Syntactic vs. Semantic Locality: How Good Is a Cheap Approximation?

Chiara Del Vescovo[1], Pavel Klinov[2], Bijan Parsia[1], Uli Sattler[1], Thomas Schneider[3], and Dmitry Tsarkov[1]

[1] University of Manchester, UK
{delvescc,bparsia,sattler,tsarkov}@cs.man.ac.uk
[2] University of Ulm, Germany
pavel.klinov@uni-ulm.de
[3] Universität Bremen, Germany
tschneider@informatik.uni-bremen.de

Abstract Extracting a subset of a given OWL ontology that captures all the ontology's knowledge about a specified set of terms is a well-understood task. This task can be based, for instance, on locality-based modules (LBMs). These come in two flavours, syntactic and semantic, and a syntactic LBM is known to contain the corresponding semantic LBM. For syntactic LBMs, polynomial extraction algorithms are known, implemented in the OWL API, and being used. In contrast, extracting semantic LBMs involves reasoning, which is intractable for OWL 2 DL, and these algorithms had not been implemented yet for expressive ontology languages.
We present the first implementation of semantic LBMs and report on experiments that compare them with syntactic LBMs extracted from real-life ontologies. Our study reveals whether semantic LBMs are worth the additional extraction effort, compared with syntactic LBMs.

1 Introduction

Extracting a subset of a given OWL ontology that captures all the ontology's knowledge about a specified set of concept and role names is an interesting task for various applications, and it is by now well-understood [2,10,11]. In general, we consider a setting where, for a given *signature*, we want to determine a (small) subset of a given ontology such that any axiom over the signature entailed by the ontology is also entailed by the subset. For expressive logics, this task can be implemented by making use of the notion of *locality*, and results in what is known as locality-based modules (LBMs) [2]. Locality comes in many different flavours, in particular there are notions of syntactic and semantic locality. A syntactic LBM is known to contain the corresponding semantic LBM, but might also contain extra axioms which are, because they are not in the semantic LBM, superfluous for entailments over the given signature. Algorithms for the extraction of syntactic LBMs are known that run in time that is polynomial in the size of the ontology (thus much cheaper than reasoning), implemented in the OWL

API, and being used. In contrast, despite the fact that algorithms for extracting semantic LBMs are known, until now and to the best of our knowledge, they had not yet been implemented. Moreover, these involve entailment checking, and are thus intractable for expressive profiles of OWL 2.

We present the first implementation of semantic LBMs and report on experiments that compare them with syntactic LBMs extracted from real-life ontologies. The contributions of this paper are as follows: we show with statistical significance that, for almost all members of a large corpus of existing ontologies, there is no difference between any syntactic LBM and its corresponding semantic LBM. In the few cases where differences occur, these differences are modest and not worth the increased computation time needed to compute semantic LBMs. In addition, we isolate two types of axioms that lead to differences, where one is a simple tautology that can, in principle, be detected by a straightforward addition to the syntactic locality checker. Furthermore, our results show that the extraction of semantic LBMs, which is in principle hard, seems feasible in practice. The lesson we learn from these results is that "Cheap is Great"!

2 Preliminaries

We assume the reader to be familiar with OWL and the underlying description logic $\mathcal{SROIQ}$ [1,8], and will define the central notions around locality-based modularity [2].

Let $\mathsf{N_C}$ be a set of concept names, and $\mathsf{N_R}$ a set of role names. A *signature* Σ is a set of *terms*, i.e., a set $\Sigma \subseteq \mathsf{N_C} \cup \mathsf{N_R}$ of concept and role names. We can think of a signature as specifying a topic of interest. Axioms that only use terms from Σ can be thought of as "on-topic", and all other axioms as "off-topic". For instance, if $\Sigma = \{\mathsf{Animal}, \mathsf{Duck}, \mathsf{Grass}, \mathsf{eats}\}$, then $\mathsf{Duck} \sqsubseteq \exists \mathsf{eats}.\mathsf{Grass}$ is on-topic, while $\mathsf{Duck} \sqsubseteq \mathsf{Bird}$ is off-topic.

Any concept, role, or axiom that uses only terms from Σ is called a *Σ-concept*, *Σ-role*, or *Σ-axiom*. Given any such object X, we call the set of terms in X the *signature of* X and denote it with $\widetilde{X}$.

Given an interpretation $\mathcal{I}$, we denote its restriction to the terms in a signature Σ with $\mathcal{I}|_\Sigma$. Two interpretations $\mathcal{I}$ and $\mathcal{J}$ are said to *coincide on a signature* Σ, in symbols $\mathcal{I}|_\Sigma = \mathcal{J}|_\Sigma$, if $\Delta^\mathcal{I} = \Delta^\mathcal{J}$ and $X^\mathcal{I} = X^\mathcal{J}$ for all $X \in \Sigma$.

There are a number of variants of the notion of conservative extensions, which capture the desired preservation of knowledge to different degrees. We focus on the deductive variant.

Definition 1. Let $\mathcal{M} \subseteq \mathcal{O}$ be $\mathcal{SROIQ}$-ontologies and Σ a signature.

(1) $\mathcal{O}$ is a *deductive Σ-conservative extension* (Σ-*dCE*) of $\mathcal{M}$ if, for all $\mathcal{SROIQ}$-axioms α with $\widetilde{\alpha} \subseteq \Sigma$, it holds that $\mathcal{M} \models \alpha$ if and only if $\mathcal{O} \models \alpha$.
(2) $\mathcal{M}$ is a *dCE-based module for* Σ of $\mathcal{O}$ if $\mathcal{O}$ is a Σ-dCE of $\mathcal{M}$.

Unfortunately, deciding in general if a set of axioms is a module in this sense is hard or even impossible for expressive DLs [6,12], and finding a minimal one

is even more so. However, "good sized" modules that are efficiently computable have been introduced [2]. They are based on the *locality* of single axioms, which means that, given Σ, the axiom can always be satisfied independently of the interpretation of the Σ-terms, but in a restricted way: by interpreting all non-Σ terms either as the empty set ($\emptyset$-locality) or as the full domain[4] (Δ-locality).

