Theoretical and practical aspects of interfacing ontologies and lexical resources

Alessandro Oltramari, Laurent Prévot and Stefano Borgo

Laboratory for Applied Ontology, ISTC-CNR, Trento, Italy.
{oltramari,prevot,borgo}@loa-cnr.it

Abstract. During the last few years, a number of works aiming at interfacing ontologies and lexical resources have been initiated. This paper aims at clarifying the current picture of this domain. It compares ontologies built following different methodologies and analyses their combination with lexical resources. A point defended in the paper is that different methodologies lead to very different characteristics for the resulting systems. We classify these methodologies and show how actual projects fit into this classification. We also take a more applicative viewpoint by presenting a PROTÉGÉ-based platform that can be exploited in the general task of interfacing ontologies and lexical resources along the different methodological lines.

1 Introduction

During the last few years, ontologies and lexical resources have been put under the spotlight for dealing with various NLP tasks such as word sense disambiguation and bridging resolution. Interfacing ontology and computational lexicon[1] has been presented as a promising approach for Human Language Technologies (HLT), from classical NLP tasks to meaning negotiation in multi-agent systems. In this paper we aim at clarifying the populated landscape of the on-going initiatives in the domain. We will introduce in section 2 our methodology classification for combining ontologies and lexical resources. In the next section, we will survey some of the most popular top-level ontologies, namely DOLCE [1], OPENCYC[2] and SUMO [2]. These ontologies are quite different although this might not be evident to the newcomer. The purpose is to highlight the methodologies used to build them. In section 4, on the ground of the first two sections we will show how actual initiatives fit into our classification. The lexical resources considered in the paper are basically those of the WORDNET family [3]. Section 5 presents two examples of populating and aligning ontologies with WORDNET while section 6 shows a PROTÉGÉ-based platform that can be exploited in this area. We will conclude with some comments on multi-linguality issues.

[1] The terms “computational lexicon” and “lexical resource” are often used as synonyms in the literature.

[2] See *http://www.opencyc.org/releases/doc/*

2 Classifying experiments in ontologies and lexical resources

The main aim of interfacing ontologies and lexical resources is the development of *machine-understandable knowledge bases* to be used in Human Language Technologies. These knowledge bases are central for the next generation tools envisaged by the Semantic Web where knowledge sharing, information integration, interoperability and semantic adequacy are main requirements. Different methods may guide the linking of ontologies and lexical resources, depending on the final result one intends to achieve, namely to enhance the coverage of an ontology or to build a system comprising properties of an ontology and a lexical resource. A generalization of these tasks suggests the following methodological options:

(i) *restructuring* a computational lexicon on the basis of ontological-driven principles;
(ii) *populating* an ontology with lexical information;
(iii) *aligning* an ontology and a lexical resource.

Option (i) concentrates on the lexical resource and involves the ontology only at the "meta-level": the ontological restructuring is carried out following formal constraints of ontological design [4], for instance introducing the ontological distinction between *role* or *type* for concepts.

In option (ii) one maps lexical units to ontological entries focusing on the "object-level" [2]: in this case the formal constraints correspond to ontological categories and relations already implemented in an existing ontology. Roughly, a computational lexicon and an ontology are taken as bare taxonomies of terms, the former contains only lexicalised concepts (i.e. `substance`)[3] and linguistic relations (i.e. *hyponymy*) while the latter provides formal structure of both lexicalised and not-lexicalised concepts (i.e. AMOUNT-OF-MATTER) and relations (*part-of*). It is clear that this method has to include a comparative analysis of the ontology and the lexical resource in order to find bridging synonymous terms and possible homonyms.

Finally (iii), the most complete of the proposed approaches, collects both the "meta-level" and "object-level" character of the previous approaches in order to produce a system that is ontologically sound and linguistically motivated [5].

The experimental perspectives focused in this paper will show that ontologies and lexical resources generally keep their own peculiarities in the process of integration: in other words, neither (ii) nor (iii) bring to an actual *merging* of ontological properties and lexical information.[4] Although it is possible for different ontologies to be coherently merged in a new one - associating semantically similar concepts and finding the points of intersection [6] - the real benefit of integrating ontologies and computational lexicons follows from keeping them as *distinct layers of semantic information*, albeit improved by their mutual linkings and features. This is the main reason to call *alignment* (and not *merging*) for the most advanced interfacing method, i.e. (iii).

[3] In the paper we will stick to the following font convention: `typewriter` for WORDNET synsets, SMALLCAPS for ontology concepts and *italics* for relations.

[4] Of course method (i) is not considered since it provides only "ontological-driven principles" without any real ontological category or relation.

Both ontologies and lexical resources may be built around a taxonomic structure but generally they include other types of information as well. An *axiomatic ontology* like DOLCE [1] provides an axiomatisation of *part-of, constitution, dependence, participation*, which are non-hierarchical relations. A lexical resource like the Princeton WORDNET [3] is organised as a *semantic network*, whose nodes (sets of synonym terms) are bound together by several lexical and conceptual relations (besides *hyponymy/hyperonymy* we have meronymy, antonymy, causation, entailment and so on). This fact suggests the introduction of another dimension here called *constraint density*, which, as far as we know, has not been considered in the literature.

Constraints density captures the density of the "network of constraints" that holds between the concepts. It can be opposed to the *concept density* that situates ontologies from top-level to domain-level (see Fig. 1). *Constraint density* deals with non-hierarchical features of ontologies and lexical resources, like extension with axioms for dependence, participation and constitution, formalization of meronymy relation, translation of glosses into axioms and consistency checks (See for instance [7]).

