Correction and Extension of WordNet 1.7 for Knowledge-based Applications

Philippe A. Martin
DSTC, The University of Queensland
Level 7, GP South, Staff House Rd
Qld. 4072, Australia
philippe.martin@gu.edu.au

ABSTRACT

This article presents the transformation of the noun-related part of WordNet (108,000 nouns, 74,500 categories representing their meanings, and 95,000 semantic links between them) into a genuine "lexical ontology", usable for knowledge representation, sharing and retrieval on the Web. To do so, (i) I generated intuitive identifiers for all the categories, (ii) introduced 353 lexical corrections (mainly by adding category names), (iii) extracted 6211 individuals (0th-order categories) to differentiate the WordNet specialization links into subtype and instance links, (iv) added a top-level ontology of about 150 concept types and 150 primitive relation types to permit normalized knowledge statements and some semantic checks on them and the ontology, (v) removed or modified 306 links between categories to repair inconsistencies or avoid redundancies, and (vi) added 159 links between categories, 50 prototypical schemas of relations from common categories, and about 1300 semantic annotations. These corrections are documented at http://www.webkb.org/doc/wn/ and the ontology is downloadable in DAML/RDF or more concise and expressive formats. Web users can also search and extend the ontology via the WebKB-2 knowledge server at http://www.webkb.org. Hence, the above numbers may increase as the ontology improves. This article illustrates these modifications and their rationales.

Categories and Subject Descriptors

A.10 [**Semantic Web**]: Ontologies, Semantic web applications

Keywords

Ontology, WordNet, Semantic Web, Knowledge Representation

1. INTRODUCTION

WordNet [1] is a lexical database that connects English words to "synonym sets" (each "synset" represents one of the meanings of the words in the set) and organizes the synsets by semantic links, e.g. specialization and partOf links. WordNet is increasingly interpreted and exploited as a lexical ontology (i.e. a set of categories connected by links having a formal semantics) despite its shortcomings for this purpose.

A natural language ontology derived from WordNet and other sources could support, ease or enhance various kinds of applications, e.g. precision-oriented information retrieval [14], query expansion and answering [5] [6], machine translation [10], and knowledge representation, sharing or brokering [3] [4]. In [13], I argued that the Semantic Web (as I understand it) cannot be achieved without the existence of at least one natural language ontology that can be extended by people and permit to give categories from different ontologies some shared formal meaning. [13] also details how the knowledge server WebKB-2 exploits WordNet 1.7 for guiding and checking knowledge representation, and for permitting Web users to share or retrieve knowledge, and further extend or correct the shared ontology if necessary. (Protocols and naming conventions prevent lexical and semantic conflicts).

This article introduces extensions and corrections of the noun-related part of WordNet 1.7 to transform it into a lexical ontology usable for knowledge-based applications, and especially the *manual* representation of natural language sentences. (Much more would be needed to support natural language parsing) [1]. No claim is made that this ontology is sufficient to support the inter-operation of fully automatic software agents, e.g. for e-commerce or database integration purposes. [8] shows that such inter-operations have strong requirements and, in the general case, are not likely to be fully supported by ontologies anytime soon. Conversely, this work may be re-used to enhance Web applications that are not always seen as knowledge-based, e.g. metadata registries and Yellow-pages like catalogs.

This article first shows how short and intuitive identifiers were generated for each WordNet category, and illustrate some of the lexical corrections. Second, it explains how types (1st-order categories) were distinguished from individuals (0th-order categories), and hence differentiated WordNet specialization links into subtype and instance links. Third, it introduces the top-level ontology of concept and relation types into which the top-level categories of WordNet were inserted in order to support the construction of normalized (i.e. better *retrievable*) knowledge statements and certain *semantic checks* on the ontology and the statements. Fourth, it illustrates the kinds of problems that led to the removal or modification of links in WordNet. Fifth, it details the kinds of additions (links, schemas, annotations) made to some WordNet categories.

WWW2003, May 20–24, 2003, Budapest, Hungary.
ACM xxx.

[1] Only the noun-related part of WordNet 1.7 is used because the use of categories representing the meanings of verbs, adverbs or adjectives has several drawbacks in knowledge representation for various inter-related reasons such as (i) using such categories with quantifiers has no real meaning (e.g. "any transformation" or "3 transformation" has a meaning but "any transform" and "3 transform"has not), (ii) organizing these categories by generalization links is difficult or impossible, (iii) these categories are shortcuts for more explicit constructions using categories for nouns (these shortcuts are rarely explicitely defined, and few ontology engines are currently able to exploit such definitions, hence using these categories reduce knowledge matching, checking and inferencing possibilities). More details and rationales can be found in [13].

