Topicality in Logic-Based Ontologies

Chiara Del Vescovo, Bijan Parsia, and Uli Sattler

The University of Manchester, Oxford Road, Manchester, M13 9PL, UK
{delvescc,bparsia,sattler}@cs.man.ac.uk

Abstract In this paper we examine several forms of modularity in logics as a basis for various conceptions of the topical structure of an ontology. Intuitively, a topic is a coherent fragment of the subject matter of the ontology. Different topics may play different roles: e.g., the main topic (or topics), side topics, or subtopics. If, at the lowest level, the subject matter of an ontology is characterized by the set of concepts of the ontology, a topic is a "coherent" subset of those concepts. Different forms of modularity induce different, more or less cognitively helpful, notions of coherence and thus distinct topical structures.

1 Introduction

When formalising a set of concepts in some logic we encounter a variety of structural issues. For example, we look for *definitions* of concepts in terms of categorizing attributes, that is, for the "internal" structure of our concepts. We also seek to discover "external" relations between concepts, e.g., of subsumption, equivalence, or disjointness. In a logic based knowledge representation system, we can hope that by giving the former, the system can discover the latter.

Common such systems are those based on description logic *ontologies*. Ontologies are (decidable) logical theories describing a shared vocabulary about a domain in terms of a set of concepts plus the relationships between those concepts. In notable examples, as SNOMED CT[1] (Systematized Nomenclature of Medicine – Clinical Terms) that contains more than $300,000$ axioms, while the internal structure of any concept may well be intelligible, the large scale structure is not. In particular, the subsumption hierarchy, as a whole, is not particular well suited to guide the interested reader toward a grasp of "what the ontology is about." All concepts – relevant or irrelevant, well or cursorily described, central or peripheral – participate, at least vacuously, in the subsumption hierarchy. One could attempt to heuristically organise concepts in the subsumption hierarchy in some larger grain way so that more strongly related terms were clustered together (e.g., as in [15]). With such a coarse grain structure, one might hope to discern the "main topics" of the ontology, various side topics, as well as topics that are neglected in the formalisation. Unfortunately, such heuristic organisation, even if based on the prior, logic-derived structural aspects of concepts, is not, itself, derived from logical features of the representation. That is, it is not

[1] http://www.ihtsdo.org/snomed-ct/

derived from, nor is it properly sensitive to, the semantics of our representation. Ideally, there would be a notion of logical topicality which, 1) comports with our intuitions and expectations about topical organization, 2) is computationally reasonable, and 3) supports ontology engineers in practice.

A promising foundation for logical topicality is the extensive recent work on *logically sensible* modules, that is, modules which offer strong logical guarantees for intuitive properties of modules [2]. For example, one such guarantee is called *coverage* of a set of terms (*signature*), and means that the module captures all the ontology's knowledge about this set. It is easy to see that such a notion is a reasonable candidate to underly topicality. The aim of this paper is to lay foundations for a theory of logical topicality in ontologies, by considering five different forms of logical modularity and possible sorts of topicality they induce.

2 Preliminaries

In this paper, we are concerned with logic-based ontologies, i.e., finite sets of axioms, formulated in a suitable logic. Our main focus is on Description Logics (DLs) [1] – decidable fragments of first order logics closely related to modal logics that form the logical underpinning of state-of-the-art ontology languages such as OWL 2[2] [4] – but our discussion is straightforwardly applicable to other logical formalisms.

A *signature* is a set of *terms*; in DLs, these are *individual* names to denote elements of the domain (constants in FOL), *concept* names to denote monadic predicates, and *role* names to denote binary predicates. A given DL, say $\mathcal{SHIQ}$, determines the set of *axioms* we can form over terms, and any finite set of axioms is an *ontology*. For $\mathcal{O}$ an ontology, we will use $\widetilde{\mathcal{O}}$ for the set of terms occurring in $\mathcal{O}$, i.e., for its signature. For example, the following axioms characterize animals, define parents, and assert facts about two individuals:

$$\begin{aligned}
&\mathsf{Animal} \sqsubseteq \mathsf{Organism} \sqcap \neg\mathsf{Plant}\\
&\mathsf{Parent} \equiv \mathsf{Animal} \sqcap \exists\mathsf{hasChild}.\mathsf{Animal}\\
&\mathsf{Animal}(\mathsf{peter}), \quad \mathsf{Parent}(\mathsf{mary}), \quad \mathsf{hasChild}(\mathsf{peter}, \mathsf{mary})
\end{aligned}$$

Traditionally, in DLs, in ontology consists of two parts, namely a *TBox* – those axioms that involve only concepts and roles, like the first two axioms above – and an *ABox* – those axioms that involve individuals, like the last three axioms above. For our considerations, however, this distinction can be mostly neglected.

We assume that our logic comes with an entailment relation $\models$. For DLs, this relation is the standard, first order entailment, which (standardly for these logics) is not only decidable but for which various decision procedures have been implemented.