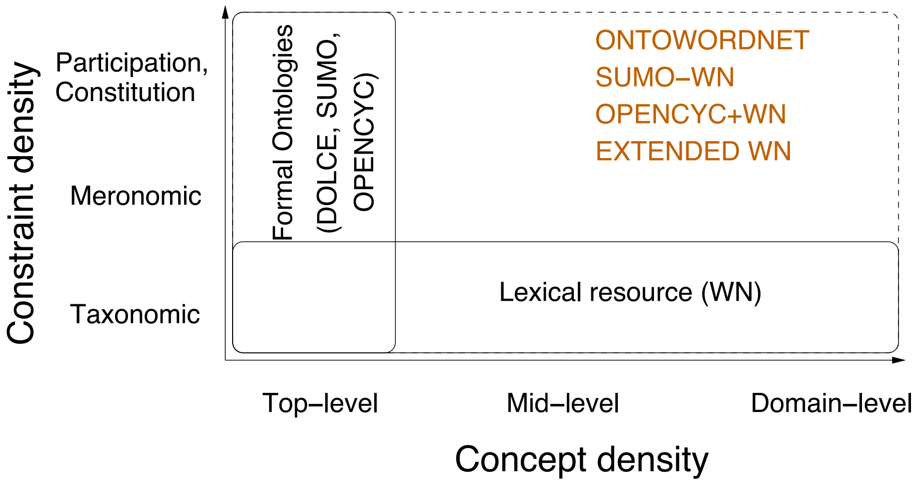


Fig. 1. Concept and constraint density

To make an analogy with the ontology development terminology, resources having very dense constraint network correspond to *heavyweight* ontologies while loose constraint network can be associated to *lightweight* ontologies [8, 9]. Lexical resources are conceptually very dense but they do not have a dense network of constraints. On the other hand, ontologies, specially top-level ones, are not densely populated but offer a dense network of constraints for their concepts.

A final remark on the nature of the lexical resources we look at. Although the experiments we consider in sections 4 and 5 concern the interfacing of ontologies with Princeton WORDNET, the methodologies we present here are general and apply to other resources like computational lexicons built on the basis of the original Princeton resource (i.e. EUROWORDNET[5] [10] modules). The three methods we isolated have not been applied to other types of lexical resources, for example FRAMENET[6]. We believe it is a question of time since nowadays in the literature "Wordnets" is the *de facto* standard for interfacing. We expect that further experience with different kinds of lexical resources will shed new light on the advantages and drawbacks of the three methodologies.

3 Ontologies and their construction

Ontology, as a branch of knowledge representation, is a young research area with several weaknesses among which the lack of established methodologies and of reliable evaluation criteria. Thus, one should not be surprised if the ontologies today available have been built following disparate approaches resulting in quite different systems. This is particularly evident in the area of *top-level ontology* by which, in this paper, we mean the research in *formal* and *foundational* ontologies. These ontologies are knowledge structures that (1) adopt a rich formal language (generally some kind of first-order logic) and (2) aim at classifying basic notions of general interest like process, event, object, quality, and so on.

Here we concentrate on three top-level ontologies, namely DOLCE, OPENCYC, and SUMO, that well indicate this variety of approaches. However, the main reason to focus on these is their attention to linguistic resources: these are the systems that have been used explicitly in relation to WORDNET.

3.1 DOLCE

DOLCE[7] (a Descriptive Ontology for Linguistic and Cognitive Engineering [1]) was released in its actual version in 2003 and has been constructed according to well documented philosophical principles. The content of the ontology is motivated from a cognitive viewpoint since the overall aim is to capture the ontological categories underlying everyday language and human commonsense. This view explains the adoption in DOLCE of a multiplicative approach which justifies the existence of co-localized (yet different) objects. For instance, DOLCE claims that a statue and the clay of which it is made, are different entities which share the same spatial (and possibly temporal) location. Co-localized entities are needed to consistently model linguistic expressions in which incompatible properties seem to be referred to the same object: a scratched statue is different (since scratched) and yet it is the same statue it was before. In DOLCE this is possible since the statue itself might not be affected by (minor) scratches, but the clay does because parts of it break up. DOLCE includes very basic and general notions only providing a total of about 40 categories which are richly axiomatized by using about a 100 relations and 80 axioms.

[5] See *http://www.illc.uva.nl/EuroWordNet/*

[6] Based on frame semantics [11]. See *http://framenet.icsi.berkeley.edu/*

[7] See *http://www.loa-cnr.it/DOLCE.html*

In the paper, we consider the “lite” version of DOLCE (aka DOLCE-LITE+), namely an extension of the axiomatic ontology that do not consider modality, temporal indexing, and relation composition. This version contains more concepts and allows for the implementation of DOLCE-based resources (i.e. the alignment of DOLCE and WORDNET called ONTOWORDNET) in languages that are less expressive than FOL e.g. OWL-DL, OWL-Lite, and RDF.

DOLCE is public resource and is released under the Lesser GNU Public License.

3.2 OPENCYC

OPENCYC is the ontology of CYC, a project initiated in 1984 with the aim of building a knowledge base comprising both scientific and commonsense knowledge. CYC grow to include hundreds of thousand elements between atomic terms, concepts, and axioms. To overcome consistency issues, CYC is now subdivided in hundreds of “microtheories”. Microtheories are, roughly speaking, bundles of assertions and rules in a specific domain of knowledge and are supposed to be locally consistent although not official claim is made in this sense. OPENCYC is a byproduct of CYC and was not part of the original project. Unfortunately the OPENCYC ontology has not been constructed according to philosophical principles nor following an ontological tested methodology. Indeed, still today the focus is on coverage: in the website one reads that OPENCYC includes “an upper ontology whose domain is all of human consensus reality”, which explains the 47,000 concepts and more than 300,000 assertions it contains, but makes one wonder what “upper” means here! Initially, it was obtained by isolating the taxonomy of the most general notions in CYC (perhaps with minor adjustments) but it was never followed by an ontological analysis and study of these notions. One can observe that OPENCYC adopts (at least in part) a cognitive viewpoint since some categories capture naïve conceptions of “reality”. For this reason, OPENCYC is compatible with the multiplicative approach (as seen in DOLCE) although this has not been followed in a systematic way. Since we lack a characterization of the ontological commitment and an analysis of the ontological choices embedded in the OPENCYC hierarchy there is not much to say about its ontological relevance. A further problem is the scarce axiomatization of OPENCYC which makes impossible to analyze the adequacy of the system in formal ontology.