2. CATEGORY IDENTIFIER GENERATION AND LEXICAL CORRECTIONS

A category may have many names (the elements of the "synset" in WordNet) that may be shared by other categories, but should have at least one "identifier" to refer to it uniquely. In WebKB-2, a category identifier is allowed to be a URL or an e-mail address, but for readability reasons, is most often composed of a short identifier for the user (or source) that created the category, and a *key name* distinguishing the category from other ones created by the same user. For example, `wn#car` refers to a WordNet category for the noun "car", while `pm#car` may represent a different notion for the user `pm`. WebKB-2 allows the prefix "wn" to be dropped, and the category creator may specify other names by appending them to the identifier; thus, `#car__auto__automobile` refers to the same category as `wn#car`. (Such categories may also be referred and accessed from outside WebKB-2 via a URL. The reader is encouraged to access http://www.webkb.org and browse from the categories referred to in this paper).

WordNet has at least two internal identifiers for each category, e.g. the category for "Friday" has for identifiers `12558316` and `friday%1:28:00::`. While some applications re-use them, others (such as [2]) generated their own identifiers by concatenating names or using suffixes, e.g. `Inessential$Nonessential` and `Cell_1`. However, for knowledge representation, exchange and sharing purposes (in knowledge bases as well as on the Web) category identifiers should be concise and clear to permit readable text-based knowledge statements (graphical interfaces should not be required and are not necessarily the best device to enter, display, debug or maintain a large amount of knowledge [13]; category identifiers should also be usable within *controlled languages* [7]). Hence, for these purposes, each category should have at least one identifier composed of a common and unambiguous word or expression for its meaning, *and as little else as possible*. This means that one of the category names should be used as key name, if possible with no suffix. Its capitalization should not be modified, in order to ease its use in controlled languages and avoid imposing on the user to add another name for specifying the exact capitalization. (This is one of the problems of the RDF naming conventions; the use of the intercap style also leads to many lexical inconsistencies, as I experienced when I integrated the TAP knowledge base [17] which follows RDF conventions).

In WordNet, the most common name for a category is supposed to be the first in the synset. This information is not very reliable and less ambiguous names may appear after the first. When one of the other names is a compound name beginning or ending with the first name (as "Steve_Martin" begins with "Steve" and ends with "Martin"), it constitutes a better choice for a key name than the first name ("Martin"). Hence, here are the first rules (ordered by decreasing order of priority) that I chose to generate key names:
1) when the 1st name of a category begins or ends with one of the other names, select this other name as key name, unless it is shared by another category that has no generated key name yet;
2) select the 1st name of a category as key, unless it is shared by another category that has no generated key name yet;
3) try the first two rules on the 2nd name instead of the 1st;
4) try the first two rules on the 3rd name instead of the 1st; etc.

To respect the decreasing order of priority on these rules, I scanned the knowledge base (KB) many times (each time, testing all remaining categories without a key name), allowing the test of a lower priority rule only when the application of rules of greater priority did not lead to any more change. This simple approach was efficient enough given WebKB-2 could scan the whole KB quite quickly (0.45 second in average). The application of the first two rules (i.e. trying to use only the 1st name of each category) permitted the assignment of key names to 75% of categories (56,074 out of 74,488). The use of the other rules, i.e. of the other names, permitted the assignment of key names to 84% of categories. This means that in the remaining 16%, each category had *all* its names in common with another category.

Hence, to go further, suffixes had to be generated. When I integrated WordNet 1.6, I used numbers, but experience in using such categories in knowledge statements led us to realize that this option was not user-friendly enough and that a much clearer option was to use the key name of the first supertype. Such a suffix often help people guess the meaning of a category without having to access its supertypes. However, I did not want to give a key name with a suffix to *all* remaining unassigned categories. Hence, I added the following rules (by decreasing order of priority and with a lower priority than the previous rules) to select the categories to which key names with a suffix would be assigned: (i) select the category with a frequency-of-use number far lower than the other categories sharing all the same names (this number is provided by WordNet and represents the frequency of appearence of the category in a few concordance documents; it is an indication but not of paramount importance; "far lower" was first set to 30 and then to decreasing values); (ii) select the category with a far lower number of subtypes than the other categories sharing all the same names. More precisely, in these last two rules, I used combinations of gradually decreasing values of frequency-of-use and number of subtypes; I also penalized the assignment of suffixes to subtypes of `#action`, as these types are more frequently used than others in knowledge statements).

After several more scans of the KB with all the rules, there were still a few dozens of unassigned categories. To fix this, I added more precise names to these categories or re-ordered their names. I also manually corrected some suffix attributions and key name choices. For example, "Republic_of_Singapore", instead of "Singapore", was selected as key name in application of the 1st rule but `#Singapore` is a more convenient identifier, and it seems that the island of Singapore and the capital of Singapore are better referred to via `#Singapore.island` and `#Singapore.capital` than via `#Singapore`. To fix such problems, before re-running the key name assignment procedure from scratch, I semi-automatically pre-attached suffixes to many key names, especially for the specializations of the category `#location`. For example, the suffixes ".capital", ".city", ".island", ".country" and ".colony" made many key names unambiguous. Sometimes, instead of using the generalizing category for the suffix, I followed the `partOf` link. For example, `#town` has three instances with only name "Bangor" but which are part of different regions. Thus, I identified them `#Bangor.Wales`, `#Bangor.Northern_Ireland` and `#Bangor.Maine`.