Definition 2. A $\mathcal{SROIQ}$-axiom α is called *$\emptyset$-local (Δ-local) w.r.t. signature Σ* if, for each interpretation $\mathcal{I}$, there exists an interpretation $\mathcal{J}$ such that $\mathcal{I}|_\Sigma = \mathcal{J}|_\Sigma$, $\mathcal{J} \models \alpha$, and for each $X \in \widetilde{\alpha} \setminus \Sigma$, $X^\mathcal{J} = \emptyset$ (for each $C \in \widetilde{\alpha} \setminus \Sigma$, $C^\mathcal{J} = \Delta$ and for each $R \in \widetilde{\alpha} \setminus \Sigma$, $R^\mathcal{J} = \Delta \times \Delta$).

It has been shown in [2] that $\mathcal{M} \subseteq \mathcal{O}$ and all axioms in $\mathcal{O} \setminus \mathcal{M}$ being $\emptyset$-local (or all axioms being Δ-local) w.r.t. $\Sigma \cup \widetilde{\mathcal{M}}$ is sufficient for $\mathcal{O}$ to be a Σ-dCE of $\mathcal{M}$. The converse does not hold: e.g., the axiom $A \equiv B$ is neither $\emptyset$- nor Δ-local w.r.t. $\{A\}$, but the ontology $\{A \equiv B\}$ is an $\{A\}$-dCE of the empty ontology.

Furthermore, locality can be tested using available DL-reasoners [2], which makes this problem considerably easier than testing conservativity. However, reasoning in expressive DLs is still complex, e.g. NExpTime-complete for $\mathcal{SHOIQ}$ [9,15]. In order to achieve *tractable* module extraction, a syntactic approximation of locality has been introduced in [2]. The following definition captures only the case of $\mathcal{SHQ}$-TBoxes and can straightforwardly be extended to $\mathcal{SHOIQ}$ ontologies.

Definition 3. An axiom α is called *syntactically $\bot$-local* ($\top$-*local*) *w.r.t. signature Σ* if it is of the form $C^\bot \sqsubseteq C$, $C \sqsubseteq C^\top$, $C^\bot \equiv C^\bot$, $C^\top \equiv C^\top$, $R^\bot \sqsubseteq R$ ($R \sqsubseteq R^\top$), or $\mathsf{Trans}(R^\bot)$ ($\mathsf{Trans}(R^\top)$), where C is an arbitrary concept, R is an arbitrary role name, $R^\bot \notin \Sigma$ ($R^\top \notin \Sigma$), and $C^\bot$ and $C^\top$ are from $\mathrm{Bot}(\Sigma)$ and $\mathrm{Top}(\Sigma)$ as defined in Part (a) (resp. (b)) of the table below.

(a) $\bot$-*Locality* — Let $A^\bot, R^\bot \notin \Sigma$, $C^\bot \in \mathrm{Bot}(\Sigma)$, $C^\top_{(i)} \in \mathrm{Top}(\Sigma)$, $\bar{n} \in \mathbb{N} \setminus \{0\}$

$$\mathrm{Bot}(\Sigma) ::= A^\bot \mid \bot \mid \neg C^\top \mid C \sqcap C^\bot \mid C^\bot \sqcap C \mid \exists R.C^\bot \mid \geqslant \bar{n}\, R.C^\bot \mid \exists R^\bot.C \mid \geqslant \bar{n}\, R^\bot.C$$
$$\mathrm{Top}(\Sigma) ::= \top \mid \neg C^\bot \mid C_1^\top \sqcap C_2^\top \mid \geqslant 0\, R.C$$

(b) $\top$-*Locality* — Let $A^\top, R^\top \notin \Sigma$, $C^\bot \in \mathrm{Bot}(\Sigma)$, $C^\top_{(i)} \in \mathrm{Top}(\Sigma)$, $\bar{n} \in \mathbb{N} \setminus \{0\}$

$$\mathrm{Bot}(\Sigma) ::= \bot \mid \neg C^\top \mid C \sqcap C^\bot \mid C^\bot \sqcap C \mid \exists R.C^\bot \mid \geqslant \bar{n}\, R.C^\bot$$
$$\mathrm{Top}(\Sigma) ::= A^\top \mid \top \mid \neg C^\bot \mid C_1^\top \sqcap C_2^\top \mid \exists R^\top.C^\top \mid \geqslant \bar{n}\, R^\top.C^\top \mid \geqslant 0\, R.C$$

It has been shown in [2] that $\bot$-locality ($\top$-locality) of an axiom α w.r.t. Σ implies $\emptyset$-locality (Δ-locality) of α w.r.t. Σ. Therefore, all axioms in $\mathcal{O} \setminus \mathcal{M}$ being $\bot$-local (or all axioms being $\top$-local) w.r.t. $\Sigma \cup \widetilde{\mathcal{M}}$ is sufficient for $\mathcal{O}$ to be a Σ-dCE of $\mathcal{M}$. The converse does not hold; examples can be found in [2].

For each of the four locality notions, modules of $\mathcal{O}$ are obtained by starting with an empty set of axioms and subsequently adding axioms from $\mathcal{O}$ that are Σ-non-local. In order for this procedure to be correct, the signature against which

[4] Or, in the case of roles, the set of all pairs of domain elements.

locality is checked has to be extended with the terms in the axioms that are added in each step, so that the resulting module $\mathcal{M}$ consists of all the non-local axioms with respect to $\Sigma \cup \widetilde{\mathcal{M}}$. Definition 4 (1) introduces locality-based modules, which are always dCE-based modules [2], although not necessarily minimal ones. Modules based on syntactic (semantic) locality can be made smaller by iteratively nesting $\top$- and $\bot$-extraction (Δ- and $\emptyset$-extraction), and the result is still a dCE-based module [2,13]. These so-called $\top\bot^*$-modules ($\Delta\emptyset^*$-modules) are introduced in Definition 4 (3).

