Will My Ontologies Fit Together?

A preliminary investigation.

Bernardo Cuenca Grau, Ian Horrocks,
Oliver Kutz, and Ulrike Sattler
School of Computer Science, The University of Manchester, UK

1 Motivation

In realistic applications, it is often desirable to integrate different ontologies[1] into a single, reconciled ontology. Ideally, one would expect the individual ontologies to be developed as independently as possible, and the final reconciliation to be seamless and free from unexpected results. This allows for the modular design of large ontologies and facilitates knowledge reuse tasks. Few ontology development tools, however, provide any support for integration, and there has been relatively little study of these issues at a fundamental level. Understanding at this fundamental level would help us predict, for example, what logical consequences to expect from the integrated ontology, and whether the integration of ontologies preserves some desirable properties of its parts.

To the best of our knowledge, the problem of predicting and controlling the consequences of ontology integration has has been tackled only in [4]. The authors propose a set of reasoning services (with decidability and complexity results) to check whether, through integration with other ontologies, desirable properties of an ontology have been destroyed.

In this paper, we propose first steps towards a different approach, the so-called *normative* approach.[2] We specify certain properties that one would like to preserve in the integration and devise a set of restrictions that, when adhered to, guarantee to preserve these properties. Thus, while the approach of [4] determines the preservation of desirable properties ex post, our methodology prescribes some restrictions that guarantee the preservation of desirable properties. We introduce two 'integration scenarios' that, we believe, capture some of the common practices in ontology engineering, and postulate desirable properties that should be satisfied by the integrated ontology. We provide syntactic restrictions on the use of shared vocabulary that guarantee the preservation of

[1] Throughout this paper, we do not distinguish between ontologies and TBoxes.

[2] Thanks to Frank Wolter for coining this expression.

these properties. The two basic ontology integration scenarios we analyze here are the following:

1. *Foundational integration*: an ontology is integrated with a foundational (or "upper") ontology. The foundational ontology describes more general terms, and may be domain independent.
2. *Broadening integration*: two ontologies describing different (and largely independent) domains are integrated to cover a broader subject area.

We define, for each scenario, *semantic* properties that one might want to be satisfied by the integrated ontology, that is, we specify how the consequences of the integrated ontology relate to those from its parts. Next, we specify *syntactic* constraints on the ontologies to be integrated and show that they guarantee these properties: these syntactic constraints are referred to as *compliance* conditions. Furthermore, we sometimes need global semantic *safety* constraints on the ontologies used in the integration. In this paper, we use a condition called localness, which is identical to a condition found in [3].

Clearly, the syntactic constraints depend on the DL used in the ontologies, and mainly concern the way the symbols occurring in the different ontologies (their *signatures*) are used. Finally, we discuss whether these constraints are realistic for the scenario, i.e., whether users could be expected to stick happily to these constraints in order to ensure that the integrated ontology will satisfy the desired semantic properties. Since all constraints are purely syntactic, they are decidable in polynomial time, and preliminary tests indicate that many ontologies satisfy the constraints already.

We assume the reader to be familiar with the basics of description logics and use, throughout this paper, *axiom* for any kind of TBox, RBox, or ABox assertion, and $\mathsf{Sig}(\mathcal{T})$ for the set of concept and roles names in $\mathcal{T}$. This paper is accompanied by a technical report [2].

2 Integration Scenarios

Suppose that two ontologies $\mathcal{T}_1, \mathcal{T}_2$ are to be integrated in some application. The ontologies may be the result of a collaborative ontology development process and may have been designed in a coordinated way by different groups of experts, or they may have been simply "borrowed" from the Web. In any case, we assume that they have both been tested and debugged individually prior to the integration and, hence, are consistent and do not contain unsatisfiable concept names. To capture this notion, we call an ontology $\mathcal{T}$ *instantiable* if there exists a model $\mathcal{I}$ of $\mathcal{T}$ s.t. $A^I \neq \varnothing \neq R^{\mathcal{I}}$ for all concept and role names A, R in the

signature of $\mathcal{T}$.[3]