Thus, *only* 5944 WordNet categories have been given a key name with a suffix. The list of these categories is accessible from [18], i.e. from http://www.webkb.org/doc/wn/. So is the list of the 353 lexical corrections: 28 modifications of category annotations, 248 category names added, and 77 manual re-orderings of category names. Here is an example showing how the corrections were documented:

```
#wn07834480|German_citizen__German
 (^ $("German_citizen" has been added as key name;
       the original annotation was:
       "a native or inhabitant of Germany")$
    a person of German nationality^)
```

This format is used for saving the KB in a backup file. `07834480` is the WordNet identifier, `German_citizen` the added key name (since "German" also refers to a language), `German` the original name, `(^...^)` the category annotation, and `$(...)$` a sub-annotation which, as a default, WebKB-2 does not show to end-users.

To conclude this section, let us highlight the fact that this work provides Web users a shared formal vocabulary to mark up their documents or the meanings of words in their documents, or to use in their knowledge statements. If a word meaning is missing, a user may easily add it to the ontology via WebKB-2, thus permitting other people to retrieve and re-use his/her categories or statements. In [13], I show how this approach (along with a mirroring strategy between knowledge servers) avoids the merging and inferencing problems inherent to the completely distributed approach often advocated for the Semantic Web [19] [9] while still retaining its advantages. The generated category identifiers may be used in RDF as well as in controlled languages, and do not suffer from the loss of information (and its associated problems) caused by the use of the intercap style. Finally, the method of generating unambiguous and readable identifiers for WordNet could be re-used on some other linguistic ontologies.

3. EXPLICITATION OF INDIVIDUALS

Distinguishing 1st-order types from their instances, often called "individuals", is important for knowledge representation, inferencing and checking. Individuals cannot have specializations, i.e. subtypes or instances. Certain individuals, often called continuants or endurants [2], can change in time without being viewed as different individuals (i.e. without loosing their identity), e.g. individuals for persons or cities. Specializing such individuals according to time might be tempting, e.g. `pm#ParisIn1995`, but better avoided: statements (facts or definitions) about individuals should represent dates and durations in an explicit way using contexts.

Distinguishing types from individuals is not always obvious. For example, [2] asserts that the WordNet category `#karate` should be an individual, but there are various kinds of karate, and furthermore, since `#karate` is a subtype of `#activity` [2], each individual practice of karate may be considered as an instance. Anything which may be specialized, or has various occurrences, or comes in different variants or versions should be represented as a type, rather than an individual; otherwise, knowledge representation possibilities and accuracy are reduced. For example, any doctrine, book, language, alphabetic character, code, diploma, sport or recurring situation should rather be represented as a type. The first character of the alphabet has many variants (e.g. its uppercase and lowercase variants) and billions of instances (occurrences) in books. An alternative view would be to consider that in certain cases a variant is not a subtype and an occurrence is not an instance, and then use different links or relations to represent this information. However, in this alternative model, information would be more complex to describe, and inferencing more complex to implement.

I chose the simplest model. However, since people often wish to use certain types without quantifiers, as if they were individuals (e.g. in English, the nouns "Monday" and "Polish" are rarely used with an article, i.e. a quantifier), WebKB-2 allows it in Frame-CG (FCG) and Formalized-English [12] (both extend and simplify the Conceptual Graph Linear Form (CGLF)) on the condition that the category has no subtype, no instance and is not a subtype of `pm#physical_entity` nor `#time_period`.

[18] lists the 6211 individuals that I manually isolated: typically, time periods, persons, organizations, places and battles. To do so, I first translated all WordNet specialization links as subtype links. Then, since WordNet categories are grouped by theme within the WordNet database files, I operated a careful but relatively quick "search and replace" of subtype links into instance links in the zones where individuals could appear. I double checked the work on categories having a name with a capitalized first letter. Here is an example in the FO notation (which I derived from the linear notation of CGs [15] and that is parsed by WebKB-2):

```
#Neolithic_Age ^ #time_period,  P #Stone_Age;
```

The character ’ˆ’ represents the instanceOf link, while ’P’ represents the partOf link.

To sum up, important formal information was added to WordNet categories (consistent with their original meanings) while adopting an approach that maximise re-use possibilities. For knowledge sharing and inferencing purposes, I also argue against the use of instance links between types (i.e. against the introduction of second-order types and second-order statements) *when* subtype links can be used instead. Indeed, subtype links are easier to use for structuring categories, and then to exploit. The logical interpretation of statements using types of different orders may also be difficult and they are not commonly exploited by inference engines. Over-uses of the instance link are frequent. For example, the TAP KB [17] categorizes certain types of magazines or books as instances of a second-order type `tap#product_type` which has no other supertype than `rdfs#class`. Even if it had, the use of a first-order type such as `#product` permits much more comparison with (or connection or inheritance of constraints from) other types, hence more retrieval and checking possibilities. Some 2nd-order types such as `daml#transitive_property` in DAML [20] are justified (transitivity is a class property: the subtypes of the class do not necessarily inherit this property). However, when possible, subtyping a 1st-order type such as `pm#transitive_relation` is preferable. The use of instance links is one of the basic issues of knowledge representation for which, recommendations should be issued by Semantic Web related organisms [19] to ease knowledge re-use.