Definition 4. Let $x \in \{\emptyset, \Delta, \bot, \top\}$, $yz \in \{\top\bot, \Delta\emptyset\}$, $\mathcal{O}$ an ontology and Σ a signature.

(1) An ontology $\mathcal{M}$ is the *x-module of $\mathcal{O}$ w.r.t. Σ* if it is the output of Algorithm 1. We write $\mathcal{M} = x\text{-mod}(\Sigma, \mathcal{O})$.
(2) An ontology $\mathcal{M}$ is the *yz-module of $\mathcal{O}$ w.r.t. Σ*, written $\mathcal{M} = yz\text{-mod}(\Sigma, \mathcal{O})$, if $\mathcal{M} = y\text{-mod}(\Sigma, z\text{-mod}(\Sigma, \mathcal{O}))$.
(3) Let $(\mathcal{M}_i)_{i \geqslant 0}$ be a sequence of ontologies such that $\mathcal{M}_0 = \mathcal{O}$ and $\mathcal{M}_{i+1} = yz\text{-mod}(\Sigma, \mathcal{M}_i)$ for every $i \geqslant 0$. For the smallest $n \geqslant 0$ with $\mathcal{M}_n = \mathcal{M}_{n+1}$, we call $\mathcal{M}_n$ the *yz^*-module of $\mathcal{O}$ w.r.t. Σ*, written $\mathcal{M} = yz^*\text{-mod}(\Sigma, \mathcal{O})$.

Algorithm 1 Extract a locality-based module

Input: Ont. $\mathcal{O}$, sig. Σ, $x \in \{\emptyset, \Delta, \bot, \top\}$ **Output:** x-module $\mathcal{M}$ of $\mathcal{O}$ w.r.t. Σ

$M \leftarrow \emptyset$; $\mathcal{O}' \leftarrow \mathcal{O}$
repeat
  changed $\leftarrow$ `false`
  for all $\alpha \in \mathcal{O}'$ **do**
    if α not x-local w.r.t. $\Sigma \cup \widetilde{\mathcal{M}}$ **then**
      $\mathcal{M} \leftarrow \mathcal{M} \cup \{\alpha\}$; $\mathcal{O}' \leftarrow \mathcal{O}' \setminus \{\alpha\}$; changed $\leftarrow$ `true`
until changed = `false`
return $\mathcal{M}$

As for (1), it has been shown in [2] that the output $\mathcal{M}$ of Algorithm 1 does not depend on the order in which the axioms α are selected.[5] Furthermore, the integer n in (3) exists because the sequence $(\mathcal{M}_i)_{i \geqslant 0}$ is decreasing (more precisely, we have $\mathcal{M}_0 \supset \cdots \supset \mathcal{M}_n = \mathcal{M}_{n+1} = \dots$). Due to monotonicity properties of locality-based modules, the dual notions of $\bot\top^*$- and $\emptyset\Delta^*$-modules are uninteresting because they coincide with those of $\top\bot^*$- and $\Delta\emptyset^*$-modules.

Roughly speaking, a Δ- or $\top$-module for Σ gives a view from above because it contains all subclasses of class names in Σ, while a $\emptyset$- or $\bot$-module for Σ gives a view from below since it contains all superconcepts of concept names in Σ.

Modulo the locality check, Algorithm 1 runs in time cubic in $|\mathcal{O}| + |\Sigma|$ [2]. Modules based on $\bot/\top$-locality are therefore a feasible approximation for modules based on $\emptyset/\Delta$-locality. In both cases, modules are extracted axiom by axiom

[5] Our algorithm is a special case of the one in [2, Figure 4].

but, as said above, the $\emptyset/\Delta$-locality check is more complex. A module extractor is implemented in the OWL API[6] and SSWAP[7]. To summarize:

1. Given an ontology $\mathcal{O}$, the semantic module $\mathcal{M}^{\text{sem}}_{\Sigma}$ for a signature Σ is contained in the corresponding syntactic module $\mathcal{M}^{\text{syn}}_{\Sigma}$ for the same seed signature.[8] This means that in principle more unnecessary axioms for preserving entailments over Σ can end up in syntactic modules rather than in semantic modules.
2. The extraction of a syntactic module can be done in polynomial time w.r.t. the size of the ontology $\mathcal{O}$. In contrast, the extraction of a semantic module is as hard as reasoning.

3 Experimental design

The main aim of this paper is to investigate how well syntactic locality approximates semantic locality. In particular, we want to see how (un)likely it is that syntactic locality-based modules are larger than semantic locality-based ones and how large these differences are. We also want to understand empirically how much more costly semantic locality is in terms of performance.

Selection of the Corpus. For our experiments, we have built a corpus containing: (1) from the TONES repository,[9] those ontologies that have already been studied in a previous work on modularity [4]: Koala, Mereology, University, People, mini-Tambis, OWL-S, Tambis, Galen; (2) all ontologies from the NCBO BioPortal ontology repository.[10]

We then filter out all those the ontologies for which at least one of the following problems occurs: the ontology is impossible to download; the .owl file is corrupted when downloaded; the file is not parseable; the ontology is inconsistent. Furthermore, due to time constraints, we exclude from this preliminary investigation all ontologies whose size exceeds $10,000$ axioms.

This selection results in a corpus of 156 ontologies, which greatly differ in size and expressivity [7], as summarized in Table 3. For a full list of the corpus, please refer to the Appendix.