Finally, we make use of the standard notion of *conservative extension*: roughly speaking, for Σ a signature and ontologies $\mathcal{O}_1 \subseteq \mathcal{O}_2$, we say that $\mathcal{O}_2$ is a (deductive) Σ-conservative extension of $\mathcal{O}_1$ if $\mathcal{O}_1 \models \alpha$ implies $\mathcal{O}_2 \models \alpha$, for every

[2] `http://www.w3.org/TR/owl2-overview/`

axiom that we can build over Σ. That is, if $\mathcal{O}_2$ is a (deductive) Σ-conservative extension of $\mathcal{O}_1$, then $\mathcal{O}_2$ says as much or as little about terms in Σ as $\mathcal{O}_1$. Conservative extensions and practical approximations have been recently used in DLs to define notions of modules that are now implemented and used in various ontology engineering tasks [10,3].

3 Topicality

Similar to, say, a book, an ontology can be about one or more things, from one or more angles, and it can be about various more or less independent topics. For example, "Zen and the Art of Motorcycle Maintenance" (by Robert M. Pirsig) can be said to be about two topics, philosophy and motorcycles, but it ties them rather closely together so that they may be viewed as dependent. In order to determine the topic of a book – or an ontology – we thus need to first agree on a suitable notion of *coherence* and its dual, *independence*.

In this paper, we are concerned with the following concepts:

- a *description*, a syntactical object, e.g., a text, a conversation, a thought, or an ontology. We assume that this object can be broken down into smaller pieces such as chapters, sentences, axioms, etc.
- a *vocabulary* and a *grammar* used to write well-formed sentences. In logics, the former is called signature and the latter determined by the syntax of the logic. We assume that the syntactical object conforms to our grammar and sticks to the signature.
- a notion of cognitive or logical *coherence*, which determines whether we take a given description as a coherent whole, i.e., one where the terms involved in it depend on each other, or whether it disaggregates into various pieces that can be said to be *independent* of each other. Clearly, coherence and independence are dual to each other, and there can be different notions of coherence.
- the *topic* or subject matter of the description, that is, a label describing the main concepts and their relationships within the description we are concerned with. We want to emphasize that the label describing a topic does not follow the same grammar rules as the elements of the description. For example, the topic addressed in the book by R. Pirsig can be described as "philosophycal discussion during motocycle riding". This concept description does not conform to the given grammar as it misses a predicate. For ontologies, we may think of a topic as a suitable selection of terms from the signature, organized in structured expressions such as "animals in terms of their energy sources and reproduction".

So, assume you want to automatically determine *the* topic of an ontology: this is clearly an underspecified task. Firstly, assume we have a high-quality ontology that we all agree is a coherent representation of the relevant concepts, regardless of how strict a notion of coherence we employ. Then, how do we represent its topic? Clearly, taking the set of all terms is unhelpful: it is probably too verbose and detailed, and also fails to distinguish between the *main* concepts, e.g., animals, and auxilliary concepts, e.g., their energy sources. Thus representing a topic also involves a measure of importance and means for structuring. Secondly, both for informal descriptions and for ontologies, we are confronted with different notions of coherence of varying granularity and with different properties. For example, it depends on our notion of coherence whether "Zen and the Art of Motorcycle Maintenance" is about motorcycle maintenance and, rather unrelated to that, about philosophy, or whether Pirsig has related both topics together so that the book is about both. Hence we are, in this paper mostly concerned with logic-based notions of coherence and leave the problem of how to represent a topic in a useful way for future work.

In what follows, we briefly introduce five logic-based notions of coherence that will later be discussed in more detail.

Signature independence The most coarse-grained notion of coherence is more easily defined in terms of indepence: two descriptions are independent if they talk about different things, i.e., if they do not share any (non-logical) terms. Thus a description is coherent if it cannot be partitioned into (independent) fragments whose signatures are disjoint. In general, we would assume each book to be coherent in this sense, and for its title to be some representation of its content.

Δ-signature independence Clearly, the above notion is so coarse-grained that most ontologies will be considered as one coherent whole, regardless of whether we could easily point out intuitively independent fragments. For example, a more detailed version of our ontology about animals and their energy sources might talk both about parts of animals involved in the metabolic process, e.g., their stomaches, and about parts of metabolic processes. If we use the same role, say hasPart to describe parthood, then anatomy and metabolism are a coherent whole w.r.t. to the first notion of coherence, whereas we might want to see them as two independent "subtopics". Hence a second notion of coherence can be obtained from the first one by loosening the signature disjointness condition and allowing independent fragments to share terms from a special, shared part of the signature, e.g., hasPart.

Natural independence Another notion of coherence can be identified by observing that, in our ontology, we may be talking about different *kinds* of things, e.g., about animals and (metabolic) processes, and that we describe the former in terms of the latter. This would allow us to decompose our ontology into a fragment about processes and one about animals, where the latter depends on the former. With this in mind, we can say that a fragment is coherent if it talks about one kind of things, and independent otherwise.