4. TOP-LEVEL ONTOLOGY

WordNet has not been built for knowledge representation purposes, nor apparently according to basic taxonomy building principles and with consistency checking tools. As noted in [2], types and individuals are not distinguished, the annotation of a category is not to be relied on as it may be contradicted by specializations of this category, direct specializations of categories often have heterogeneous levels of generality, role types (e.g. `#student`) are not distinguished from natural types (e.g. `#person`) and may generalize them. I also found that (i) specialization links are sometimes used where "location" or "similar" links should be used, (ii) the "part" and "member" links between types are not used in a consistent way (most seem to mean that all instances of the source type have for part/member at least one instance of the destination type, but this is not *always* the case), (iii) some of these transitive links are redundant (and there were even a few directed cycles in previous versions of WordNet), and (iv) exclusion links (the rare constraints that WordNet provides to check its taxonomy) are sometimes broken, i.e. some exclusive categories have common specializations. Table 1 shows the top WordNet categories for nouns and their direct subtypes, using the FO notation. The lack of structure is clear.

This work seems the first to have isolated individuals, generated intuitive category identifiers, corrected and documented a large number of problems, and permitted Web users to further extend and correct this ontology (the links and names added by Web users must not introduce detectable inconsistencies, and for search or presentation purposes, can be selectively filtered since the creators of links and names are also stored). I have not attempted to bring more structure to the whole of WordNet, as this would probably take many years of work. However, like others, this work inserts the top-level categories of WordNet into a better structured top-level ontology. In 1994, Sensus [10] was created by manually merging the

Table 1: The WordNet 1.7 top-level types for nouns ***('>' introduces subtypes, brackets enclose exclusive subtypes)***

#human_action__act__human_activity
> #action #nonaccomplishment #leaning #assumption #rejection #forfeit {#activity #inactivity} #wearing #judgment #production.human_action #judgment #stay #residency #laughter #hindrance #stoppage #group_action #distribution #permissive_waste #communicating #speech_act;

#state > #skillfulness #cognitive_state #cleavage.state #medium.state #condition #condition.state #conditionality #state_of_affairs #relationship #relationship.state #tribalism.state {#utopia #dystopia} #wild #isomerism #degree.state #office.state #status {#beingness #nonbeing} #death.state {#employ #unemployment} {#order.state #disorder} #enmity #conflict.state #illumination #freedom #representation.state #dependance {#motion #motionlessness} #non-issue {#action.state #inaction.state} #temporary_state #imminence #preparedness #kalemia #union.state {#matureness #immaturity} #state_of_grace #eternal_damnation #omniscience #omnipotence {#flawlessness #imperfection} #unity #receivership.state #ownership.state #end.state #sale.state #turgor #polyvalence;

#event > #might-have-been #nonevent #happening #social_event #miracle.event #Fall;

#phenomenon > #natural_phenomenon #levitation #metempsychosis #outcome #luck.phenomenon #luck #process;

#entity
> #self-contained_entity #whole_thing #living_thing #cell #causal_agent #holy_of_holies #physical_object #location #depicted_object #unnamed_thing #imaginary_place #anticipation #body_of_water #natural_enclosure #expanse {#inessential #essential} #physical_part #sky #building_block #variable;

#group__grouping > #arrangement #straggle #kingdom.group #biological_group #biotic_community #human_race #people #social_group #aggregation #edition.group #electron_shell #ethnic_group #race.group #association.group #subgroup #sainthood #citizenry #population.group #masses #circuit #system #series.group;

#possession > #belongings #territorial_dominion #white_elephant.possession #transferred_property #circumstances #assets #treasure.possession #liabilities;

#psychological_feature > #cognition #motivation #feeling;

#abstraction > #time #space #attribute #relation #measure #set;

WordNet top-level into Ontos and the Generalized Upper Model, and then semi-automatically merging WordNet with the Longmann Dictionary of Contemporary English. Sensus was created for machine translation purposes. At about the same period, for knowledge acquisition and representation purposes, I extended Sowa's first top-level ontology [15] and used it for structuring WordNet 1.5 top-level [11]. In 2001, for the Semantic Web and other knowledge sharing purposes, the OntoClean ontology and methodology was used to re-structure WordNet 1.6 top-level [2]. In October 2002, I integrated the last version of the OntoClean ontology, DOLCE (D17) [21], into WebKB-2 ontology but found most of the 40 DOLCE top categories *too specific* to specialize them with WordNet categories. The next section presents two examples.