OPENCYC is publicly available under the GNU Lesser General Public License.

3.3 SUMO

SUMO (Suggested Upper Merged Ontology [2]) began as a potpourri of theories in the knowledge representation area among which [12–15]. The ontology was created for computer applications (data interoperability, information retrieval, etc.) with no philosophical concerns and did not adopt ontological principles. This attitude is still present today (notwithstanding sporadic claims that SUMO is “rooted in metaphysical naturalism”), and the overall system is ontologically unclear as pointed out several times in the SUO mailing list.[8] Still, one can recognize some ontological choices in the system

[8] See *http://suo.ieee.org/index.html*

like the distinction between objects and events, and the adoption of a realistic approach. However, there is no guarantee that these have been consistently exploited in the whole ontology. The last version was released in 2005 and consists of about 4,000 assertions and 1,000 concepts. Several domain ontologies, linked to SUMO, are also available.

In the paper, we consider also the middle level ontology called MILO. MILO is written in the same language of SUMO and is provided as a "bridge" system between the general ontology and a number of domain ontologies. The latest version available on the web has been released in July 2004 and is marked "provisional and incomplete". We consider it since it is an integral part of the research in ontology and linguistic resources based on SUMO.

SUMO was initially distributed under the GNU Licence. Now it is subject to other restrictions;[9] in particular, SUMO "must not be utilized for any conformance/compliance purposes" and "[...] entities seeking permission to reproduce this document, in whole or in part, must obtain permission." However, it is claimed that these restrictions do not apply to research work.

4 How actual resources fit the classification

Generally speaking, projects interfacing ontologies and lexical resources are not easy to compare since often only generic statement are provided; the objectives are rarely addressed and the results are not homogeneously evaluated. Our classification of the methodologies is an attempt to put some order and to situate these resources. It is not meant to be a measure for ranking the resources.

4.1 ONTOWORDNET

The work underlying the ONTOWORDNET project is rooted in early proposals about upper levels of lexical resources [16]. More recent presentations can be found in [5, 17]. The program of ONTOWORDNET includes:

1 reengineering WORDNET lexicon as a formal ontology, and in particular:
  1.1 to distinguish synsets that can be formalized as classes from those that can be formalized as individuals;
  1.2 to interpret lexical relations from WORDNET as ontological relations.
2 aligning WordNet's top-level to the ontology by allowing re-interpretation of hyperonymy if needed;
3 consistency check of the overall result and consequent corrections;
4 learning and revising formal domain relations (from glosses or from corpora).

The first point corresponds to the restructuring task mentioned in section 2, points (2) and (3) deal with populating an ontology. Point (4) addresses the orthogonal issue of *constraint density* (axiomatizing the glosses).

The ONTOWORDNET project relies on the ONTOCLEAN methodology [4]. This methodology consists in determining the meta-properties of the given property. Very

[9] See *http://ontology.teknowledge.com/IEEE_license.htm*

roughly, a *rigid* property is a property that is essential to all its instances while a *non-rigid* property is not and an *anti-rigid* is essential to none of them. Some properties (called sortals) carry an *identity* criterion. A property ϕ can be said to be *dependent* on a property ψ if for all instances of ϕ some instance of ψ must exist (without being a part or a constituent).[10] Finally, another meta-property we will use in section 5 is *unity*: "a property ϕ is said to carry *unity* (+**U**) if there is a *common* unifying relation R such that all the instances of ϕ are essential wholes under R. A property carries *anti-unity* (∼**U**) if all its instances can possibly be non-wholes" [19].

In the second step of the methodology, one checks that a series of constraints on these meta-properties are satisfied. For example, unitarian properties cannot subsume anti-unitarian ones and properties subsuming rigid properties must be rigid themselves. Other constraints follows automatically from these. E.g. roles cannot subsume types. More precisely, from [18] roles are *non-rigid*, they do not supply their *identity criterion* but might carry one, and they are *dependent* on other properties. Types, on the other habd, are *rigid* and supply their own *identity criterion*. (The first version of ONTOWORDNET required the removal of roles from the ontology while the new version softens this constraint and requires only to label roles for separating them from types.)

This constraint checking is a crucial aspect of the ONTOWORDNET project. It is at this step that the lexical resource benefits from some ontological cleaning. ONTOWORDNET does not simply populate the top-level ontology by attaching WORDNET terms under ontology concepts. It determines which constraints have to be satisfied for integrating a WORDNET synset in an ontology in order to preserve its properties. ONTOWORDNET also claims that WORDNET itself benefits from the re-organization and from the application of the constraints. A full description of these constraints can be found in [5, 17]. Note that the re-structuration has been systematically performed only up to the third (somewhere fourth) upper level of WORDNET. The current ONTOWORDNET comprises now a re-structured and cleaned upper level, and a bare copy of WORDNET at the lower levels (without any ONTOCLEAN check).

Finally, the axiomatisation of WORDNET glosses (in the spirit of XWN as described in section 4.4) is an active area of research for the ONTOWORDNET project as shown in [7].