(Ir)relevance Clearly, the above notion is still rather coarse-grained in that it cannot distinghuish between, e.g., "organisms in terms of their metabolism" and

"organisms in terms of their reproduction", even though these topics might be largely independent: we can talk about the former without mentioning the latter, i.e., the latter does not come up naturally when talking about the former. In this sense, we can say that a fragment is coherent if it completely covers a set of terms, e.g., Organism and livesOf. In contrast to the former notion, this now allows us to consider the same things, e.g., organisms, as being the subject of independent topics.

Minimal (ir)relevance Now this last notion is a rather loose one: taking it literally, it would allow us to consider anything that "covers" a set of terms as a coherent whole, regardless of how loosely connected these terms are. It is hard to think of an ontology where the fragment of organisms and tax forms could be considered coherent: we would expect it to fall apart into two fragments. Thus a further condition leads us to our final notion of coherence: we say that a fragment is coherent if it completely covers a set of terms, e.g., Animal and hasChild, and is not simply the union of two or more such covering fragments.

Each of these notions of coherence allows us to determine *how many* topics an ontology is about. Some of them even give rise to a possibly interesting structure of topics, i.e., a topic can be a subtopic of another one.

4 Logic-based notions of coherence

In what follows, we relate the five notions of coherence sketched above with the corresponding logical formalisations.

4.1 Signature independence

The logical formalisation of this first notion of coherence was introduced by Parikh in 1999 in the context of Belief Revision [14]. The question addressed originally is essentially the following: if we want to revise a theory with a new piece of knowledge that contradicts some of what is entailed, do we have to check it against the whole theory? Or do we have some kind of safety that allows us not to touch those parts that are *independent* from this new finding? Hence Parikh was mainly concerned with independence, and thus his definition of coherence was only a byproduct. He formalises a way to split a logical theory $\mathcal{T}$ into independent parts each of which is, in a maximally fine-grained split, coherent.

Definition 1. *Let $\mathcal{O}$ be a logical theory over the signature $\widetilde{\mathcal{O}}$ and let $\{\widetilde{\mathcal{O}}_1, \widetilde{\mathcal{O}}_2\}$ be a partition of $\widetilde{\mathcal{O}}$. We say that $\widetilde{\mathcal{O}}_1, \widetilde{\mathcal{O}}_2$* split *the theory $\mathcal{O}$ if there are formulae α over $\widetilde{\mathcal{O}}_1$ and β over $\widetilde{\mathcal{O}}_2$ such that the logical closure of α and β concides with the logical closure of $\mathcal{O}$. In this case, we say that $\{\widetilde{\mathcal{O}}_1, \widetilde{\mathcal{O}}_2\}$ is a $\mathcal{O}$-splitting. In general, we say that (mutually disjoint) signatures $\widetilde{\mathcal{O}}, \ldots, \widetilde{\mathcal{O}}_n$ split $\mathcal{O}$ if there exist formulae $\alpha_i \in \widetilde{\mathcal{O}}_i$ for $i = 1, \ldots, n$ such that $\mathcal{O}$ is the logical closure of $\alpha_1, \ldots, \alpha_n$.*

Please note that Parikh's splitting is a *signature* splitting: it may be the case that $\mathcal{O}$ is a single formula, and thus we cannot identify any coherent *subsets* of $\mathcal{O}$, even though we can identify independent subsets of its signature: please note how, in first order logic, we can take any finite set of axioms and express it equivalently in a single axiom.

A nice property of $\mathcal{O}$-splittings is that they are unique.

Lemma 2. *Given a theory $\mathcal{O}$ over the signature $\widetilde{\mathcal{O}}$, there is a unique finest $\mathcal{O}$-splitting of $\widetilde{\mathcal{O}}$, i.e. one which refines every other $\mathcal{O}$-splitting.*

As mentioned above, Parikh's notion of coherence is rather loose: any ontology whose signature cannot be decomposed in *disjoint*, independent subsets is coherent, and would thus give rise to only a single topic.

4.2 Δ-signature independence

To obtain a more fine-grained view of topics of an ontology, we can loosen Parikh's notion of independence, i.e., we can say that any (part of an) ontology is coherent if its signature *apart from some special, common terms* cannot be decomposed in disjoint, independent subsets. These special, common terms can be distinctive of the whole area, but can be used in different contexts. They can be be considered as descriptive of patterns and methodologies of a theory, but not specific of a topic: roughly speaking, they belong to a "meta-topic". In our introductory example, hasPart was such a special term.

In [12] this approach has been formalized by introducing Δ-decompositions. The authors assert that some terms, like `hasPart` in our example, behave as logical symbols under certain points of view. So, the idea is to identify the set Δ of these terms in order to decompose the ontology into "signature-but-Δ"-disjoint sub-ontologies.

In contrast to Parikh's approach, we now have an approach that depends on the choice of Δ.