In the simplest case, one would construct an integrated ontology $\mathcal{T}$ by simply taking the *union* of the two ontologies $\mathcal{T}_1, \mathcal{T}_2$. In general, the ontologies $\mathcal{T}_1$ and $\mathcal{T}_2$ may be related and share symbols in their signatures $\mathsf{Sig}(\mathcal{T}_1)$ and $\mathsf{Sig}(\mathcal{T}_2)$.[4] We will first postulate the semantic properties that $\mathcal{T}$ should satisfy in order to capture the modeling intuitions of each scenario, and then to investigate "acceptable" syntactic restrictions on the $\mathcal{T}_i$ that make sure that $\mathcal{T}$ will behave as expected. The intuition is simple: the more liberal the syntactic constraints including the use of shared symbols, the more freedom is given to the modeler, but the less likely it is that the integrated ontology will behave as desired.

2.1 Foundational Integration

Often, interoperability between different *domain* ontologies $\mathcal{T}_{dom}$ and their data is achieved through the use of a *foundational* (or "upper") ontology $\mathcal{T}_{up}$. A well designed foundational ontology should provide a carefully conceived high level axiomatization of general purpose concepts. Foundational ontologies, thus, provide a structure upon which ontologies for specific subject matters can be based.

A prominent example of an ontology conceived as the integration of a foundational ontology and a set of domain ontologies is GALEN [8], a large medical ontology designed for supporting clinical information systems. The foundational ontology contains generic concepts, such as *Process* or *Substance*. The domain ontologies contain concepts such as *Gene* or *Research Institution*, which are specific to a certain subject matter. The domain ontologies in GALEN are connected to the foundational ontology through subsumption relations between concept and role names. For example, $\mathsf{Microorganism}$ in one of the domain ontologies is a subconcept of $\mathsf{Organism}$ in the foundational ontology: $\mathsf{Microorganism} \sqsubseteq \mathsf{Organism}$. Some prominent ontologies, such as CYC, SUMO and DOLCE have been designed specifically to be used in applications as foundational ontologies. For example, given a large dataset about chemicals annotated with concepts in a certain biomedical ontology, e.g., the National Cancer Institute Thesaurus (NCI) [5], one may want to annotate it semi-automatically with concepts of a different ontology. For such a purpose, one may align organic chemicals in NCI to substances in SUMO using the axiom: $\mathsf{Organic\_Chemical} \sqsubseteq \mathsf{Substance}$. Similarly, one may want to use a foundational ontology to generalize the roles of a given domain ontology. For example, a University ontology may use SUMO to

[3] For $\mathcal{T}$ a TBox in a logic whose models are closed under disjoint unions, such as $\mathcal{SHIQ}$, $\mathcal{T}$ is instantiable if all concept and role names in $\mathcal{T}$ are satisfiable.

[4] There may be some previous reconciliation w.r.t. symbols, e.g., to identify different symbols in the two ontologies that have the same intended meaning [7]. This is a separate problem, often referred to as *ontology alignment*, which we do not address here.

generalize the role writes as follows: writes $\sqsubseteq$ authors , where authors is defined in SUMO and does not occur in the University ontology.

Foundational ontologies are well-established ontologies that one does not control and, typically, does not fully understand. When one of these ontologies is borrowed from the Web and integrated in an application, it is especially important to make sure that the merge preserves their semantics. In particular, we do not want the classification tree in $\mathcal{T}_{up}$ to change as a consequence of the merge. This property can be conveniently formalized by the notion of a *conservative extension* [4].

Definition 1 *The TBox $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2$ is a* conservative extension *of $\mathcal{T}_1$ if, for every axiom α in the signature of $\mathcal{T}_1$: $\mathcal{T} \models \alpha$ implies $\mathcal{T}_1 \models \alpha$.*

Clearly, if $\mathcal{T}$ is a conservative extension of $\mathcal{T}_1$ and $\mathcal{T}_1, \mathcal{T}_2$ are consistent, then so is $\mathcal{T}$. However, conservativeness is a much stronger condition than instantiability: even if $\mathcal{T}$ is instantiable, new (and probably unintended) subsumptions between (possibly complex) concepts in $\mathcal{T}_1$ may still occur as a consequence of the merge.