In conclusion, ONTOWORDNET is a costly methodology that hasn't been applied to the totality of WORDNET but that offers general rules to clean the lexical resource and populate the ontology. This methodology falls into the third category: an *alignment* between a lexical resource and an ontology.

4.2 OPENCYC and WORDNET

The next proposal we present is the integration of OPENCYC with WORDNET. The integration is obtained by adding in OPENCYC a synonym relationship between OPENCYC concepts and WORDNET synsets [20]. The purpose is to enrich the ontology with WORDNET information.

In our classification, this work falls into the *populating an ontology* option since there is no interest in restructuring the lexical resource nor in merging the two systems.

[10] For a detailed account see [18]. For an overview of ONTOCLEAN see [4].

4.3 SUMO-WN

We call the third approach SUMO-WN [21], i.e. the integration of SUMO with WORDNET. This integration has been performed for nouns, verbs, adverbs and adjective synsets. The result is a new resource whose entries are WORDNET synsets tagged by SUMO categories. At first sight, this project seems to address the three methodologies we identified: (i) Re-structuring a lexical resource (tagging WORDNET entries with SUMO categories might constitute a first step for re-structuring WORDNET), (ii) Populating an ontology (tagging also allows to present WORDNET synsets as synonyms, hyponyms and instances of SUMO concepts), (iii) Aligning an ontology and a lexical resource because SUMO-WN concerns both methodologies.

This brief description of SUMO-WN integration makes it sound very complete. However we need to look closer at the methodology in order to understand exactly what is done in SUMO-WN.

The result of the interfacing between SUMO and WORDNET is a list of synset annotated with SUMO concepts. The main task is therefore the annotation. In [21] three unproblematic annotation cases are presented:

- the WORDNET synset is a *synonym* of an existing SUMO concept
- the WORDNET synset is an *hyponym* of an existing SUMO concept
- the WORDNET synset is an *instance of* of an existing SUMO concept

Unfortunately, the examples given in [21] are a bit confusing as we will see in our discussion of practical example (section 5). The ontology has been recently improved but since our focus is on the methodology we look at problems that arise from its application disregarding subsequent *ad hoc* solutions.

Another problem in SUMO-WN is the absence of verification during the integration process. The quality of the resulting resource relies totally on the quality of WORDNET and SUMO. This is problematic since structural problems of WORDNET are now well-known and we saw in section 3 that the methodology for building SUMO jeopardizes its use as a well-founded reference for annotating the resource. We believe that a more careful restructuring of WORDNET is required before populating the ontology, and only then an annotation with SUMO concepts might have its interest. SUMO-WN links are rather *ad-hoc* and it is unlikely that such an approach can improve the accuracy of WORDNET or SUMO.

In conclusion, SUMO-WN addresses only the second category of our classification (*populating*) although the annotation of WORDNET entries could be seen as a preliminary step for re-structuring the resource. Moreover, since there is no clear methodology for determining how to perform the tagging, it seems not advisable to use this tagging for modifying the resource.

4.4 Axiomatizing glosses (EXTENDED WORDNET)

The EXTENDED WORDNET(XWN) project started with the objective of improving several weaknesses of WORDNET. These weaknesses are described in [22] and include in particular the need for more conceptual relations such as *causation* and *entailment* which are absent or not developed enough in WORDNET.

The proposal [23] consists in *"translating"* WordNet glosses into logical formulas with the help of natural language analysis. WordNet glosses are in a first step parsed to produce "logical forms". The second step consists in the transformation of the "logical form" into "semantic forms" by taking into account finer semantic aspects such as thematic relations. WordNet glosses eventually become axioms that can be manipulated in a more precise and efficient way than current natural language glosses. Furthermore, the disambiguation of the terms in the glosses and their systematic linking to other WordNet entries or to terms in other glosses, augment dramatically the connectivity between WordNet synsets.

This work is very promising and is complementing the approaches presented in sections 4.1 and 4.3 which at this point provides mainly taxonomic axioms.

In our terminology, XWN wants to increase the constraint density since the axioms derived from this method are potentially of all types. XWN is not properly speaking proposing to interface an ontology and a lexical resource because it does not involve explicitly an existing ontology. Since the ontological input is only implicit, XWN does not enter into our classification. However, if this ontological input was coming from an existing ontology, XWN would belong to the *re-structuring* methodological option.

	Level	**Examples**
Re-structurating	Meta	OntoClean
Populating	Object	OPENCYC, SUMO-WN
Aligning	Object&Meta	OntoWordNet

Fig. 2. Methodology classification

4.5 Summary

The result of our classification is summarized in Figure 2. Among the initiatives we looked at, OPENCYC is a clear example of a *populating* methodology, SUMO-WN falls also into this category while OntoWordNet includes both the *re-structuring* methodology through the application of OntoClean and the *populating* one by linking WordNet synsets to DOLCE-LITE+ categories. Finally, SUMO-WN offers a complete integration of WordNet while OntoWordNet and OPENCYC are, for different reasons, incomplete.