Definition 3. *Let $\mathcal{O}$ a finite theory of formulae in SO, $\Delta \subseteq \widetilde{\mathcal{O}}$ and $\mathcal{L}$ a fragment of SO. A partition $\Sigma_1, \ldots, \Sigma_n$ of $\widetilde{\mathcal{O}} \setminus \Delta$ is called a* signature Δ-decomposition *of $\mathcal{O}$ in $\mathcal{L}$ if there are $\mathcal{O}_1, \ldots, \mathcal{O}_n$ theories of formulae in $\mathcal{L}$ such that*

- $\widetilde{\mathcal{O}_i} \subseteq \Sigma_i \cup \Delta$ *for* $i = 1, \ldots, n$
- $\mathcal{O}_1 \cup \ldots \cup \mathcal{O}_n \equiv \mathcal{O}$.

$\mathcal{O}_1, \ldots, \mathcal{O}_n$ *is called a* realization *of the signature Δ-decomposition $\Sigma_1, \ldots, \Sigma_n$ in $\mathcal{L}$.*

This definition involves second order logic for technical reasons, namely to ensure that the realization of a Δ-decomposition always exists. As for Parikh's approach, there is a unique finest Δ-decomposition.

Theorem 4. *Let $\mathcal{O}$ a finite theory of SO formulae, $\Delta \subseteq \widetilde{\mathcal{O}}$, and let $\Sigma_1, \ldots, \Sigma_n$ and $\Pi_1, \ldots, \Pi_m$ be Δ-decompositions of $\mathcal{O}$ in SO. Then, the partition $\Sigma_i \cap \Pi_j$ for all i, j with $\Sigma_i \cap \Pi_j \neq \emptyset$ of $\widetilde{\mathcal{O}} \setminus \Delta$ is a Δ-decomposition of $\mathcal{O}$ in SO. Thus, there exists a unique finest Δ-decomposition of $\mathcal{O}$ in SO.*

Also as in Parikh's approach, we decompose a signature and not necessarily an ontology: given an ontology $\mathcal{O}$ in a given DL $\mathcal{L}$, it can be the case that there exists a suitable, non-trivial Δ-decomposition of $\mathcal{O}$ but we are not able to automatically compute corresponding subsets of $\mathcal{O}$ in the given DL. That is, in contrast to second order logic, other logics like DLs do not allow the so-called *unique decomposition realization* (UDR).

Another challenge is the suitable selection of a set Δ. As the authors say, they "do not expect signature decompositions to be a push-button technique, but rather envision an iterative and interactive process of understanding and improving the structure of an ontology, where the designer repeatedly chooses sets Δ and analyzes the impact on the resulting decomposition."

This remark leaves us with the open question: can we use this technique to extract the topics of an ontology? Since the decomposition depends on the selection of terms in Δ, the result obtained does not reflect an intrinsic logical coherence of topics – at least, we still do not have conditions to ensure such a property.

4.3 Natural independence

A totally different approach to the partitioning of an ontology $\mathcal{O}$ is carried out by identifying fragments that respect validity of axioms that they contain. In 1989 Garson [9] proposed that a *logical module* $\mathcal{M}$ should be:

- *logically correct*, i.e. any axiom entailed by $\mathcal{M}$ should be entailed by $\mathcal{O}$
- *logically complete*, i.e. any axiom over $\widetilde{\mathcal{M}}$ that is entailed by $\mathcal{O}$ should be entailed by $\mathcal{M}$

The intuition is that a logical module should preserve all the entailments that involve the signature the logical module "deals with". This means that different logical modules can share terms. In [5] the authors apply $\mathcal{E}$-connections to *decompose* ontologies, instead than to compose them as originally defined in [13]. The (computable) notion of module they are searching for is such that no subsumption relations exist between concepts (as in DLs, i.e. unary predicates) inside the module and concepts outside the module. This intuition leads to the following notion of module.

Definition 5. *A TBox* $\mathcal{M}_{\mathtt{A}} \subseteq \mathcal{O}$ *is a* module *for a concept* $\mathtt{A} \in \widetilde{\mathcal{O}}$ *if:*

- $\mathcal{M}_{\mathtt{A}}$ *is a logical module in* $\mathcal{O}$
- *for every concept* $\mathtt{B} \in \widetilde{\mathcal{O}}$*, the following holds:*
  (a) $\mathcal{M}_{\mathtt{A}} \models \{\mathtt{A} \sqsubseteq \mathtt{B}\} \iff \mathcal{O} \models \{\mathtt{A} \sqsubseteq \mathtt{B}\}$
  (b) $\mathcal{M}_{\mathtt{A}} \models \{\mathtt{B} \sqsubseteq \mathtt{A}\} \iff \mathcal{O} \models \{\mathtt{B} \sqsubseteq \mathtt{A}\}$
- *there are no concepts* $\mathtt{C}, \mathtt{D} \in \widetilde{\mathcal{O}}$ *such that* $\mathtt{C} \in \widetilde{\mathcal{M}_{\mathtt{A}}}$, $\mathtt{D} \notin \widetilde{\mathcal{M}_{\mathtt{A}}}$ *and either* $\mathcal{O} \models \mathtt{C} \sqsubseteq \mathtt{D}$ *or* $\mathcal{O} \models \mathtt{D} \sqsubseteq \mathtt{C}$.