In general, it may still be tolerable, and even desirable, to allow new subsumptions to occur in the domain ontology as a consequence of the integration and, in such a case, $\mathcal{T}$ will not be a conservative extension of $\mathcal{T}_{dom}$.

Also, the notion of a conservative extension is not sufficient to capture all the intended and unintended consequences. In particular, one would not expect concept names originally in $\mathcal{T}_{up}$ to be subsumed by concepts originally in $\mathcal{T}_{dom}$. In other words, the rôles of the foundational and domain ontologies should not be inverted after the merge. In contrast, new subsumptions may and should be entailed between concepts (respectively roles) in $\mathcal{T}_{dom}$ and concepts (roles) in $\mathcal{T}_{up}$. For example, since the shared concept Substance is subsumed by SelfConnectedObject in SUMO, it is expected that $\mathcal{T} = \mathcal{T}_{NCI} \cup \mathcal{T}_{SUMO}$ will entail the subsumption Organic_Chemical $\sqsubseteq$ SelfConnectedObject, where Organic_Chemical occurs in NCI, but not in SUMO, whereas SelfConnectedObject occurs in SUMO, yet not in NCI.

Next, we specify syntactic restrictions that will ensure these "nice" properties of $\mathcal{T} = \mathcal{T}_{up} \cup \mathcal{T}_{dom}$. Given the examples, it seems reasonable to limit the coupling between $\mathcal{T}_{up}$ and $\mathcal{T}_{dom}$ to subsumptions relating concept (role) names in $\mathcal{T}_{dom}$ and concept (role) names occurring in $\mathcal{T}_{up}$.

Definition 2 *The pair $\Im = \langle \mathcal{T}_{up}, \mathcal{T}_{dom} \rangle$ is* f-compliant[5] *if, for $\mathbf{S} = \mathsf{Sig}(\mathcal{T}_{up}) \cap \mathsf{Sig}(\mathcal{T}_{dom})$ the shared signature, concept and role names $A, R \in \mathbf{S}$ occur in $\mathcal{T}_{dom}$ only in axioms of the form $B \sqsubseteq A$ and $S \sqsubseteq R$ respectively, where $B, S \in \mathsf{Sig}(\mathcal{T}_{dom}) \setminus \mathbf{S}$.*

[5] "f" stands for "foundational".

f-compliance suffices for capturing the coupling between the foundational and the domain ontologies in GALEN. However, is f-compliance enough to guarantee our "nice" properties for $\mathcal{T} = \mathcal{T}_{up} \cup \mathcal{T}_{dom}$? A simple example will provide a negative answer: just assume that $\mathcal{T}_{dom}$ contains a GCI of the form $\top \sqsubseteq A$; after the merge, every concept in $\mathcal{T}_{up}$ will be subsumed by $A \in \mathsf{Sig}(\mathcal{T}_{dom})$ and, thus, the foundational ontology does not act as such anymore.

As mentioned above, we use an additional safety condition—called localness—which is defined as follows [3]: if $\mathcal{I} = (\Delta^{\mathcal{I}}, .^{\mathcal{I}})$ and $\mathcal{J} = (\Delta^{\mathcal{J}}, .^{\mathcal{J}})$ are interpretations such that $\Delta^{\mathcal{J}} = \Delta^{\mathcal{I}} \cup \nabla$, where ∇ is a non-empty set disjoint with $\Delta^{\mathcal{I}}$, $A^{\mathcal{J}} = A^{\mathcal{I}}$ for each concept name, and $R^{\mathcal{J}} = R^{\mathcal{I}}$ for each role name, then $\mathcal{J}$ is called the *expansion* of $\mathcal{I}$ with ∇. Intuitively, the interpretation $\mathcal{J}$ is identical to $\mathcal{I}$ except for the fact that it contains some additional elements in the interpretation domain. These elements do not participate in the interpretation of concepts or roles. Now, *local* ontologies are precisely those whose models are closed under domain expansions, i.e., $\mathcal{T}$ is *local* if, for every $\mathcal{I} \models \mathcal{T}$ and every set ∇ disjoint with $\Delta^{\mathcal{I}}$, the expansion $\mathcal{J}$ of $\mathcal{I}$ with ∇ is a model of $\mathcal{T}$.

