarding non-logical axioms, which can easily be added to the extracted logical module. A *signature* is a set of concept and role names. We can think of it as specifying a topic of interest. Given a concept or role name, axiom, or ontology X, we call the set of terms in X the *signature of* X, denoted by $\widetilde{X}$.

Conservative extensions and locality. Conservative extensions (CEs) capture encapsulation of knowledge: a CE-based module for a signature Σ of an ontology $\mathcal{O}$ preserves all entailments of $\mathcal{O}$ that can be formulated using symbols in Σ only. For more precise definitions, see e.g., (Konev et al. 2009; Lutz, Walther, and Wolter 2007; Del Vescovo et al. 2010b).

CEs are hard/impossible to decide for many DLs (Ghilardi, Lutz, and Wolter 2006; Konev et al. 2009), but approximations have been found, such as *syntactic locality* (here for short: locality). Locality-based modules can be efficiently computed and provide coverage (Cuenca Grau et al. 2008; Jiménez-Ruiz et al. 2008). We use the notion of locality and of $\top$-, $\bot$-, $\top\bot^*$-modules from (Sattler, Schneider, and Zakharyaschev 2009, Def. 3,4). Module signature and seed signature can be orthogonal.

Genuine modules. In order to limit the overall number of modules, we introduce the notion of a *genuine module*. A given module $\mathcal{M}$ of an ontology is *fake* if it can be partitioned into a set $\{\mathcal{M}_1, \ldots, \mathcal{M}_n\}$ of smaller modules such that each entailment of $\mathcal{M}$ is an entailment of some $\mathcal{M}_i$. All other modules are called *genuine*. We give details in (Del Vescovo et al. 2010b). In particular, if the whole ontology has a *partition* into modules, then every entailment can be obtained from some of those modules. Fake modules are uninteresting: different seed signatures of the $\mathcal{M}_i$ do not interact with each other. Given that often the overall number of modules appears to grow exponentially with the size of the subontology, a natural question arising is whether only the number of *fake* modules is exponential.

3. Experiments and results

We have extracted all modules from real ontologies and their subsets. In the worst case, the module number can be exponential in the number of terms or axioms in the ontology—even for genuine modules of very simple families of ontologies, for instance $\mathcal{T}_n = \{B \sqsubseteq A\} \cup \{C_i \sqsubseteq B \mid 1 \leqslant i \leqslant n\}$, or $\mathcal{O}_n = \{B_i \sqsubseteq A,\ C_i \sqsubseteq B_i \mid 1 \leqslant i \leqslant n\} \cup \{B_i \sqsubseteq \neg B_j \mid 1 \leqslant i < j \leqslant n\}$. All other examples that we are aware of rely on the class hierarchies having unbounded width. In contrast, there are ontologies of arbitrary size with exactly one or at most quadratically many modules. Thus, real ontologies might still have a reasonable number of modules. Unfortunately, empirically, as discussed in the following, this does not seem to be the case. See (Del Vescovo et al. 2010b) for all omitted details.

Full modularization. The table below shows the full modularization of Koala and Mereology for the four module types, where $\top\bot^*_g$ denotes genuine $\top\bot^*$ modules. "Size" refers to the number of logical axioms—a syntax-dependent measure. We will look at alternatives in future work.

We observe that the number of modules increases from $\top$- via $\bot$- to $\top\bot^*$-modules as expected because $\top$-modules tend to be bigger and apparently too coarse-grained for our purposes. For a more fine-grained modularization, we pay with an increased module number and extraction time.