To obtain such modules from an ontology, the authors describe a 3-steps algorithm: a safety-check, a partitioning algorithm, and the identification and extraction of modules.

The safety-check enforces a (mild) limitation in the ontologies that can be modularized. If an ontology is not safe, then this algorithm cannot be applied.

The partition algorithm, instead, aims at creating groups of concepts that can be interpreted independently from the other groups (see Theorem 3 in [5]). In particular, we obtain a partition of the *domain*, whose parts can be of three types: (Red) those which import vocabulary from others, (Blue) those whose vocabulary is imported, and (Green) isolated parts. Intuitively, this property means that either the parts correspond to actual non-overlapping subject matters, or the ontology is underspecified and some of the parts correspond to "unused information". In both cases, this seems to reflect a logical structure of the ontology. Moreover, any of these partitions can be automatically labelled with the highest common concept name in the concept hierarchy.

The step that identifies and extracts the modules is needed to determine which partitions are modules, or if an aggregation of some of these is necessary to ensure that what we have is a logical module. However, this step is irrelevant for the purpouses investigated in this paper. In Fig. 1 it is shown the partitioning of the toy ontology Koala,[3] that contains 42 axioms.

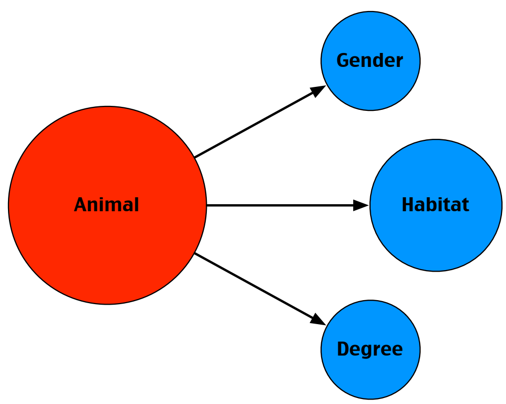

Figure 1. $\mathcal{E}$-connections of the ontology Koala

Using the procedure described, when $\mathcal{E}$-connections succeeds, it generates modules that correspond to an intuitive partition of the ontology. Moreover, it reveals also a structure: parts that need to import other parts to be defined. In particular, distinct parts can share part of the vocabulary, but no instances: for example, we can have a part dealing with `Pets` and another with `PetOwners`. Unfortunately, this technique fails sometimes in partitioning an ontology, and it returns a unique block, even if the ontology seems in principle well structured; for example, this happens with the ontology Periodic.[4]

[3] `http://protege.stanford.edu/plugins/owl/owl-library/koala.owl`
[4] `www.cs.man.ac.uk/~stevensr/ontology/periodic_full_06012009.owl`

4.4 (Ir)relevance

Another approach to the identification of topics within an ontology starts from a different point of view: the same individual can belong to different topics, as for example if we talk about a person, we can mean she in terms of her job, or she in terms of her family, etc. Intuitively, we think to these as different topics. In other words, a topic is identified not from a single label, but from the relationships between all the concepts it deals with.

In [2] the authors define *locality-based* modules: given an arbitrary set of terms Σ, such a module contains all axioms that "know everything" about Σ, where "know everything" depends on the notion of locality-based module used. These modules are an approximation of conservative extensions, as defined in Sect. 2: in other words, an ontology is always a conservative extension of any locality-based module. Among the strong logical properties that such modules satisfy, we have that they are logical modules as defined before. Moreover, such modules can be efficiently extracted. Hence, it would be interesting to investigate the family of all possible modules of an ontology $\mathcal{O}$.

By definition, locality-based modules can overlap. In principle this could lead to an exponential number (w.r.t. the size of the ontology) of modules; in [6] the trendline of the number of modules has been empirically studied for a selection of different ontologies, and the exponentiality of the family of modules seems to be confirmed by the experiments. Hence, studying the whole family of all locality-based modules is in general an infeasible task to be approached by brute force.

4.5 Minimal (ir)relevance

In [7] we introduced a new approach to represent the whole family $\mathfrak{F}_{\mathcal{O}}$ of locality-based modules of an ontology $\mathcal{O}$. The key point is observing that some axioms appear in a module only if other axioms do. In this spirit, we defined a notion of "logical dependence" between axioms: the idea is that an axiom α depends on another axiom β if whenever α occurs in a module $\mathcal{M}$ then β also belongs to $\mathcal{M}$.