Repository	Range Expr.	Range axs.	Range sig.
BioPortal	$\mathcal{ALCN}$-$\mathcal{SHIN}(\mathcal{D})/\mathcal{SOIN}(\mathcal{D})$	38 - 4,735	21 - 3,161
TONES	$\mathcal{AL}$-$\mathcal{SROIF}(\mathcal{D})/\mathcal{SHOIQ}(\mathcal{D})$	13 - 9,629	14 - 9,221

Table 1. Ontologies corpus

[6] `http://owlapi.sourceforge.net`
[7] `http://sswap.info`
[8] Recall that $\bot$-syntactic modules approximate $\emptyset$-semantic modules, while $\top$-syntactic modules approximate Δ-semantic modules.
[9] `http://owl.cs.manchester.ac.uk/repository/`
[10] `http://bioportal.bioontology.org`

Comparing Syntactic and Semantic Locality. In order to compare syntactic and semantic locality, we want to understand:

1. whether, for a given seed signature Σ, the semantic Σ-module is likely to be smaller than the syntactic Σ-module, and if so by how much,[11]
2. how feasible the extraction of semantic modules is.

Here, we focus on the two corresponding notions of $\emptyset$-semantic locality and $\bot$-syntactic locality. In particular, $\bot$-syntactic locality has been throughly investigated in previous work [3], and it has proven to have many interesting properties. A completion of the investigation described in this paper for all fundamental notions of modules is planned in our future work.

Due to the recursive nature of the locality-based module extraction algorithm, we want to investigate locality both on a

- per-axiom basis: given an axiom α and a signature Σ, is it likely that α is semantically $\emptyset$-local w.r.t. Σ but not syntactically $\bot$- local w.r.t. Σ?
- per-module basis: given a signature Σ, is it likely that $\bot\text{-mod}(\Sigma, \mathcal{O}) \neq \emptyset\text{-mod}(\Sigma, \mathcal{O})$? If yes, is it likely that the difference is large?

Hence we need to pick, for each ontology in our corpus, a suitable set of signatures, and this poses a significant problem. First, we do not yet have enough insight into what typical seed signatures are for module extraction. One could assume that large ones are rarely relevant for module extraction—why bother with extracting a large module—but this still leaves a large, i.e., exponential space of possible seed signatures. If $m = \#\widetilde{\mathcal{O}}$, there are 2^m possible seed signatures for which axioms can be tested for locality and for which modules can be extracted. Hence a full investigation is infeasible.

One could assume that the comparison between semantic and syntactic modules could be easier since many signatures can lead to the same module. In other words, the statistically significant number of modules w.r.t. the total number of modules is not larger than that of seed signatures needed w.r.t. the total number of seed signatures. In previous work [4,5], however, modules have been studied with respect to how numerous they are in real-world ontologies. The experiments carried out suggest that the number of modules in ontologies is, in general, exponential w.r.t. the size of the ontology. Moreover, the extraction of enough *different* modules can be hard, because by looking just at seed signatures there is no chance to avoid the extraction of the same module many times. In particular, for a module $\mathcal{M}$ there can be exponentially many seed signatures w.r.t. $\#\widetilde{\mathcal{M}}$ that generate $\mathcal{M}$ [3].

As a consequence, we compare the two kinds of locality of axioms—both on a per-axiom basis and a per-module basis—w.r.t. random signature. To avoid any bias, we select a random signature as follows: we set each named entity `E` in the ontology to have probability $p = 1/2$ of being included in the signature. Thus each seed signature has the same probability to be chosen. For ontologies whose signature exceeds 9 entities, in order to get get results where the true

[11] Recall that the semantic Σ-module is always a subset of the syntactic Σ-module.

proportion of differences between the two notions of locality lies in the confidence interval (±5%) with confidence level 95%, we have to select only 400 random signatures [14]. That is, we need to test only 400 random signatures to have a confidence of 95% (±5%) that the differences/equalities we observe reflect the real ones.

Non-random seed signatures. A module, in general, does not necessarily show any internal coherence: intuitively, if we had an ontology describing some knowledge from both the domains of Geology and of Philosophy, we can still extract the module for the signature $\Sigma = \{\texttt{Epistemology}, \texttt{Mineral}\}$. It is likely that the resulting module is going to be the union of the two disjoint modules for $\Sigma_1 = \{\texttt{Epistemology}\}$ and $\Sigma_2 = \{\texttt{Mineral}\}$. This combinatorial behaviour can lead to exponentially many modules in the size of the signature of the ontology and indeed, as mentioned above, the number of modules in ontologies seems to be exponential [4,5].

In contrast to *general* modules, *genuine modules* can be said to be coherent: they are defined as those modules that cannot be decomposed into the union of two different modules. Notably, there are only linearly many genuine modules in the size of the ontology $\mathcal{O}$, and the set of genuine modules is a base for all general modules. The linear bound on genuine modules is due to the fact that, for each genuine x-module $\mathcal{M}$, there is an axiom α such that $\mathcal{M} = x\text{-mod}(\tilde{\alpha}, \mathcal{O})$.

Thus genuine modules can be said to be interesting modules that we can fully investigate. Hence in addition to the above mentioned investigation of $\bot$- and $\emptyset$-modules for random signatures, we also look at all axiom signatures.

In summary, we test:

(T1) for random seed signatures Σ,
- **(a)** for each axiom α in our corpus, is α semantically $\emptyset$-local w.r.t. Σ but not syntactically $\bot$- local w.r.t. Σ?
- **(b)** is $\bot\text{-mod}(\Sigma, \mathcal{O}) \neq \emptyset\text{-mod}(\Sigma, \mathcal{O})$? If yes, we determine the difference and its size.

(T2) for each axiom signature from our corpus, is $\bot\text{-mod}(\tilde{\alpha}, \mathcal{O}) \neq \emptyset\text{-mod}(\tilde{\alpha}, \mathcal{O})$? If yes, we determine the difference and its size.

4 Experimental comparison

No differences. The main result of the experiment is that, for 151 of the 156 ontologies we tested, no difference between $\bot$- and $\emptyset$-locality can be observed. These 151 ontologies exclude the two NCBO BioPortal ontologies EFO (Experimental Factor Ontology) and SWO (Software Ontology), as well as Koala, miniTambis, and Tambis. More specifically, for every generated seed signature, the corresponding $\bot$- and $\emptyset$-module agree, and every axiom is either $\bot$- and $\emptyset$-local, or neither. This statement applies to all randomly generated seed signatures as well as for *all* axiom signatures – which are seed signatures for all genuine modules. We can therefore draw the following conclusions for the 151 ontologies with respect to (T1) and (T2) above.