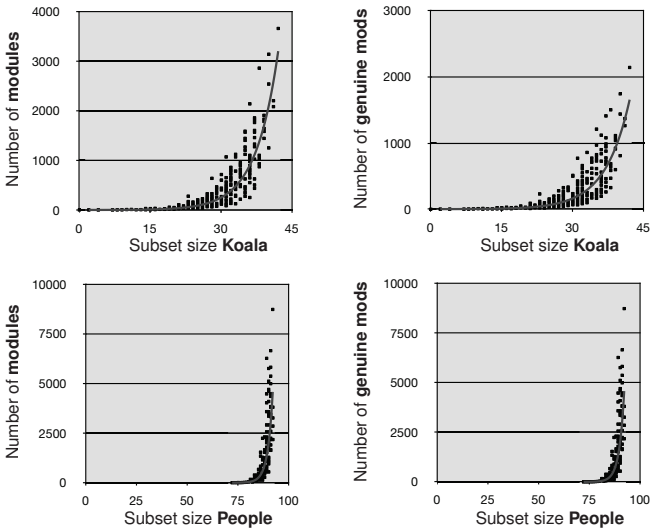
	Koala				Mereology			
	$\top$	$\bot$	$\top\bot^*$	$\top\bot^*_g$	$\top$	$\bot$	$\top\bot^*$	$\top\bot^*_g$
#Modules	**12**	**520**	**3,660**	**2,143**	**40**	**552**	**1,952**	**272**
Time [s]	*0*	*1*	*9*	*34*	*0*	*6*	*158*	*158*
Min size	29	6	0	0	18	0	0	0
Avg size	35	27	23	23	26	25	20	22
Max size	42	42	42	42	40	40	40	38
Std. dev.	4	6	6	6	6	7	8	8

A full modularization of larger ontologies did not succeed. We cancelled these computations after several hours, when thousands of modules have been extracted.

Although 3,660 and 1,952 are much smaller than the theoretical upper bound of 2^{25}, they are still too big for inspection. We therefore tried two more ways to reduce modules to fewer "interesting" ones; both showed no significant impact.

Subset sampling. In order to test whether it is plausible that other, bigger, ontologies have an exponential number of modules, we sampled subontologies, ordered them by size, and modularised them in increasing order until a single modularisation exceeded a pre-set timeout. We are convinced that most of the ontologies examined exhibit the feared exponential behavior: the figure below shows scatterplots of the number of $\top\bot^*$ modules (genuine $\top\bot^*$ modules) versus the size of the subset for People and Koala. Each chart shows an exponential trendline, the least-squares fit through

the data points by using an exponential equation. For more charts and spreadsheets, see (Del Vescovo et al. 2010a).



The equations and their determination coefficients (R^2 values) are given in the table below, which also includes the other ontologies and the estimated number of modules for the full ontology as per the trendline equation. The results for the first six ontologies strongly suggest an exponential dependence of the module number on the subset size.

	Confidence		Trendline equation		Estimate	
Ontology	R^2	R^2_g	$\top\bot^*$	$\top\bot^*_g$	$\top\bot^*$	$\top\bot^*_g$
People	.95	.95	$2 \cdot 10^{-13} e^{.41n}$		10^6	10^6
Mereology	.87	.94	$1.2e^{.16n}$	$1.1e^{.13n}$	10^3	10^2
Koala	.90	.88	$.45e^{.21n}$	$.50e^{.19n}$	10^3	10^3
Galen	.94	.86	$1.2e^{.24n}$	$1.6e^{.16n}$	$\gg 10^{99}$	$\gg 10^{99}$
University	.84	.83	$1.7e^{.19n}$	$1.6e^{.14n}$	10^4	10^3
OWL-S	.82	.84	$.0027e^{.17n}$	$.0032e^{.16n}$	10^{17}	10^{17}
Tambis	.75	.70	$1.1e^{.22n}$	$1.4e^{.13n}$	10^{58}	10^{33}
miniTambis	.47	.52	$2.6e^{.18n}$	$2.5e^{.14n}$	10^{14}	10^{10}

R^2, R^2_g Determination coefficient of trendlines ($\top\bot^*$, $\top\bot^*_g$)
Estimate Module numbers for full ontology as per trendline

4. Discussion and outlook

The fundamental conclusion is that even the number of *genuine* modules is exponential in the size of the ontology for real ontologies. Our estimates show that full modularization is practically impossible already for midsize ontologies.