Then, by using some notions of algebra, we have identified clumps of highly inter-related axioms, called *atoms*, defined to be maximal disjoint subsets of ontologies such that their axioms either appear always together in modules, or none of them does. In other words, atoms never split over two or more modules. Therefore, we can slightly extend the definition of "logical dependency" to atoms as in the following definition.

Definition 6. *Let $\mathfrak{a}$ and $\mathfrak{b}$ be two distinct atoms of an ontology $\mathcal{O}$. Then:*

- $\mathfrak{a}$ *is* dependent *on* $\mathfrak{b}$ *(written* $\mathfrak{a} \succeq \mathfrak{b}$*) if, for every module* $\mathcal{M} \subseteq \mathcal{O}$ *containing* $\mathfrak{a}$*, we have* $\mathfrak{b} \subseteq \mathcal{M}$*.*
- $\mathfrak{a}$ *and* $\mathfrak{b}$ *are* independent *if there exist two disjoint modules* $\mathcal{M}_1, \mathcal{M}_2$ *of* $\mathcal{O}$ *such that* $\mathfrak{a} \subseteq \mathcal{M}_1$ *and* $\mathfrak{b} \subseteq \mathcal{M}_2$*.*

- $\mathfrak{a}$ *and* $\mathfrak{b}$ *are* weakly dependent *if, they are neither independent, nor dependent; in such case, there exists an atom* $\mathfrak{c}$ *which both* $\mathfrak{a}$ *and* $\mathfrak{b}$ *are dependent on.*

Without loss of generality, we can remove from the ontology *syntactic tautologies*, i.e. always-local axioms, and *global axioms*, i.e. axioms that belong to all modules. We can always remove these unwanted axioms and consider them separately. As a consequence, the empty set is a module of the ontology. More importantly, the set of atoms is a partitioning of the ontology, hence linear w.r.t. its size.

The computation of the AD is polynomial w.r.t. the size of the ontology (provided that the extraction of a module is polynomial), and the algorithm to obtain the AD given an ontology and a (suitable) notion of module is discussed in [7]. The first (and fundamental) step can be described as follows: for each axiom α of the ontology, the algorithm takes as input its signature $\widetilde{\alpha}$, and returns the module $\top\bot^*$-mod$(\widetilde{\alpha}, \mathcal{O})$. These modules are non empty, since we already removed syntactic tautologies, and consequently at least α is non-local w.r.t. $\widetilde{\alpha}$. By definition, then, this module contains the atom $\mathfrak{a}$ that α belongs to. Moreover, the strong logical properties of locality-based modules imply that this module is the smallest that contains $\mathfrak{a}$. As a consequence, over the set $\mathcal{A}(\mathfrak{F}_\mathcal{O})$ of atoms, called *Atomic Decomposition* (AD), is induced a structure of partially ordered set. Hence the AD of an ontology can be represented by means of a Hasse diagram.

Beside covering all atoms, the modules we obtained following this procedure do not fall apart into two modules, and are also called *genuine*. As a consequence, they form a basis of the family of modules $\mathfrak{F}_\mathcal{O}$.

Definition 7. *A module is called* fake *if there exist two uncomparable (w.r.t. set inclusion) modules* $\mathcal{M}_1, \mathcal{M}_2$ *with* $\mathcal{M}_1 \cup \mathcal{M}_2 = \mathcal{M}$*; a module is called* genuine *if it is not fake.*

It is clear that there is a 1-1 correspondence between atoms and genuine modules. In particular, given an atom $\mathfrak{a}$ the corresponding genuine modules can be retrieved by considering all atoms $\mathfrak{a}$ is dependent on.

Definition 8. *The* principal ideal of an atom $\mathfrak{a}$ *is the set* $(\mathfrak{a}] = \{\alpha \in \mathfrak{b} \,|\, \mathfrak{b} \preceq \mathfrak{a}\} \subseteq \mathcal{O}$.

We summarize in the following table the ways described so far to look at ontologies' fragments.

Structure	$\mathcal{O}$	$\mathfrak{F}_\mathcal{O}$	$\mathcal{A}(\mathfrak{F}_\mathcal{O})$
Elements	axioms α	modules $\mathcal{M}$	atoms $\mathfrak{a}, \mathfrak{b}, \dots$
Maximal size	baseline	exponential	linear
Mathem. object	set	family of sets	poset

The decomposition of an ontology in atoms seems to capture a very fine-grained notion of coherence; however, in order to determine the topic of an atom we still need a way to identify a suitable label. Obviously, such suitable label will be chosen within the signature of the principal ideal of the atom itself. One possible way consists of labelling each atom $\mathfrak{a}$ with the vocabulary used in its axioms, and then, to express the "logical dependency", of recursively removing the terms already used in some atom that $\mathfrak{a}$ is dependent on. In this way, each atom is labelled only with the "new terms" introduced. Notice that in this case some atoms can have empty labels. In Fig. 2 is represented the Hasse diagram for the AD of the ontology Koala, whose labels are picked as described. Please note that the heights of the nodes vary: the heigher, the more numerous their axioms are.