5 Two practical examples

5.1 `Christian_Science` and `Underground_Railroad` examples

The first example comes from the SUMO-WN presentation [21]. It concerns the *hyponym* case. It is claimed that the SUMO concept ReligiousOrganization is a hypernym of WordNet synset `Christian_Science` *(gloss: "religious system based on the teachings of Mary Baker Eddy emphasizing spiritual healing").*

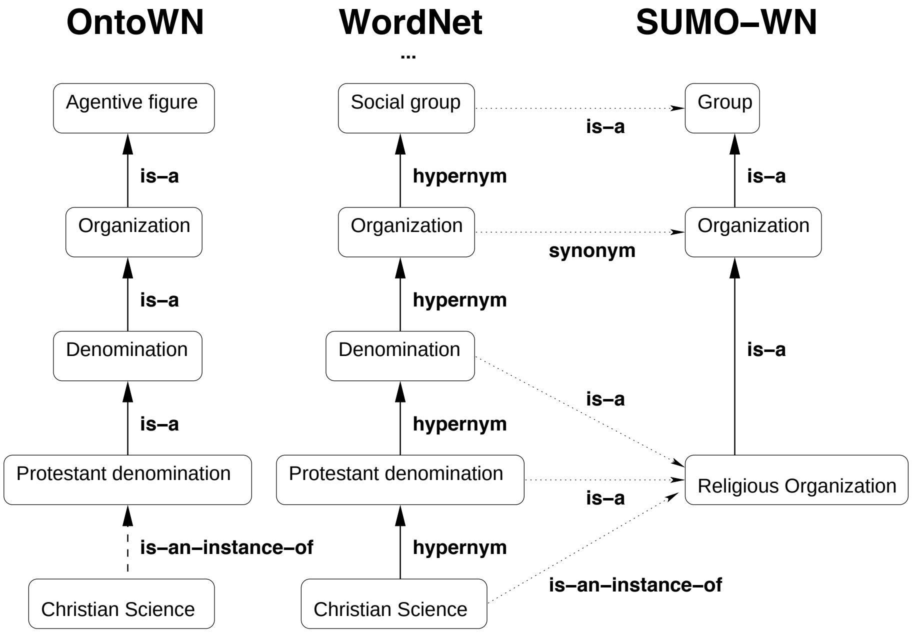


Fig. 3. `Christian_Science` example

Since RELIGIOUSORGANIZATION are ORGANIZATIONS, there is no clear reason for setting `Christian_science` to be an organization because SUMO organizations are *"corporate or similar institutions (...)"*. The corresponding category for `Christian_Science` should be something like *Christian_Science_Church*. There are actually another WORDNET synset for `Christian_Science` *(gloss: "Protestant denomination founded by Mary Baker Eddy in 1866.")*. But even accepting this conceptual shortcut, it is still not clear why `Christian_Science` is an *hyponym* of RELIGIOUSORGANIZATION and not an *instance-of* it. For `Christian_Science` to be a sub-type of ORGANIZATION, there must be at least two instances of `Christian_Science`. The WORDNET gloss describes it more as a general doctrine and therefore as an instance of something like a *Religious_System*. The example provided fits better the second WORDNET synset.

There is a lack of information about the notions of religious systems and organization to pursue further the investigation but it is clear to us that the choices that have been made in SUMO on these topics are dubious. The current version we can find online[11] corrected some of these problems as we can see in Figure 3. In the downloadable file, we can find both synsets. The first one is now an *instance-of* RELIGIOUSORGANIZATION while the second one is a sub-type of SUMO's PROPOSITION.

The example provided for illustrating the *instance-of* case is very similar to the previous one. In this case, `UndergroundRailroad` *(gloss: abolitionist secret aid to*

[11] See *http://ontology.teknowledge.com/*

escaping slaves) is taken to be an *instance-of* and not an *hyponym of* SUMO ORGANIZATION. In the end, it remains difficult to understand the methodology adopted to classify terms in these two examples, one wonders if the tagging relies essentially on the intuitions of the SUMO-WN developers.

In ONTOWORDNET `Christian_Science` and its hyperonyms are integrated in the resource as shown in Figure 3. The hierarchy corresponds to that of WORDNET up to the top-level. About the first sense, the last WORDNET hypernym is `Organization` and there is ORGANIZATION present in DOLCE-LITE+. The second sense is more tricky because of a double inheritance in the WORDNET hierarchy.

Regarding the `Underground_Railroad`, the ONTOWORDNET version proposed it as a subtype of `Escape`. It is a clear example that shows that the application of the methodology is incomplete in the current version of ONTOWORDNET. Because of its development cost, the checking and the restructuration of WORDNET couldn't go deeper than the first four upper levels of the hierarchy. As a result, `Underground_Railroad` hasn't been checked and therefore not corrected yet in ONTOWORDNET.

5.2 `Cement` example

The second example concerns the need for WORDNET restructuration (Figure 4). In WORDNET `cement` (*gloss: "a building material that is a powder made of a mixture of calcined limestone and clay") is situated under* `building_material` *and further under* `artifact` *(see Fig. 4). On the meta-properties level,* `Artefact` *presents therefore both unitarian concept such as regular artefacts (*`chair, hammer,...`*)* and non-unitarian object such as `cement`. This constitutes a formal violation in terms of ONTOCLEAN methodology.

In SUMO-WN this violation is repeated since `building_material` *is-a* SELFCONNECTEDOBJECT, which is an unitarian concept (+**U**) and SELFCONNECTEDOBJECT include FOOD which subsumes itself BEVERAGE, that is clearly non-unitarian (∼**U**).

On the other hand, ONTOWORDNET performs a re-structuration at this level which forces to distinguish unitarian and non-unitarian concepts as explained in [5]. `building_material` is therefore removed from the `artefact` category and put under FUNCTIONAL-MATTER which *is-subsumed* by AMOUNT_OF_MATTER (∼**U**). The `artefact` synset is put under ORDINARY_OBJECT. Finally, we do not discuss, specific examples involving OPENCYC for lack of public material.

6 A PROTÉGÉ-based platform for interfacing ontologies and lexical resources

So far we have described the general methodologies that underlie the process of interfacing ontologies and computational lexicons and we have shown concrete examples of populating and aligning distinct ontologies with WORDNET. Here we take an application perspective and discuss a PROTÉGÉ-based platform that can be exploited for such tasks.