Intuitively, local ontologies contain only GCIs with a limited "global" effect. Examples of non-local axioms are GCIs that fix the size of the domain in every model of the ontology (e.g. $\top \sqsubseteq \mathit{bob}$ for a nominal *bob*), or GCIs that establish the existence of a "universal" named concept (e.g. $\top \sqsubseteq \mathit{Car}$). In contrast, role domain and range and concept disjointness are local. In order to show that localness is a reasonable condition to impose, we have implemented a localness checker, tested it on about 800 ontologies available on the Semantic Web, and found that less than 1% of them contain non-local axioms. In [2], we provide the proofs of our initial results, a precise syntactic characterisation of localness for $\mathcal{SHIQ}$ and further details on our experimental results.

Theorem 1 *Let $\Im = \langle \mathcal{T}_{up}, \mathcal{T}_{dom} \rangle$ be f-compliant. If $\mathcal{T}_{dom}$ is a local $\mathcal{SHOIQ}$ TBox, $\mathcal{T}_{up}$ is a $\mathcal{SHIQ}$ TBox (not necessarily local), and $\mathcal{T} = \mathcal{T}_{up} \cup \mathcal{T}_{dom}$ is instantiable, then*

1. *$\mathcal{T} = \mathcal{T}_{up} \cup \mathcal{T}_{dom}$ is a conservative extension of $\mathcal{T}_{up}$,*
2. *$\mathcal{T} \not\models A \sqsubseteq B$, for all concept names $A \in \mathsf{Sig}(\mathcal{T}_{up})$ and $B \in \mathsf{Sig}(\mathcal{T}_{dom}) \setminus \mathbf{S}$, and*
3. *$\mathcal{T} \not\models R \sqsubseteq S$, for all role names $R \in \mathsf{Sig}(\mathcal{T}_{up})$ and $S \in \mathsf{Sig}(\mathcal{T}_{dom}) \setminus \mathbf{S}$.*

This theorem states that our desirable properties are indeed preserved after the merge and, most importantly, the rôles of the foundational and domain ontologies are preserved (Items 2 and 3). Note, however, that f-compliance does not suffice for ensuring the instantiability of the merge: only if $\mathcal{T}$ is consistent and free from unsatisfiable names do the guarantees provided by the theorem apply. Although instantiability, as opposed to conservative extensions, can be easily checked using a reasoner, one might want to strengthen the theorem (and

the corresponding syntactic restrictions) to ensure the instantiability of $\mathcal{T}$ as well. Also, note that localness certainly is a too restrictive safety condition since it rules out "harmless" GCIs as well. The investigation of new f-compliance conditions that ensure the instantiability of the integrated ontology and of less strict safety conditions is the focus of our ongoing work.

2.2 Broadening Integration

In this scenario, an ontology $\mathcal{T}_1$ is to be integrated with another $\mathcal{T}_2$ that describes in more detail one or more of the domains that are only touched on in $\mathcal{T}_1$. For example, we may wish to integrate the Wine Ontology [9] with an ontology describing, in more detail, the regions in which wines are produced or the kinds of grapes they contain.

The Wine Ontology illustrates a common pattern: although ontologies usually refer to a *core* application domain, they also refer to other *secondary* domains that deal with different objects. This modeling paradigm is not only characteristic of small and medium sized ontologies, but also occurs in large, high-quality knowledge bases, written by groups of experts. A prominent example is the NCI Thesaurus [5], a huge ontology covering areas of basic and clinical science. The core of NCI is focused on genes; other subject matters described in the ontology include diseases, drugs, chemicals, diagnoses treatments, professional organizations, anatomy, organisms, and proteins.