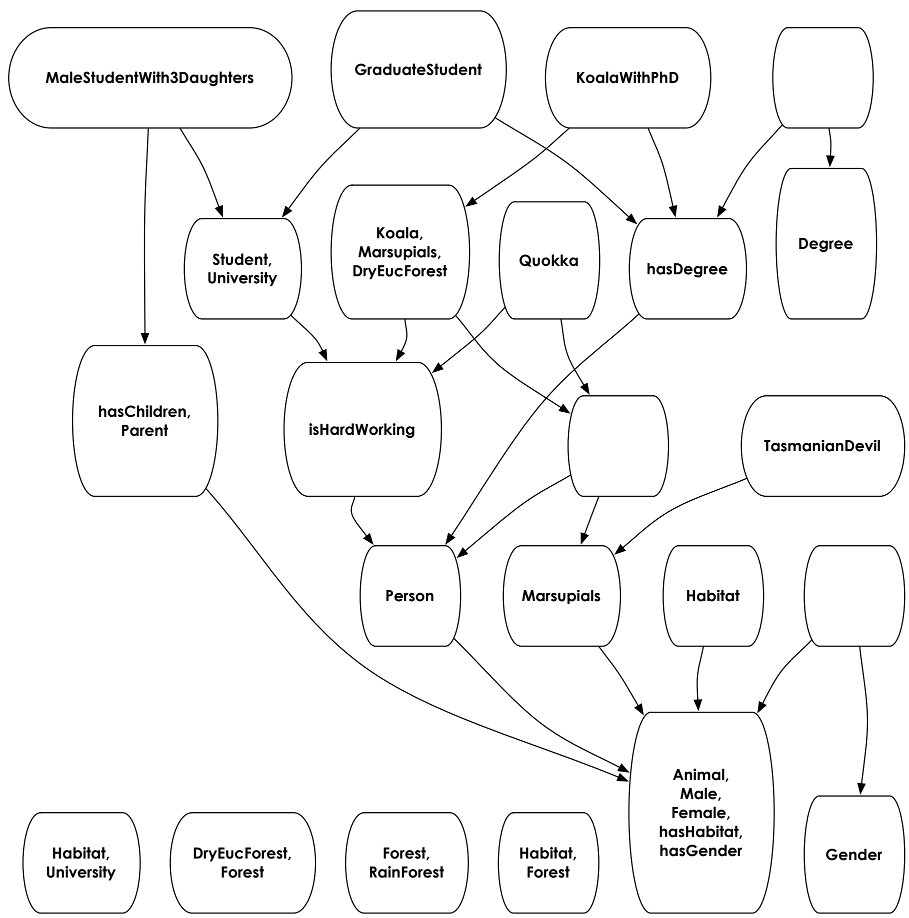

Figure 2. Locality-based decomposition of part of the ontology Koala

The intuitition suggests that the "core" atoms of an ontology can be recognized by looking at this structure, because many other atoms depend on them.

However, we still need to carry out real experiments to validate how cognitively significant this way of representing an ontology and of labelling it is.

5 Discussion and Outlook

The more an ontology reflects faithfully the knowledge about a domain, the more flexible and reusable it can be. However, ontologies are complex systems, since the tasks that ontology engineers have to perform are *per se* complex, and include the design, the implementation and the maintenance of ontologies. The design deals with the understanding of the domain, the implementation is a translation of the knowledge about the domain into a suitable ontology language, and the maintenance consists of updating the knowledge about the domain, as well as testing the consistency between the intended modelling of the domain and the actual modelling of the ontology.

Hence, the complexity of ontology engineering demands for the development of tools and methodologies to support engineers in manipulating ontologies. A crucial task is to support ontology engineers in the comprehension of ontologies. In particular, we have addressed the problem of topicality in ontologies, that is, "what the ontology is about", by means of modularity. However, to support comprehension of an ontology, we need to identify not only coherent fragments of the ontology but also their meaning.

The first two approaches we described here define coherence as "sharing terminology", and this, as we have seen, lead to a coarse decomposition of the ontology, the second being a refinement of the first. In principle, this can provide support in the understanding of the big picture of the ontology. Whilst the signature independence is straightforward to determine, the computation of Δ-independence can involve rewriting part of the ontology: complexity results for different languages are also carried out in [12]. However, both techniques are still merely theoretical, that is, there is no implementation, and we do not have any evaluation of their cognitive usefulness: these approaches can display a weakness in applicability if ontology engineers need only tools to support a fine tuning of the ontology. Moreover, since these methods do not involve any extraction of topic, the ontology engineers still have to look at all axioms of a partition. Finally, we recall that the selection of terms to be included in the set Δ is not automatic or guided: for any selection of terms, the engineer has to evaluate if the decomposition obtained matches with her understanding of the domain.