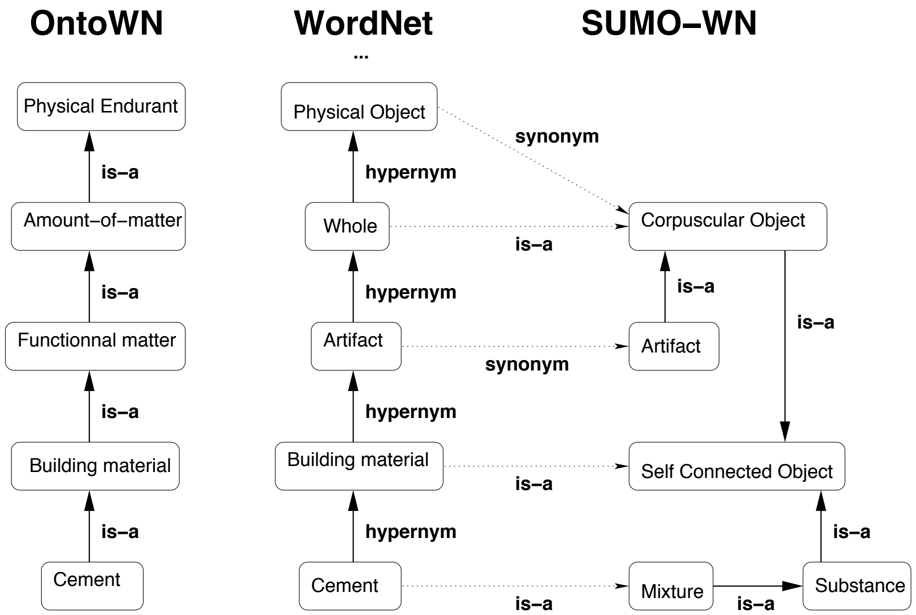


Fig. 4. `Cement` example

PROTÉGÉ[12] is the most successful tool for creating, editing and visualizing ontologies; and the recent implementation of an OWL plug-in makes PROTÉGÉ a *de facto* standard for Semantic Web research and development. The platform we are using here is constituted by three modules/stages: the first one concerns the creation (and possibly the import) of an ontology and exploits the standard PROTÉGÉ interface (release 3.1) together with the above-mentioned OWL plug-in; the second one deals with the process of augmenting the ontology with a lexical resource using ONTOLING, a tool for PROTÉGÉ created by the University of Rome "Tor Vergata"[13]; finally, the third one adopts "PAL Constraints"[14] to implement ONTOCLEAN metaproperties and to check for formal violations throughout the considered lexicon. In the next paragraphs we focus on the last-two issues only since the first concerns standard practice in PROTÉGÉ.

6.1 Enriching ontologies with ONTOLING

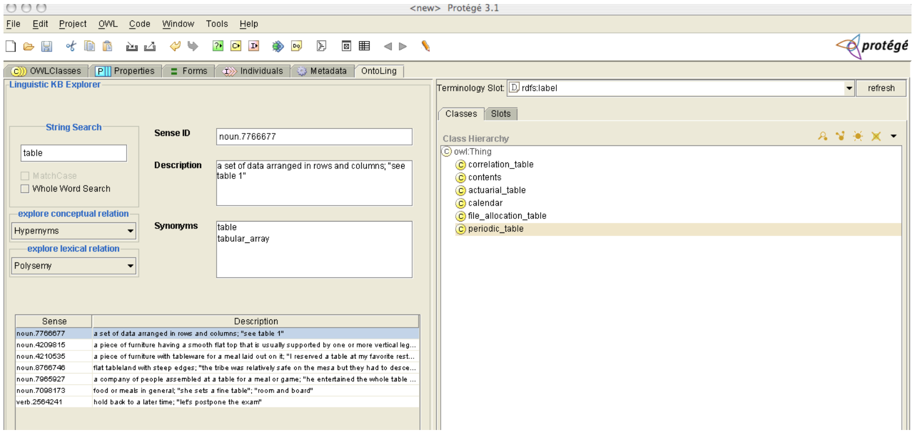
ONTOLING (see Fig. 5) allows the user to populate the categories of a given ontology with any WORDNET-like computational lexicon. Recall that the basic structure of ‘a wordnet’ is the taxonomy which is then enriched with other semantic relations: synsets are mainly organized via hyponymy (equivalent to *is-a* relationship for ontologies), potentially providing a huge amount of lexical sub-classes to ontological nodes. For exam-

[12] Detailed information about PROTÉGÉ can be found at *http://protege.stanford.edu*

[13] For more information see *http://ai-nlp.info.uniroma2.it/software/OntoLing/*

[14] PAL stands for PROTÉGÉ Axiom Language.

ple, suppose one wants to attach the *children* of the `substance` synset in WORDNET to the concept `amount of matter` in DOLCE. By means of ONTOLING, one simply "moves" single synsets or even branches –that is, a *node* with its *children*– in a given ontology and this move includes sense identifiers, glosses and all other information available from the computational lexicon[15]. A user can also change the "name" of a certain concept in the ontology according to a suitable term in the lexicon, this is done by selecting the appropriate lexical entry from a dedicated window. The ONTOLING interface is user-friendly and can also be used for simple navigation in the lexical resource, easing the access with a minimal but effective combination of widgets. We think that this plug-in is a necessary tool for the developer who wants to create semi-automatically bridges between ontologies and lexicons in a intuitive way.