Of course, there might be principled ways to reduce the target number of modules, such as a coarser approximation, though that would be hard to justify on logical grounds. Attempts to use "less minimal" modules or to heuristically merge modules turned out not to be helpful.

We believe that this conclusion is robust, even though our experiments on Tambis and miniTambis did not uncover exponential behavior. We expect that a longer timeout will finally reveal it and large number of unsatisfiable classes causes these ontologies to have relatively few modules.

Attempts at estimating the module number statistically by randomly sampling seed signature turned out unhelpful too.

While the outcome of the experiments means that we cannot use the complete modularization in order to analyze the ontology, it does suggest interesting lines of future work. First, we have seen correlations between several features of ontologies and a large/small number of modules, but cannot fully explain them yet. Thus, for example, we need to get a precise picture of the relationship between justificatory and modular structure. Second, even if we cannot compute all modules, we may be able to better estimate their number. We intend to explore sources of module number increase or reduction, such as the shape of the inferred class hierarchy and patterns of axioms, using artificial ontologies.

References

Cuenca Grau, B.; Parsia, B.; Sirin, E.; and Kalyanpur, A. 2006. Modularity and web ontologies. In *Proc. of KR-06*, 198–209.

Cuenca Grau, B.; Horrocks, I.; Kazakov, Y.; and Sattler, U. 2008. Modular reuse of ontologies: Theory and practice. *J. Artif. Intell. Res.* 31:273–318.

Cuenca Grau, B.; Halaschek-Wiener, C.; and Kazakov, Y. 2007. History matters: Incremental ontology reasoning using modules. In *Proc. of ISWC/ASWC-07*, volume 4825 of *LNCS*, 183–196.

Cuenca Grau, B.; Parsia, B.; and Sirin, E. 2006. Combining OWL ontologies using $\mathcal{E}$-connections. *JWebSem* 4(1):40–59.

Del Vescovo, C.; Parsia, B.; Sattler, U.; and Schneider, T. 2010a. Experimental evaluation. `http://owl.cs.manchester.ac.uk/modproj/meat-experiment`.

Del Vescovo, C.; Parsia, B.; Sattler, U.; and Schneider, T. 2010b. The modular structure of an ontology: an empirical study. Technical report, University of Manchester. `http://www.cs.man.ac.uk/%7Eschneidt/publ/modstrucreport.pdf`.

Ghilardi, S.; Lutz, C.; and Wolter, F. 2006. Did I damage my ontology? A case for conservative extensions in description logics. In *Proc. of KR-06*, 187–197.

Horrocks, I.; Kutz, O.; and Sattler, U. 2006. The even more irresistible $\mathcal{SROIQ}$. In *Proc. of KR-06*, 57–67.

Horrocks, I.; Patel-Schneider, P. F.; and van Harmelen, F. 2003. From $\mathcal{SHIQ}$ and RDF to OWL: The making of a web ontology language. *JWebSem* 1(1):7–26.

Jiménez-Ruiz, E.; Cuenca Grau, B.; Sattler, U.; Schneider, T.; and Berlanga Llavori, R. 2008. Safe and economic re-use of ontologies: A logic-based methodology and tool support. In *Proc. of ESWC-08*, volume 5021 of *LNCS*, 185–199.

Jimeno, A.; Jiménez-Ruiz, E.; Berlanga, R.; and Rebholz-Schuhmann, D. 2008. Use of shared lexical resources for efficient ontological engineering. In *SWAT4LS-08*, volume 435 of *CEUR*.

Konev, B.; Lutz, C.; Walther, D.; and Wolter, F. 2009. Formal properties of modularization. In Stuckenschmidt, H.; Spaccapietra, S.; and Parent, C., eds., *Ontology modularization*, volume 5445 of *LNCS*. Springer. 25–66.

Lutz, C.; Walther, D.; and Wolter, F. 2007. Conservative extensions in expressive description logics. In *Proc. of IJCAI-07*, 453–458.

Sattler, U.; Schneider, T.; and Zakharyaschev, M. 2009. Which kind of module should I extract? In *DL 2009*, volume 477 of *CEUR*.