4.1 Minimizing Re-categorization

Example 1. OntoClean/DOLCE distinguishes "qualities" (like size, color, redness, smell and duration) from "quales" (quality regions/spaces, i.e. categories of values for qualities, e.g., according to [2], #red, #past_times and #Greenwich_Mean_Time). They are categorized under the exclusive categories dolce#quality and dolce#region__quale. However, in WordNet, such categories (about 8900) are inter-related by specialization links, e.g. #red specializes #chromatic_color and #color, while #past_times specializes #time. Specializing the types dolce#quality and dolce#region by WordNet categories, as suggested in [2], is problematic: (i) this classification has to be done for most of the 8900 categories, *not just for their most general categories*; (ii) a great number of WordNet specialization links have to be *broken*, hence this structure is *lost* and the meaning of a great number of WordNet categories is *modified*; (iii) it is often difficult to decide whether a WordNet category should be *interpreted* as a quality or as a quale; as opposed to [2], I consider #Greenwich_Mean_Time, #work_time and #red as quality types (the authors of [2] argue for the representation of red and other adjectives for colors as quales, but #red_redness represents the meaning of the nouns "red" and "redness"). In the integration of WordNet, I have added or refined but *not removed or modified* links – except for 306 (out of 74,488) in order to fix inconsistencies. From an Ontoclean perspective, this is possible by interpreting most of the above cited 8900 categories as qualities. However, I have not explicitely categorized their upper types as specializations of dolce#quality in order to permit WebKB-2 users to classify certain WordNet categories as subtypes of dolce#region when this does not introduce inconsistencies. I have generalized these upper types, plus dolce#quality and dolce#region, by pm#attribute_or_measure (this type name is due to the fact that the things I call "measures" may specialize the things that are usually called "attributes"). Here is a statement in FCG (a graph-based notation parsed by WebKB-2) showing how knowledge representation can be done in an intuitive and normalizing way with the interpretation of WordNet attributes or measures as qualities: [a #car, #color: a #red, #weight:900 #kg]. In Formalized-English, a notation equivalent to FCG, this statement can be written: there exists a #car that has for #color some #red and for #weight 900 #kg. Both #red and #kg are quantified (I give KIF definitions for the FCG numerical quantifiers in [12]). As in Ontoseek [3] (a WordNet-based knowledge retrieval system built by the team that designed OntoClean), the types #color and #weight are used as if they were relation types. WebKB-2 checks that these categories are subtypes of the type pm#thing_that_can_be_seen_as_a_relation and that they respectively generalize #red and #kg. No quale is explicitely referred to in this statement. If #red and #kg were categorized as quales, more complex statements would have to be written, e.g.:
[a #car, #color: (a #color,
  pm#measure: a #red), #weight: (a #weight,
  pm#measure: 900 #kg)]. Checking this graph would also

be more complex and would require additional information on categories acceptable as measures for colors and weight.

Example 2. In [2], dolce#amount_of_matter is exclusive with dolce#physical_object and given #substance as subtype. However, #substance has many subtypes which are also subtypes of dolce#physical_object. An example is the type #olive.relish which specializes #fruit (#physical_object) and #relish (#condiment, #substance). Another example is #glass_wool, subtype of #artifact (#physical_object) and #insulator (#substance). Since these WordNet links do not appear as clear mistakes, it seems that in [2], #substance has been over-interpreted (or adapted) to fit the meaning of the type dolce#amount_of_matter. Instead, I categorize #substance (along with types like #building_block and #physical_part) as subtype of pm#physical_part_or_substance which, like dolce#physical_object and dolce#amount_of_matter, is a direct subtype of pm#physical_entity. Since this last type covers both substances and physical objects with unity, it may be seen as an adequate candidate for classifying a "statue of clay". It may also be used for signatures of relations, e.g. relations representing physical attributes such as color or mass (although as hinted in Example 1, this is discouraged in WebKB-2).

4.2 Summary of the Approach and its Results

The distinctions made by DOLCE and other top-level ontologies are important and their integration may be used by knowledge servers to guide the users to represent knowledge in more precise and re-usable ways. The precision of DOLCE categories and their associated constraints are also intended to ease the automatic matching of categories from (Semantic Web) ontologies independently developed but re-using the DOLCE ontology. Although this precision makes the current set of DOLCE categories difficult to use for structuring WordNet top-level (and other distinctions are also required), it is valuable (including the distinction between qualities and quales, although I are more interested in the more general distinction between concept types that can be used for relations and those that can be used as destinations of those relations; Section 6 will show how I have prefered to make the distinction).

However, not all distinctions are equally useful. For example, I also integrated Sowa's recent top-level ontology [16], one of the basis of which is C.S. Peirce's distinction between (i) things that can be seen as "independent of any relationships to other entities" (e.g. a person), (ii) things that can be seen as being "in a relationship to some other entities" (e.g. a spouse), (iii) things that "create a relationship to some other entities" (e.g. a mariage). The problem is that no constraint or particular relation can be associated to those categories and they cannot be used to classify natural language categories since almost anything can be seen as being in relation with, or creating relations between, other entities.