(T1) Given an arbitrary seed signature Σ, there is no difference (a) between $\bot$- and $\emptyset$-locality of any given axiom w.r.t. Σ and (b) between the $\bot$- and $\emptyset$-modules for Σ, both times at a significance level of 0.05.
(T2) Given *any* axiom signature Σ, there is no difference between the $\bot$- and $\emptyset$-modules for Σ.

In the case of the 151 ontologies, the extraction of a $\emptyset$-module (with tautology tests performed by FaCT++) often took considerably longer than the extraction of the corresponding $\bot$-module. For example, for MoleculeRole, the largest of the 151 ontologies, times to extract a $\bot$-module (test all axioms for $\bot$-locality, respectively) ranged between 27 and 169ms (21 and 77ms, respectively), while the extraction of a $\emptyset$-module (test of all axioms for $\emptyset$-locality, resp.) took up to $6\times$ as long, on average $2.7\times$ ($2.0\times$, resp.). It is also worth noting that the ontologies Galen and People, which are renowned for having particularly large $\bot$-modules [2,5], are among those without differences between $\bot$- and $\emptyset$-locality.

Differences. For the five ontologies where differences between $\bot$- and $\emptyset$-modules (or -locality) occur, we isolated two types of culprits – axioms which are not $\bot$-local w.r.t. some signature Σ, but which are $\emptyset$-local w.r.t. Σ. Type-1 culprits are simple tautologies that have accidentally entered the "inferred view" – i.e., closure under certain entailments – of two ontologies. They do not occur in the original "asserted" versions and can, in principle, be detected by a slightly refined syntactic locality check. Type-2 culprits are definitions of concept names via a conjunction that satisfies certain conditions explained below. There are not many type-1 and type-2 axioms in the affected ontologies, and the observed differences are comparably small. Table 2 gives an overview of the differences observed.

Type-1 culprits are axioms `InverseObjectProperties(P, InverseOf(P))`, where `P` is a role. This translates into the tautology $\mathtt{P} \equiv (\mathtt{P}^-)^-$ in DL notation. Such an axiom is therefore $\emptyset$-local w.r.t. any signature. However, it behaves differently for $\bot$-locality: if the signature Σ contains `P`, then both sides of the equation are neither in $\mathrm{Bot}(\Sigma)$ nor in $\mathrm{Top}(\Sigma)$, hence the axiom is considered non-local; otherwise, both sides are $\bot$-equivalent, hence the axiom is local.

Type-1 axioms occur in the "inferred view" of the ontologies EFO and SWO. Table 2 shows the relatively modest differences caused by these axioms. In all cases, there are no other axioms in the differences. This means that no differences occur for the non-inferred original versions of EFO and SWO.

Type-2 culprits are complex definitions $A \equiv C$ of a concept name A where C is a disjunction that contains both a universal and an existential (or minimum cardinality) restriction on the same role. This affects the ontologies Koala, miniTambis, and Tambis. The effect is best illustrated for Koala, which contains exactly one such axiom, namely $\mathtt{M} \equiv \mathtt{S} \sqcap \forall \mathtt{c}.\mathtt{F} \sqcap \forall \mathtt{g}.\{\mathtt{m}\} \sqcap {=}3\,\mathtt{c}.\top$, where we have abbreviated the concept names `MaleStudentWith3Daughters`, `Student`, `Female`, the roles `hasChildren`, `hasGender`, and the nominal `male`. Now if the signature against which the axiom is tested for locality contains $\{\mathtt{S}, \mathtt{c}, \mathtt{g}\}$ but neither `M` nor

Ontology	#axs	#differences		difference sizes #axs	rel.	time ratio avg.	culprit type and frequency
SWO	3446	T1 a	400	6–22	0–1%	3.31	1 (30×)
		T1 b	400	23–29	1–2%	5.11	
		T2	3446	3–1	1–5%	5.86	
EFO	6008	T1 a	400	8–24	0–1%	1.42	1 (32×)
		T1 b	400	13–30	0–1%	1.38	
		T2	128	1–4	9–17%	—	
Koala	42	T1 a	0	0	0%	—	2 (1×)
		T1 b	2	1	3%	—	
		T2	0	0	0%	—	
miniTambis	170	T1 a	68	1–2	1–3%	—	2 (3×)
		T1 b	93	1–4	1–3%	—	
		T2	26	1–7	6–75%	—	
Tambis	592	T1 a	58	1–3	0–1%	3.31	2 (11×)
		T1 b	229	2–11	0–2%	5.01	
		T2	191	4–41	2–26%	—	

Table 2. Overview table of differences observed. The columns show: the ontology name; the overall number of axioms; the name of the test (see list on Page 7); the number of cases with differences; the number of axioms in the differences (absolute and relative to the $\bot$-case); the average time ratio $\emptyset : \bot$ ("—" indicates that no reliable statement is possible: the time for $\bot$ is only a few, often 0, milliseconds); the type of culprit present and the number of axioms of this type.

F, then this axiom is not $\bot$-local because none of the conjuncts on the right-hand side is in $\mathrm{Bot}(\Sigma)$. On the other hand, this axiom is a tautology when M and F are replaced by $\bot$: the conjunction $\forall \mathsf{c}.\bot \sqcap =3\,\mathsf{c}.\top$ cannot have any instances, regardless of how c is interpreted.

For Koala, this effect only causes two singleton differences between sets of local axioms for the randomly generated seed signatures, as shown in Table 2. For axiom signatures, there is no difference. Interestingly, this effect does not propagate to modules: for all signatures, $\bot$- and $\emptyset$-modules are the same. The reason might be that (a) g is used in many axioms and is thus very likely to contribute to the extended signature during module extraction, and (b) then the axiom defining F is no longer local, which "pulls" F into the extended signature, preventing the observed effect.