String Search: table

□ MatchCase
□ Whole Word Search

explore conceptual relation: Hypernyms

explore lexical relation: Polysemy

Sense ID: noun.7766677

Description: a set of data arranged in rows and columns; "see table 1"

Synonyms: table
tabular_array

Sense	Description
noun.7766677	a set of data arranged in rows and columns; "see table 1"
noun.4209815	a piece of furniture having a smooth flat top that is usually supported by one or more vertical leg...
noun.4210535	a piece of furniture with tableware for a meal laid out on it; "I reserved a table at my favorite rest...
noun.8766746	flat tableland with steep edges; "the tribe was relatively safe on the mesa but they had to desce...
noun.7965927	a company of people assembled at a table for a meal or game; "he entertained the whole table ...
noun.7098173	food or meals in general; "she sets a fine table"; "room and board"
verb.2564241	hold back to a later time; "let's postpone the exam"

Terminology Slot: rdfs:label

owl:Thing
correlation_table
contents
actuarial_table
calendar
file_allocation_table
periodic_table

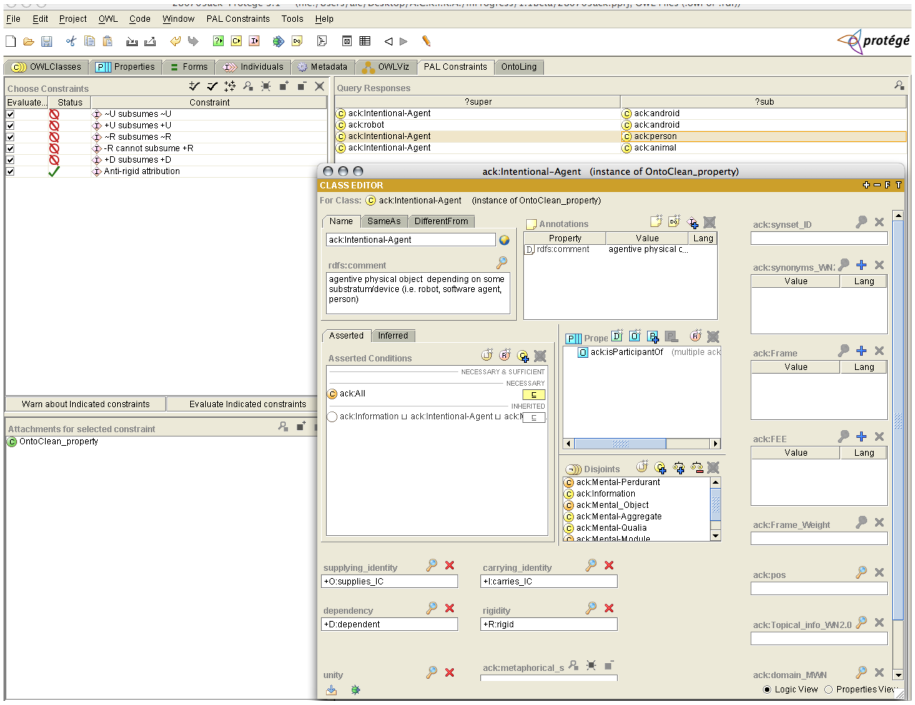
Fig. 5. PROTÉGÉ and ONTOLING screenshot

6.2 ONTOCLEAN in PAL

An essential phase of interfacing ontologies and lexical resources consists in the evaluation of the latter according to suitable ontological principles (we called it *restructuring* in the previous sections). Our experience in ONTOWORDNET showed that ONTOCLEAN provides such a principled methodology, helping the modeler to understand problems and guiding her in finding suitable conceptual solutions. Nevertheless, one disadvantage of restructuring a lexicon with ONTOCLEAN is its cost. Labelling lexical concepts with meta-properties and checking for formal violations through the *is-a* arcs is dramatically time-consuming, especially from a practical perspective. For example, in building ONTOWORDNET we couldn't check all 100000 synsets. Therefore, we needed to assume that the restructuring of higher levels of the WORDNET taxonomy

[15] In this sense, lexical items actually become OWL classes.

would effect on the lower ones too. A possible way to overcome this kind of problems could be to automatize ONTOCLEAN-labelling or/and ONTOCLEAN-checking. However, since performing ONTOCLEAN-labelling in an automatic way is extremely difficult (if possible) and opens thorny artificial intellingence issues[16], the efforts have been concentrated on the ONTOCLEAN-checking. By importing ONTOCLEAN formal rules into PAL[17] the knowledge engineer can exploit the PROTÉGÉ internal language and reasoner to check for ONTOCLEAN violations. The "PAL Constraints" widget looks like a splitted-window: the left side shows ONTOCLEAN rules paraphrased in natural language (i.e. ∼**U** cannot subsume +**U**); the right side elicits the formal violations visualizing every couple *father-son* which exemplifies a violation-type. For example, in Fig.6, ONTOCLEAN detected a problem in the dependence relation between `Intentional-Agent` and `android` [18].



Evaluate...	Status	Constraint
☑	⊘	~U subsumes ~U
☑	⊘	+U subsumes +U
☑	⊘	~R subsumes ~R
☑	⊘	-R cannot subsume +R
☑	⊘	+D subsumes +D
☑	✓	Anti-rigid attribution

?super	?sub
ack:Intentional-Agent	ack:android
ack:robot	ack:android
ack:Intentional-Agent	ack:person
ack:Intentional-Agent	ack:animal

ack:Intentional–Agent (instance of OntoClean_property)

For Class: ack:Intentional-Agent (instance of OntoClean_property)

Name: ack:Intentional-Agent

rdfs:comment: agentive physical object depending on some substratum/device (i.e. robot, software agent, person)

Property	Value	Lang
rdfs:comment	agentive physical c...	

ack:All ⊑

ack:Information ⊔ ack:Intentional-Agent ⊔ ack:M ⊑

ack:isParticipantOf (multiple ack

Disjoints: ack:Mental-Perdurant, ack:Information, ack:Mental_Object, ack:Mental-Aggregate, ack:Mental-Qualia, ack:Mental-Module

supplying_identity: +O:supplies_IC

carrying_identity: +I:carries_IC

dependency: +D:dependent

rigidity: +R:rigid

Fig. 6. ONTOCLEAN in working with PROTÉGÉ PAL: a problem of dependence

[16] Such a goal would require a terribly huge amount of common-sense and natural language knowledge to be inserted in a machine.