Table 2 presents a summary of the top-level concept types in WebKB-2. Many types from WordNet, DOLCE and Sowa are shown. The catch-all categories #entity and #abstraction do not appear but their direct subtypes have been categorized in various places. Most of the upper types, e.g. pm#spatial_entity or pm#description, are relatively intuitive types required for the signatures of the relation types (Table 3). These concept types have been given constraints (mainly exclusion links) and prototypes (typical relations) that are inherited by all their specializations.

The relation types proposed by WebKB-2 are mainly for primitive binary relations and intended to support an explicit and normalized way of representing natural language sentences (in [13], I give rationales against the use of non-binary relations and complex relations, e.g. relations representing processes). I also integrated argumentation relations and the relations of DAML, RDF, RDFS, Dublin Core and the core of KIF. Table 3 shows the overall organization, although it also deepens in the case relations. The grouping by source category proved to be the cleanest and most intuitive structure, and WebKB-2 exploits it when generating menus to guide knowledge representation.

The "Suggested Upper Merged Ontology" (SUMO) [22] has similarities with WebKB-2 ontology in the sense that it has mappings with categories of WordNet 1.6, and includes some spatial and case relations, and various top concept types from various top-level ontologies, e.g. Sowa's last top-level ontology. Its integration into WebKB-2 ontology has begun.

On the other hand, I do not plan to integrate the HPKB top-level ontology [23] which has been created in 1998 by merging the top-level ontologies of CYC and Sensus. It seems preferable and easier to integrate some elements of the last release of CYC top-level [24] (only some elements since on the one hand, there is already some overlap, and on the other hand different approaches have been adopted in CYC, e.g. it includes many non-binary relations and relations representing processes).

To conclude, WebKB-2 and its ontology may help people avoid the difficult task of finding, integrating and extending adequate ontologies, especially top-level ontologies (a task that some Semantic Web researchers, seem to think the knowledge providers to the future Semantic Web are able to do, have the time to do and will *have to* do! [9]). Instead, the WebKB-2 user is simply supposed to find adequate categories by typing words and browsing from the proposed categories for these words, and then fill cascading menus adapted to the categories s/he selected or entered (this point is detailed in Section 6). Knowledge precision and normalization is encouraged by the various proposed distinctions, the adopted approach (e.g. the proposed basic binary relations) and the proposed notations (e.g. their extended quantifiers).

5. SEMANTIC CORRECTIONS

Up to March 2003, 117 links have been removed, and the types or destinations of 198 links have been modified. Of these 315 links, 41 were redundant and about 230 were inconsistent with other links. Most of the inconsistencies were automatically detected thanks to the exclusion links in WebKB-2 top-level ontology. For example, some categories in WordNet were classified as *both human action and* causal agent, instrument or result of action (e.g. #relaxant and #interpretation) or of communication medium/content (e.g. #epilog and #thanksgiving), or *as both communication medium/content and* physical entity (e.g. #book_jacket) or attribute (e.g. #academic_degree). Some specialization links in WordNet were also used instead of "member" links, (e.g. between categories for species and genus of species). Similarly, WordNet does not have "location", "similarTo" and "identity" links, and uses subtype links instead of location links (e.g. many city/regions where battles have occured were classified both as city/regions and battles), similarTo links (e.g. for a Greek god and its Roman counterpart) and identity links (WordNet sometimes introduces a different category to represent obsolete names).

Redundancy was detected by exploiting the transitivity of specialization, part and member links. Apart from exclusion and specialization links, Only the combination of exclusion and specialization links was exploited to detect inconsistencies or redundancies. More could be done. For example, the fact that "if t2 specializes t1, and t1 is member of t0, then t2 is member of t0" should be exploited to detect more redundant links, e.g. in WordNet both #dog and its subtype #hound_dog are member of #pack.animal_group