In miniTambis and Tambis, this effect is much stronger and affects a large proportion of modules, as shown in Table 2. The differences in these cases do not only consist of culprit axioms, but also of axioms that become non-local after the signature has been extended by the terms in the culprit axioms. Still, the size of the differences is mostly modest while, for Tambis, the $\emptyset$-locality test ($\emptyset$-module extraction) takes on average over three times (five times) as long as the $\bot$-locality test ($\bot$-module extraction).

5 Conclusion and Outlook

Summary. We obtain two main observations from the experiments carried out.

- In practice, there is no or little difference between semantic and syntactic locality. That is, the computationally cheaper syntactic locality is a good approximation of semantic locality.
- Though in principle hard to compute, semantic modules can be extracted rather fast in practice.

These results suggest that it is questionable to conclude that semantic locality should be preferred to syntactic locality. In terms of computation time, there is often a benefit in using syntactic locality: the average speed-up compared to the extraction of a semantic-locality based module is by a factor of up to 6. For some particular module pairs, it is higher by an order of magnitude. The gain in module size is zero or so small that it is hard to justify the extra time spent. In particular, there is no gain in size for the ontologies Galen and People, which are "renowned" for having disproportionately large modules [2,5].

Our results are interesting not only because they provide an evaluation of how good the cheap syntactic locality approximates semantic locality, but also because they enabled us to fix bugs in the implementation of syntactic modularity. For example, earlier data from the experiment have shown that reflexivity axioms had been treated incorrectly by the syntactic locality checker.

Future Work. It is evident that this work is preliminary. It investigates only the differences between the related notions of $\bot$- and $\emptyset$-locality. We plan to extend the same study to other notions of locality, in particular, nested modules ($\top\bot^*$- vs. $\Delta\emptyset^*$-modules) – these notions are the most economical in terms of module size. Moreover, we want to extend the investigation to the remaining larger ontologies in the BioPortal repository and further large ontologies, e.g., some versions of the NCI Thesaurus[12]. Preliminary results with a version outside the regular history show differences due to type-2 culprits, but we have not included them here because the differences disappear after removing axioms that were introduced due a problem with object and annotation properties when the ontology file is parsed by the OWL API. This behaviour is yet to be investigated and explained.

Another interesting extension is to modify the seed signature sampling. Currently, the random variable "size of the seed signature generated" follows the binomial distribution with expected value $m/2$ and variance $m/4$. Hence, most signatures in the sample have size around $m/2$; small and large signatures are underrepresented. For example, for one ontology with 915 terms, all signature sizes lay between 422 and 509. One might argue that, for big ontologies, the typical module extraction scenario does not require large seed signatures – but it does sometimes require relatively small seed signatures, for example, when a module is extracted to efficiently answer a given entailment query of typically small size.

[12] Downloadable from `http://evs.nci.nih.gov/ftp1/NCI_Thesaurus`

On the other hand, large modules resulting from larger seed signatures may be more likely to differ. We therefore plan an alternative seed signature sampling via bins for average signature sizes: repeat the current sampling procedure scaled to several subintervals of the range of possible signature sizes.

Our current results answer the question whether there is a significant difference between the two locality notions *with respect to a given signature*. It is also interesting to ask the same question relative to a given module. To answer it, the sampling of modules instead of seed signatures requires further investigation.

Acknowledgment. We thank Rafael Gonçalves for helpful comments.

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. Del Vescovo, C., Gessler, D., Klinov, P., Parsia, B., Sattler, U., Schneider, T., Winget, A.: Decomposition and Modular Structure of BioPortal Ontologies. In: Proc. ISWC-11 (2011)
4. Del Vescovo, C., Parsia, B., Sattler, U., Schneider, T.: The modular structure of an ontology: an empirical study. In: Proc. of WoMO-10. Frontiers in AI and Appl., vol. 211, pp. 11–24. IOS Press (2010)
5. Del Vescovo, C., Parsia, B., Sattler, U., Schneider, T.: The modular structure of an ontology: atomic decomposition and module count. In: Proc. of WoMO-11. Frontiers in AI and Appl., vol. 230, pp. 25–39. IOS Press (2011)
6. Ghilardi, S., Lutz, C., Wolter, F.: Did I damage my ontology? A case for conservative extensions in description logics. In: Proc. of KR-06. pp. 187–197 (2006)
7. Horridge, M., Parsia, B., Sattler, U.: The state of bio-medical ontologies. In: Proc. of 2011 ISMB Bio-Ontologies SIG (2011)
8. Horrocks, I., Kutz, O., Sattler, U.: The even more irresistible $\mathcal{SROIQ}$. In: Proc. of KR-06. pp. 57–67 (2006)
9. Horrocks, I., Sattler, U.: A tableau decision procedure for $\mathcal{SHOIQ}$. J. of Automated Reasoning 39, 249–276 (2007)
10. Konev, B., Lutz, C., Walther, D., Wolter, F.: Semantic modularity and module extraction in description logics. In: Proc. of ECAI-08. Frontiers in AI and Appl., vol. 178, pp. 55–59. IOS Press (2008)
11. Kontchakov, R., Wolter, F., Zakharyaschev, M.: Logic-based ontology comparison and module extraction, with an application to DL-Lite. Artificial Intelligence 174(15), 1093–1141 (2010)
12. Lutz, C., Walther, D., Wolter, F.: Conservative extensions in expressive description logics. In: Proc. of IJCAI-07. pp. 453–458 (2007)
13. Sattler, U., Schneider, T., Zakharyaschev, M.: Which kind of module should I extract? In: Proc. of DL 2009. `ceur-ws.org`, vol. 477 (2009)
14. Smithson, M.: Confidence Intervals. Quantitative Applications in the Social Sciences, Sage Publications (2003)
15. Tobies, S.: Complexity Results and Practical Algorithms for Logics in Knowledge Representation. Ph.D. thesis, RWTH Aachen (2001)