[17] See http://protege.stanford.edu/ontologies/ontoClean/ontoCleanOntology.html for further details about the translation performed by Nancy Ide.

[18] In the knowledge base considered - an extension of DOLCE - `android` was wrongly modeled as independent and `Intentional-Agent` as dependent, causing an ONTOCLEAN violation.

This concludes our brief sketch of the characteristics of two important tools that turn PROTÉGÉ from an ontology-oriented application into an integrated platform for interfacing ontologies and lexical resources. By means of ONTOLING plug-in and ONTOCLEAN in PAL, we showed the basic features of an implementation of the methodological and experimental issues introduced in the previous sections of the paper. Future work will concern the improvement of such tools mainly regarding new functionalities to perform better and deeper interfacing. As of now, this seems to be the only platform available with a reasonable set of implemented features.

7 Conclusion

We proposed a way of classifying the work done in interfacing ontologies and lexical resources. It consists in a clear separation between the restructuration of a lexical resource on the ground of an existing ontology hosting ontological principles, and the process of populating an ontology with lexical resources terms. A third option, called *alignment*, is a combination of these two aspects for the benefit of both the lexical resource and the ontology. We have shown how actual on-going work fits this classification through some examples. In the light of these clarifications, we discussed the issue of *constraint density* for lexical resources and related it to the light-weight/heavy-weight distinction established in kwnowledge representation. In this paper, we showed that different construction methodologies leads to different features in the resulting resources. We emphasized the need for selecting top-level ontologies and lexical resources according to their reliability. Finally, we overviewed a PROTÉGÉ-based platform for actual interfacing between ontologies and computational lexicons.

Future work concerns in particular the practical evaluation of the resources developed with the different methods that have been presented. This evaluation has to be done task by task in order to understand better which task requires which features. Such an evaluation constitutes a crucial step for the integration of ontological enhancement for lexical resources.

References

1. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A.: Wonderweb deliverabled18, ontology library (final). Technical report, LOA-ISTC, CNR (2003)
2. Niles, I., Pease, A.: Towards a standard upper ontology. In Welty, C., Smith, B., eds.: Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Ogunquit, Maine (2001)
3. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press (1998)
4. Guarino, N., Welty, C.: An overview of ontoclean. In Staab, S., Studer, R., eds.: Handbook of ontologies. Springer Verlag (2004) 151–159
5. Oltramari, A., Gangemi, A., Guarino, N., Masolo, C.: Restructuring wordnet's top-level: The ontoclean approach. In Simov, K., ed.: Workshop Proceedings of OntoLex'2, Ontologies and Lexical Knowledge Bases, LREC2002, Las Palmas, Spain (2002) 17–26
6. Taboada, M., Martinez, D., Mira, J.: Experiences in reusing knowledge resources using Protégé and PROMPT. International Journal of Human-Computer Studies **62** (2005) 597–618

7. Gangemi, A., Navigli, R., Velardi, P.: The ontowordnet project: extension and axiomatisation of conceptual relations in wordnet. In: International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE), Catania, (Italy) (2003)
8. Guarino, N.: Formal ontology in information systems. In Press, I., ed.: Proceedings of FOIS'98, Trento, Italy (1998) 3–15
9. Gómez-Pérez, A., Fernández-López, M., Corcho, O.: Ontological Engineering. Springer (2004)
10. Vossen, P., ed.: EuroWordNet: a multilingual database with lexical semantic networks. Kluwer Academic Publishers (1998)
11. Fillmore, C.: Frame semantics and the nature of language. In: Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech. (1976)
12. Sowa, J.: Knowledge Representation. Brooks/Cole, Pacific Grove, CA (1995)
13. Borgo, S., Guarino, N., Masolo, C.: A pointless theory of space based on strong connection and congruence. In Kaufmann, M., ed.: Proceedings of Knowledge Representation and Reasoning (KR96). (1996)
14. Allen, J.F.: Towards a general theory of action and time. Artificial intelligence **23** (1984) 123–154
15. Smith, B.: Mereotopology: A theory of parts and boundaries. Data and Knowledge Engineering **20** (1996) 287–303
16. Guarino, N.: Some ontological principles for designing upper level lexical resources. In Rubio, A., Gallardo, N., Castro, R., Tejada, A., eds.: Proceedings of First International Conference on Language Resources and Evaluation, ELRA (1998) 527–534
17. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A.: Sweetening WordNet with DOLCE. AI Magazine **3**(24) (2003) 13–24
18. Guarino, N., Welty, C.: A formal ontology of properties. In: in 12th International Conference on Knowledge Engineering and Knowledge Management: Methods, Models and Tools. (2000)
19. Gangemi, A., Guarino, N., Oltramari, A.: Conceptual analysis of lexical taxonomies: The case of wordnet top-level. In Welty, C., Smith, B., eds.: 2nd International Conference on Formal Ontology in Information Systems. (2001) 285–296
20. Reed, S., Lenat, D.: Mapping ontologies into cyc. In: AAAI 2002 Conference Workshop on Ontologies For The Semantic Web, Edmonton, Canada (2002)
21. Niles, I., Pease, A.: Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In: Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03), Las Vegas, Nevada (2003)
22. Harabagiu, S.M., Moldovan, D.I.: Knowledge processing on an extended wordnet. In Fellbaum, C., ed.: WordNet, An electronic lexical Database. The MIT Press (1998) 379–406
23. Harabagiu, S.M., Miller, G.A., Moldovan, D.I.: Wordnet 2 - a morphologically and semantically enhanced resource. In: SIGLEX 1999. (1999)