Table 2: Some of the 160 top-level concept types in WebKB-2

>: subtype link; =: identity link; /: complementOf link; {(...)}: close subtype partition; {...}: open subtype partition

```
pm#thing__something__top_concept_type (ˆany object is instance of this typeˆ)
 > {(pm#situation pm#entity)} {(pm#thing_playing_some_role sowa#independent_thing)}
   {(sowa#physical_thing sowa#abstract_thing)} {(sowa#continuant sowa#occurrent)},  = dolce#entity,  / daml#nothing;

   pm#situation (ˆsomething that ”occurs” in a real/imaginary region of time and spaceˆ)
    > {(pm#state pm#process)} {(dolce#stative dolce#event)}
      pm#phenomenon sowa#process sowa#situation #event pm#situation_playing_some_role,  = dolce#perdurant__occurence;

      pm#state (ˆsituation not changing and not making a change during a given period of timeˆ)
       > #state #feeling pm#state_playing_some_role;

      pm#process (ˆsituation that makes a change during some period of timeˆ)
       > pm#event pm#problem_solving_process #unconscious_process
         #cognitive_process #human_action pm#process_playing_a_role;

   pm#entity (ˆsomething that can be ”involved” in a situationˆ)
    > {(pm#spatial_object pm#nonspatial_object)} {(pm#undivisible_entity pm#divisible_entity)}
      dolce#endurant pm#entity_playing_some_role;

      pm#spatial_object (ˆspace region or thing occupying a space regionˆ) > pm#space dolce#physical_endurant sowa#object;

         pm#space (ˆpoint or extent in spaceˆ) > dolce#feature #space #location #natural_enclosure #expanse #sky #shape;

         dolce#physical_endurant > {(pm#physical_entity dolce#feature)};

            pm#physical_entity (ˆspatial entity made of matterˆ)
             > {dolce#physical_object dolce#amount_of_matter} pm#physical_part_or_substance;

               dolce#physical_object > {(dolce#agentive_physical_object dolce#non_agentive_physical_object)};

                  dolce#agentive_physical_object (ˆe.g. an animal, a cellˆ) > pm#living_entity #living_thing #cell;

                  dolce#non_agentive_physical_object (ˆe.g. a bottleˆ) > pm#dead_entity #physical_object;

               pm#physical_part_or_substance > #physical_part #building_block #substance;

      pm#nonspatial_object (ˆe.g. knowledge, motivation, language, measureˆ)
       > pm#psychological_entity {pm#description_content/medium/container pm#attribute_or_measure}
         pm#collection dolce#abstract;

         pm#psychological_entity (ˆfeature/product of mental activity, e.g. feelingˆ)
          > dolce#mental_object #psychological_feature;

         pm#description_content/medium/container > {pm#description pm#container_of_description};

            pm#description (ˆdescription (content/medium) of an entity or a situationˆ)
             > pm#description_content pm#description_medium sowa#form;

               pm#description_content__information (ˆe.g. a narration, an hypothesisˆ)
                > sowa#proposition sowa#intention dolce#fact kads#role rdf#description #code.laws
                  #subject_matter #written_material #public_knowledge #cognitive_factor
                  #perception.cognition #cognitive_content #history.cognition #mental_attitude;

               pm#description_medium (ˆe.g. a syntax, a language, a script, a structureˆ)
                > pm#abstract_data_type #structure #communication #language_unit #symbolic_representation;

            pm#container_of_description (ˆfile, image, ... but not a disk or a piece of paperˆ)
             > pm#document_element #representation_container;

         pm#collection (ˆsomething gathering separated things (entities/situations)ˆ)
          > #group #set dolce#set dolce#arbitrary_sum pm#structured_ADT sowa#structure pm#type;

            pm#type > rdfs#class rdf#property;

      pm#entity_playing_some_role (ˆe.g. an agent, an ownerˆ)
       > pm#owned_entity pm#entity_part #variable pm#situation_result pm#process_recipient
         pm#process_object pm#causal_entity pm#imaginary_entity {#essential #inessential}
         #self-contained_entity #anticipation #unnamed_thing #holy_of_holies;

   pm#thing_playing_some_role (ˆcategory to classify things according to roles/viewpoints; this is application-dependantˆ)
    > pm#created_thing pm#thing_needed_for_some_process pm#thing_that_can_be_seen_as_a_relation
      pm#situation_playing_some_role pm#entity_playing_some_role {sowa#mediating_thing sowa#relative_thing};

      pm#thing_that_can_be_seen_as_a_relation (ˆtype usable as relation typeˆ)
       > pm#attribute_or_measure pm#contact_point #relation #psychological_feature #information #national #maker
         #creator #employee #employer #seller #user #relative #peer;
```