In this scenario, concepts in the core application domain can be defined in terms of concepts in the secondary domains. For example, in the Wine Ontology, a Bordeaux is described as a Wine produced in France, where France is defined in the Regions ontology: $\mathsf{Bordeaux} \sqsubseteq \mathsf{Wine} \sqcap \exists \mathsf{producedIn}.\mathsf{France}$ In NCI, the gene ErbB2 is an Oncogene that is found in humans and is associated with a disease called Adrenocarcinoma.

$$\mathsf{ErbB2} \sqsubseteq \mathsf{Oncogene} \sqcap \exists \mathsf{foundIn}.\mathsf{Human} \sqcap \exists \mathsf{associatedWith}.\mathsf{Adrenocarcinoma}$$

Concepts in secondary ontologies, however, do not use the core concepts in their definitions, i.e., regions are not defined in terms of wines or diseases in terms of genes. Note, in this connection, that a 'broadening scenario' in this interpretation is closely related to the way ontologies would be integrated using the framework of $\mathcal{E}$-connections, but is rather mimicking than directly adopting the syntax and semantics of $\mathcal{E}$-connections [6].

Ontologies following this pattern can evolve by expanding their domain of discourse with knowledge about new subject matters. For example, we may extend the Wine ontology by representing the kinds of dishes each wine is most appropriate for, or NCI by adding information about useful publications on cancer research. This evolution process will typically consist of adding a new

"secondary" ontology, either developed by a group of experts, or borrowed directly from the Web. As a consequence, this ontology should be "good" as it is, and thus we want to make sure that it will not be affected by the integration, i.e., we should require $\mathcal{T} = \mathcal{T}_{core} \cup \mathcal{T}_{side}$ to be a conservative extension of $\mathcal{T}_{side}$.

Furthermore, since we assume $\mathcal{T}_{core}$ and $\mathcal{T}_{side}$ to cover different aspects of the world, we require that the merged ontology $\mathcal{T}$ does not entail subsumptions in any directions between non-shared concept names $A \in \mathsf{Sig}(\mathcal{T}_{core})$ and $B \in \mathsf{Sig}(\mathcal{T}_{side})$. This condition ensures that the ontologies actually describe different objects.

Let $\mathcal{T}_{core}$ and $\mathcal{T}_{side}$ be ontologies with signatures $\mathbf{S}_{core} = \mathbf{C}_{core} \cup \mathbf{R}_{core}$ and $\mathbf{S}_{side} = \mathbf{C}_{side} \cup \mathbf{R}_{side}$, let the shared signature $\mathbf{S} = \mathbf{S}_{core} \cap \mathbf{S}_{side}$ contain only concept names, and let $\mathbf{R}_{out} \subseteq \mathbf{R}_{core}$ be a distinguished subset of roles. Intuitively, the roles in $\mathbf{R}_{out}$ connect objects in different ontologies. Some concepts in $\mathcal{T}_{core}$ are defined in terms of restrictions on these roles; for example, the Bordeaux wines are related to France via the role producedIn and the ErbB2 oncogenes with organisms and diseases through the roles foundIn and associatedWith, respectively.

Definition 3 *The pair $\Im = \langle \mathcal{T}_{core}, \mathcal{T}_{side} \rangle$ is* b-compliant *if:* ***1)*** *$\mathbf{S} = \mathbf{S}_{core} \cap \mathbf{S}_{side} = \mathbf{C}_{core} \cap \mathbf{C}_{side}$, $\varnothing \neq \mathbf{R}_{out} \subseteq \mathbf{R}_{core}$;* ***2)*** *for every role inclusion axiom $R \sqsubseteq S \in \mathcal{T}_{core}$, either both $R, S \in \mathbf{R}_{out}$ or both $R, S \notin \mathbf{R}_{out}$;* ***3)*** *for every GCI $C_1 \sqsubseteq C_2 \in \mathcal{T}_{core}$, C_1, C_2 can be generated using the following grammar:*

$$C_i \leftarrow A | C \sqcap D | \neg C_i | \exists R.C_i | \exists P.A' | \geq nR.C_i | \geq nP.A'$$

where $A \in \mathbf{C}_{core} \setminus \mathbf{C}_{side}$, C, D and C_i are concepts generated using the grammar, $A' \in \mathbf{C}_{side}$, $R \notin \mathsf{RoI}(\mathbf{R}_{out})$, and $P \in \mathbf{R}_{out}$.

As a consequence, concept names in $\mathcal{T}_{side}$ can *only* be used in $\mathcal{T}_{core}$ through restrictions on the "outgoing" relations. Condition **2)** makes sure that the hierarchies for the two kinds of roles are disconnected from each other. It turns put that the Wine Ontology and the "modules" that can be extracted from NCI [3] are local and b-compliant. As in the foundational scenario, the theorem requires the instantiability (and thus the consistency) of the merged ontology $\mathcal{T}$.