Appendix: overview of the ontologies used

Ontology	Expressivity	#Axioms	Sig. size
aba-adult-mouse-brain	$\mathcal{ALCI}$	3,441	915
adverse-event-reporting-ontology	$\mathcal{SHOIN}(\mathcal{D})$	574	503
african-traditional-medicine	$\mathcal{ALE}$	208	225
amino-acid	$\mathcal{ALCF}(\mathcal{D})$	477	52
amphibian-gross-anatomy	$\mathcal{ALE}$	2,673	1,647
anatomical-entity-ontology	$\mathcal{ALE}$	352	359
ascomycete-phenotype-ontology	$\mathcal{AL}$	294	329
basic-formal-ontology	$\mathcal{ALC}$	95	39
basic-vertebrate-anatomy	$\mathcal{SHIF}$	388	231
bilateria-anatomy	$\mathcal{ALEH}+$	138	121
bioinformatics-data-formats-identifiers...	$\mathcal{ALE}+$	3,803	2,844
biological-imaging-methods	$\mathcal{S}$	548	626
biomedical-resource-ontology	$\mathcal{SHIF}(\mathcal{D})$	681	672
biopax	$\mathcal{SHIN}(\mathcal{D})$	391	165
biotop	$\mathcal{SRI}$	680	404
birnlex	$\mathcal{AL}$	3,572	3,589
bleeding-history-phenotype	$\mathcal{ALCIF}(\mathcal{D})$	1,925	582
body-system	$\mathcal{AL}$	28	30
breast-tissue-cell-lines	$\mathcal{ALCH}(\mathcal{D})$	2,734	412
brenda-tissue-enzyme-source	$\mathcal{ALE}$	6,284	5,272
c-elegans-development	$\mathcal{AL}$	71	73
c-elegans-phenotype	$\mathcal{AL}+$	2,279	2,026
cao	$\mathcal{SHIQ}(\mathcal{D})$	476	290
cell-behavior-ontology	$\mathcal{ALUO}$	13	14
cell-type	$\mathcal{ALC}$	2,975	2,012
cereal-plant-development	$\mathcal{ALE}$	235	237
cereal-plant-gross-anatomy	$\mathcal{ALE}+$	1,839	1,173
cognitive-atlas	$\mathcal{ALC}$	3,622	1,585
common-anatomy-reference-ontology	$\mathcal{ALE}+$	54	54
common-terminology-criteria-for-adverse...	$\mathcal{AL}(\mathcal{D})$	6,940	3,889
dendritic-cell	$\mathcal{ALC}$	313	192
dikb-evidence-ontology	$\mathcal{ALCHOIN}(\mathcal{D})$	640	251
drosophila-development	$\mathcal{ALEH}+$	410	138
electrocardiography-ontology	$\mathcal{ALCIF}(\mathcal{D})$	1,274	1,171
environment-ontology	$\mathcal{S}$	1,807	1,574
epilepsy	$\mathcal{ALH}(\mathcal{D})$	145	148
event-inoh-pathway-ontology	$\mathcal{ALEH}+$	7,131	3,836
evidence-codes	$\mathcal{ALE}$	342	268
exo	$\mathcal{ALE}+$	85	121
experimental-factor-ontology	$\mathcal{ALHIF}+$	6,008	4,869
fda-medical-devices-2010	$\mathcal{AL}$	4,907	4,941
fly-taxonomy	$\mathcal{AL}$	6,587	6,599
flybase-controlled-vocabulary	$\mathcal{ALE}+$	659	771
fungal-gross-anatomy	$\mathcal{ALEI}+$	106	86
gene-regulation-ontology	$\mathcal{ALCHIQ}(\mathcal{D})$	962	544

Continued on next page

Table 3 – *Continued from previous page*

Ontology	Expressivity	#Axioms	Sig. size
general-formal-ontology	$\mathcal{SHIQ}$	212	86
hom-datasource_oshpd	$\mathcal{AL}$	351	361
hom-datasource_oshpdsc	$\mathcal{AL}$	351	360
hom-dxprocs_mdcdrg	$\mathcal{AL}$	774	784
hom-harvard	$\mathcal{AL}$	189	191
hom-icd9_procs_oshpd	$\mathcal{AL}$	4,642	4,652
hom-icd9cm-ecodes	$\mathcal{AL}$	1,490	1,500
hom-icd9cm_procedures	$\mathcal{AL}$	4,644	4,656
hom-mdcdrg_oshpd	$\mathcal{AL}$	773	782
hom-oshpd-sc	$\mathcal{AL}$	266	278
hom-oshpd_usecase	$\mathcal{AL}$	393	408
hom-procs2_oshpd	$\mathcal{AL}$	4,642	4,652
hom-ucare	$\mathcal{AL}$	64	75
hom_mdcs-drgs	$\mathcal{AL}$	774	780
homerun-ontology	$\mathcal{AL}$	1,194	1,094
host-pathogen-interactions-ontology	$\mathcal{SHI}$	403	319
human-developmental-anatomy-abs...	$\mathcal{ALE}$	2,335	2,316
human-developmental-anatomy-tim...	$\mathcal{ALE}$	8,339	8,343
human-disease	$\mathcal{AL}$	6,753	8,625
hymenoptera-anatomy-ontology	$\mathcal{SR}$	8,493	4,324
imgt-ontology	$\mathcal{ALCIN}(\mathcal{D})$	1,112	122
infectious-disease-ontology	$\mathcal{SROIF}$	1,221	640
information-artifact-ontology	$\mathcal{SHOIN}(\mathcal{D})$	294	197
interaction-network-ontology	$\mathcal{ALC}$	1,034	981
ixno	$\mathcal{AL}$	39	53
leukocyte-surface-markers	$\mathcal{AL}+$	472	473
linkingkin2pep	$\mathcal{SHIF}(\mathcal{D})$	30	17
lipid-ontology	$\mathcal{ALCHIN}$	2,375	762
loggerhead-nesting	$\mathcal{ALE}$	347	314
maize-gross-anatomy	$\mathcal{ALE}$	217	184
mass-spectrometry	$\mathcal{SH}$	4,447	4,492
medaka-fish-anatomy-and-dev...	$\mathcal{ALE}$	4,402	4,363
mego	$\mathcal{ALE}+$	421	370
minimal-anatomical-terminology	$\mathcal{ALE}$	504	481
molecule-role-inoh-protein-name...	$\mathcal{ALE}+$	9,629	9221
mouse-adult-gross-anatomy	$\mathcal{ALE}+$	3,776	2,984
mouse-pathology	$\mathcal{ALE}+$	808	757
multiple-alignment	$\mathcal{ALE}+$	168	174
neomark-oral-cancer-ontology	$\mathcal{SHIQ}$	399	352
neural-electromagnetic-ontologies	$\mathcal{SHIQ}(\mathcal{D})$	2,578	1766
neural-immune-gene-ontology	$\mathcal{SH}$	8,835	4,843
neuro-behavior-ontology	$\mathcal{AL}$	768	733
nif-dysfunction	$\mathcal{SROIF}(\mathcal{D})$	2,635	2,951
nmr-instrument-specific-component...	$\mathcal{AL}$	290	301
obo-relationship-types	$\mathcal{ALR}+$	33	26
ontology-for-drug-discovery-investigations	$\mathcal{SHOIN}(\mathcal{D})$	996	837
ontology-for-general-medical-science	$\mathcal{ALCO}$	216	162