The approach based on $\mathcal{E}$-connections aims at identifying fragments of the ontology dealing with different kinds of individuals. The method has been implemented and it is available on the web.[5] In [5] the authors prove that the modularisation algorithm is polynomial in the size of the ontology. Moreover, when the algorithm succeeds in partitioning the ontology, the result corresponds to the intuition of users, and since it provides also labels, it helps in the understanding the structure of the ontology. Ontology engineers use this tool in real

[5] `http://www.mindswap.org/2004/SWOOP`

applications. However, it does not capture the finer-grained notion of topic, that occurs when we want to focus on individuals in terms of a specific aspects. And this is the intuitive reason for returning just one part – the ontology itself – for highly interrelated ontologies.

The extraction of a locality-based module is a well-understood and starting to be deployed in standard ontology development environments, such as Protégé 4,[6] and online.[7] It is used, for example, in the field for ontology reuse [11]: the modules extracted are quite small, and capture the knowledge of the ontology about the signature provided for the extraction. In [6] the authors tried to extract all such modules from ontologies in order to get insight into the modular structure of the ontology, but for medium size ontologies the algorithm, although highly optimised, did not succeed in the task. However, such modules capture the notion of topicality described as relationships between terms.

The atomic decomposition takes advantage of the nice logical properties of locality-based modules and separate parts of the ontology that show a minimal irrelevance. In [8] the authors prove that computing the atomic decomposition of an ontology is polynomial in the size of the ontology, provided that the extraction of a module is polynomial. Although methods for automatically labelling the atoms are future work, from a first evaluation it seems that this kind of decomposition can be used by ontology engineers in fine tuning ontologies. Moreover, this structure suggests a different notion of relevance from just counting the number of axioms of a part: an atom is more relevant if it is needed by many atoms. However, for big ontologies the decomposition looks too fragmented. This problem can be solved by using one of the previous methods, even if other formal techniques to group together some atoms into meaningful but coarser parts are included in our future work.

References

1. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press (2003)
2. Cuenca Grau, B., Horrocks, I., Kazakov, Y., Sattler, U.: Modular reuse of ontologies: Theory and practice. J. of Artif. Intell. Research 31, 273–318 (2008)
3. Cuenca Grau, B., Horrocks, I., Kazakov, Y., Sattler, U.: Extracting modules from ontologies: A logic-based approach. LNCS, vol. 5445, pp. 159–186. Springer-Verlag (2009)
4. Cuenca Grau, B., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P.F., Sattler, U.: OWL 2: The next step for owl. J. of Web Sem. 6(4), 309–322 (2008)
5. Cuenca Grau, B., Parsia, B., Sirin, E., Kalyanpur, A.: Modularity and web ontologies. In: Proc. of KR-06. pp. 198–209. AAAI Press (2006)
6. Del Vescovo, C., Parsia, B., Sattler, U., Schneider, T.: The modular structure of an ontology: an empirical study. In: Proc. of DL 2010. `ceur-ws.org`, `CEUR-WS.org/Vol-573/` (2010)

[6] `http://www.co-ode.org/downloads/protege-x`
[7] `http://owl.cs.manchester.ac.uk/modularity`

7. Del Vescovo, C., Parsia, B., Sattler, U., Schneider, T.: The modular structure of an ontology: Atomic decomposition. In: Proc. of IJCAI-11, accepted (2011)
8. Del Vescovo, C., Parsia, B., Sattler, U., Schneider, T.: The modular structure of an ontology: atomic decomposition. Tech. rep., The University of Manchester (2011), available at `http://bit.ly/i4olY0`
9. Garson, J.: Modularity and relevant logic. Notre Dame Journal of Formal Logic 30(2), 207–223 (1989)
10. Ghilardi, S., Lutz, C., Wolter, F.: Did I damage my ontology? A case for conservative extensions in description logics. In: Doherty, P., Mylopoulos, J., Welty, C.A. (eds.) Proc. of KR-06. pp. 187–197. AAAI Press (2006)
11. Jimeno, A., Jiménez-Ruiz, E., Berlanga, R., Rebholz-Schuhmann, D.: Use of shared lexical resources for efficient ontological engineering. In: SWAT4LS-08. `ceur-ws.org`, vol. 435 (2008)
12. Konev, B., Lutz, C., Ponomaryov, D., Wolter, F.: Decomposing description logic ontologies. In: Proc. of KR-10. pp. 236–246 (2010)
13. Kutz, O., Lutz, C., Wolter, F., Zakharyaschev, M.: $\mathcal{E}$-connections of abstract description systems. Artificial Intelligence 156(1), 1–73 (2004)
14. Parikh, R.: Beliefs, belief revision, and splitting languages. In: L. S. Moss, J.G., de Rijke, M. (eds.) Logic, language and computation, vol. 2, pp. 266–278. Center for the Study of Language and Information, USA (1999)
15. Stuckenschmidt, H., Klein, M.: Structure-based partitioning of large concept hierarchies. In: Proc. of ISWC-04. LNCS, vol. 3298, pp. 289–303. Springer-Verlag (2004)