Theorem 2 *Let $\Im = \langle \mathcal{T}_{core}, \mathcal{T}_{side} \rangle$ be* b*-compliant. If $\mathcal{T}_{core}$ is a local $\mathcal{SHOIQ}$ TBox, $\mathcal{T}_{side}$ is a local $\mathcal{SHIQ}$ TBox, and $\mathcal{T} = \mathcal{T}_{core} \cup \mathcal{T}_{side}$ is instantiable, then*

1. *$\mathcal{T} = \mathcal{T}_{core} \cup \mathcal{T}_{side}$ is a conservative extension of $\mathcal{T}_{side}$,*
2. *For all $A \in \mathsf{Sig}(\mathcal{T}_{core}) \setminus \mathbf{S}$ and $B \in \mathsf{Sig}(\mathcal{T}_{side})$: $\mathcal{T} \not\models A \sqsubseteq B$ and $\mathcal{T} \not\models B \sqsubseteq A$,*
3. *For all $R \in \mathsf{Sig}(\mathcal{T}_{side})$ and $S \in \mathsf{Sig}(\mathcal{T}_{core})$: $\mathcal{T} \not\models R \sqsubseteq S$ and $\mathcal{T} \not\models S \sqsubseteq R$,*
4. *For all $R \in \mathbf{R}_{out}$ and $S \in \mathbf{R}_{core} \setminus \mathbf{R}_{out}$: $\mathcal{T} \not\models R \sqsubseteq S$ and $\mathcal{T} \not\models S \sqsubseteq R$.*

3 Outlook

So far, the problem of predicting and controlling the consequences of ontology integration has been largely overlooked by the Ontology Engineering and Semantic Web communities.

In this paper, we have formalized two basic scenarios for ontology integration. In each case, we have identified a set of *semantic* properties that the integrated ontology should satisfy and, under certain simplifying assumptions, we have shown how these properties can be guaranteed by imposing certain *syntactic* constraints on the ontologies to be integrated.

So far we have been very conservative in both the (syntactic) compliance and safety conditions (localness) in the scenarios. In the future, we aim at investigating how these can be relaxed in each case without losing the nice properties of the integrated ontology. We expect that our results will constitute the basis for a normative methodology for ontology integration that is both well-founded and understandable to modelers, and that can be supported by ontology development tools.

References

[1] F. Baader, C. Lutz, H. Sturm, and F. Wolter. Fusions of Description Logics and Abstract Description Systems. *JAIR*, 16:1–58, 2002.

[2] B. Cuenca-Grau, I. Horrocks, O. Kutz, and U. Sattler. Will my Ontologies Fit Together? Technical report, University of Manchester, 2006. Available at `http://www.cs.man.ac.uk/~bcg/Integration-TR.pdf`.

[3] B. Cuenca-Grau, B. Parsia, E. Sirin, and A. Kalyanpur. Modularity and Web Ontologies. In *Proc. of KR-2006*, 2006.

[4] S. Ghilardi, C. Lutz, and F. Wolter. Did I Damage My Ontology? A Case for Conservative Extensions in Description Logics. In *Proc. of KR-2006*, 2006.

[5] J. Golbeck, G. Fragoso, F. Hartel, J. Hendler, B. Parsia, and J. Oberthaler. The National Cancer Institute's Thesaurus and Ontology. *J. of Web Semantics*, 1(1), 2003.

[6] O. Kutz, C. Lutz, F. Wolter, and M. Zakharyaschev. $\mathcal{E}$-connections of Abstract Description Systems. *Artificial Intelligence*, 1(156):1–73, 2004.

[7] N. Noy. Semantic Integration: A Survey on Ontology-based Approaches. *SIGMOD Record*, 2004.

[8] A. Rector. Modularisation of Domain Ontologies Implemented in Description Logics and Related Formalisms, including OWL. In *Proc. of FLAIRS*, 2003.

[9] M.K. Smith, C. Welty, and D.L. McGuinness. OWL Web Ontology Language Guide. *W3C Recommendation*, 2004.