Continued on next page

Table 3 – *Continued from previous page*

Ontology	Expressivity	#Axioms	Sig. size
ontology-for-genetic-interval	$\mathcal{SHIN}(\mathcal{D})$	509	298
ontology-for-microrna-target-prediction	$\mathcal{ALCI}(\mathcal{D})$	415	338
ontology-for-parasite-lifecycle	$\mathcal{SHOIF}$	855	415
ontology-of-general-purpose-datatypes	$\mathcal{ALCHOI}$	459	193
ontology-of-geographical-region	$\mathcal{AL}$	38	39
ontology-of-glucose-metabolism-disorder	$\mathcal{AL}$	132	132
ontology-of-medically-related-social-entities	$\mathcal{ALCO}$	157	99
ontology-of-physics-for-biology	$\mathcal{ALCHIQ}(\mathcal{D})$	795	545
pathogen-transmission	$\mathcal{AL}$	24	28
pediatric-terminology	$\mathcal{AL}$	894	891
phare	$\mathcal{ALCHIF}(\mathcal{D})$	459	312
phenotypic-quality	$\mathcal{SH}$	1,831	2,282
phylogenetic-ontology	$\mathcal{AL}$	77	83
physicalfields	$\mathcal{ALI}$	136	78
physico-chemical-process	$\mathcal{ALE}$	734	560
pilot-ontology	$\mathcal{ALCIF}(\mathcal{D})$	85	39
pko_re	$\mathcal{ALCF}$	771	770
plant-environmental-conditions	$\mathcal{AL}$	499	501
plant-growth-and-development-stage	$\mathcal{ALE}+$	240	285
plant-ontology	$\mathcal{S}$	2,215	1,460
plant-trait-ontology	$\mathcal{ALE}$	1,290	1,124
platynereis-stage-ontology	$\mathcal{ALE}$	31	18
protein-modification	$\mathcal{ALE}+$	1,986	1,346
protein-ontology	$\mathcal{ALCF}(\mathcal{D})$	689	226
protein-protein-interaction	$\mathcal{ALE}+$	1,007	962
pseudogene	$\mathcal{AL}$	19	23
quantitative-imaging-biomarker...	$\mathcal{ALUIF}(\mathcal{D})$	1,697	1,381
rat-strain-ontology	$\mathcal{ALE}$	4,122	3,004
reproductive-trait-and-phenotype...	$\mathcal{AL}$	91	96
sample-processing-and-sep...	$\mathcal{AL}$	193	194
sequence-types-and-features	$\mathcal{SHI}$	2,545	2,167
sleep-domain-ontology	$\mathcal{SHIF}(\mathcal{D})$	363	256
smoking-behavior-risk-ontology	$\mathcal{ALEI}+$	185	135
software-ontology	$\mathcal{SHOIQ}(\mathcal{D})$	3,446	1,039
spatial-ontology	$\mathcal{ALEHI}+$	235	172
spider-ontology	$\mathcal{ALE}+$	778	581
student-health-record	$\mathcal{ALH}(\mathcal{D})$	418	382
symptom-ontology	$\mathcal{AL}$	839	935
syndromic-surveillance-ontology	$\mathcal{ALIF}(\mathcal{D})$	1,679	364
sysmo-jerm	$\mathcal{SI}(\mathcal{D})$	417	280
systems-biology	$\mathcal{AL}$	587	558
systems-chemical-biology-chemogenomics	$\mathcal{SHIN}(\mathcal{D})$	489	216
taxonomic-rank-vocabulary	$\mathcal{AL}$	58	59
tick-gross-anatomy	$\mathcal{ALE}+$	948	630
tissue-microarray-ontology	$\mathcal{ALI}(\mathcal{D})$	60	32
tok_ontology	$\mathcal{SRIQ}(\mathcal{D})$	466	331
translational-medicine-ontology	$\mathcal{SRIN}(\mathcal{D})$	499	389

Continued on next page

Table 3 – *Continued from previous page*

Ontology	Expressivity	#Axioms	Sig. size
units-of-measurement	$\mathcal{ALE}$	343	336
units-ontology	$\mathcal{SHIF}$	105	88
vertebrate-anatomy-ontology	$\mathcal{ALER}+$	340	234
vertebrate-homologous-organ-groups	$\mathcal{ALE}+$	1,689	1,186
vertebrate-trait-ontology	$\mathcal{AL}+$	3,586	3,072
web-service-interaction-ontology	$\mathcal{ALER}+$	29	39
wheat-trait	$\mathcal{AL}$	175	176
xenopus-anatomy-and-development	$\mathcal{ALE}+$	2,243	1,051
yeast-phenotypes	$\mathcal{AL}$	266	300