Table 3: Some of the 150 primitive relation types in WebKB-2

>: subtype link;   ^: instanceOf link;   -: reverse link;   (...): signature;   ?: any type;   {...}: open subtype partition ;   //:comment

```
pm#relation__related_with (*) (^type for any relation (unary, binary, ..., *-ary) and instance of rdf#property^)
 >  {pm#relation_from_situation  pm#relation_from_spatial_object  pm#relation_from_description_content/medium/container
     pm#relation_from_type}  {dc#Type  dc#Description}  kif#subst
    pm#relation_from_collection {pm#relation_to_collection  pm#relation_to_time_measure}
    pm#attributive_relation  {pm#different  pm#ordering_relation}  pm#relation_for_an_application  dc#Relation,    ^  rdf#property;

    pm#relation_from_situation (pm#situation,*)   //'*': 0 or more of any type
     >  pm#relation_from_situation_to_time_measure  pm#relation_from_situation_to_situation  pm#case_relation  pm#within_group;

        pm#relation_from_situation_to_time_measure (pm#situation,pm#time_measure)
         >  pm#time  pm#duration  pm#from_time  pm#until_time  pm#before_time;

        pm#relation_from_situation_to_situation (pm#situation,pm#situation)  >  pm#later_situation;

            pm#later_situation (pm#situation,pm#situation)  >  pm#next_situation  pm#consequence;

        pm#case_relation__thematic_relation (pm#situation,*)
         >  pm#cause/object/result/place  pm#experiencer  pm#recipient  pm#relation_from_process_only;

            pm#cause/object/result/place (pm#situation,*)  >  pm#cause/object/result  pm#place  pm#from/to_place;

                pm#cause/object/result (pm#situation,*)  >  pm#agent  pm#initiator  pm#object/result;

                    pm#agent__doer (pm#situation,pm#entity)  >  pm#organizer  pm#participant;

                        pm#organizer (pm#situation,pm#causal_entity);      pm#participant (pm#situation,pm#causal_entity);

                    pm#object/result (pm#situation,?)  >  pm#object  pm#instrument  pm#result;

                        pm#object__patient__theme (pm#situation,?)  >  pm#input  pm#input_output;

                            pm#input (pm#process,?)  >  pm#material  pm#parameter;

                            pm#input_output (pm#process,?)  >  pm#modified_object  pm#deleted_object;

                        pm#instrument (pm#situation,pm#entity);             pm#result (pm#situation,?)  >  pm#output;

                pm#from/to_place (pm#process,pm#spatial_object)  >  pm#from_place  pm#to_place  pm#via_place  pm#path;

            pm#experiencer (pm#situation,pm#causal_entity);                    pm#recipient (pm#situation,pm#entity)  > pm#beneficiary;

            pm#relation_from_process_only (pm#process,?)  >  pm#purpose  pm#triggering_event  pm#ending_event  pm#precondition
                pm#postcondition  pm#input  pm#input_output  pm#sub_process  pm#method  pm#from/to_place  pm#process_attribute;

                pm#triggering_event (pm#process,pm#event);                pm#ending_event (pm#process,pm#event);

                pm#precondition (pm#process,pm#situation);                pm#postcondition (pm#process,pm#situation);

                pm#sub_process (pm#process,pm#process);                   pm#method (pm#process,pm#description);

                pm#process_attribute (pm#process,pm#process_attribute_or_measure)  >  pm#manner;

    pm#relation_from_spatial_object__relation_from_a_spatial_object (pm#spatial_object,*)  >  pm#location;

        pm#location (pm#spatial_object,pm#spatial_object)
         >  pm#address  pm#on  pm#above  pm#in  pm#near  pm#interior  pm#exterior  pm#before_location;

    pm#relation_from_description_content/medium/container (pm#description_content/medium/container,*)
     >  pm#relation_from_description  pm#version dc#Coverage dc#Contributor dc#Source dc#Publisher dc#Rights  pm#authoring_time
        pm#author  dc#Language  dc#Format  pm#description_instrument  pm#description_object  pm#physical_support
        pm#rhetorical_relation  pm#argumentation_relation;

        pm#relation_from_description (pm#description,*)  > pm#description_container  pm#logical_relation  pm#contextualizing_relation;

            pm#logical_relation (pm#description,pm#description)  >  pm#and  pm#contextualizing_logical_relation;

            pm#contextualizing_relation (pm#description,*)  >  pm#contextualizing_logical_relation  pm#modality  pm#believer
                  pm#corrective_specialization  pm#corrective_generalization  pm#correction  pm#overriding_specialization;

        pm#argumentation_relation (pm#description_content/medium/container,pm#description_content/medium/container)
         >  pm#answer  pm#contribution  pm#replacement  pm#confirmation  pm#reference  pm#argument  pm#contradiction;

    pm#relation_from_type (pm#type,*) (^type of relations from a type, i.e. in RDF terminology, from a class or a property^)
     >  pm#exclusive_type  pm#relation_from_property  pm#relation_from_class;  //DAML, RDF, RDFS relations are categorized here

    pm#relation_from_collection (pm#collection,*)    //many kif#relations are categorized here
     >  pm#member  pm#size  pm#minimal_size  pm#maximal_size  pm#percentage  pm#average  pm#relation_between_collections;

    pm#relation_to_collection (*,pm#collection)  > kif#listof kif#setof pm#parts pm#relation_from_class_to_collection kif#item kif#cons;

    pm#different__different_from (?,?)  >  daml#different_individual_from  pm#exclusive_class,   /  pm#equal;

    pm#ordering_relation (?,?) (^e.g.  pm#kind,  pm#part,  pm#inferior_to^)  >  pm#partial_order_relation  pm#equivalence_relation;
```

Table 4: Examples of corrections

<: subtypeOf;   `l`: location; $(...)$: sub-annotation
#wn12347769 | Payne's_gray (^$('<' #blue removed since exclusive with #pigment, subtype of #substance)$ any pigment that produces a grayish to dark grayish blue^) < #pigment;
#wn07130190 | Anglia (^$('<' #England replaced by '=' #England)$ the Latin name for England^) = #England;
#wn07799755 | Mancunian (^$('<' Manchester replaced by '`l`' #Manchester)$ a resident of Manchester^) < #English_person,   `l` #Manchester;
#wn05168522 | transmission (^$('<' #communicating replaced by '<' #communication since the subtypes of this category indicate that it represents a transmission medium, not a process)$ communication by means of transmitted signals^) < #communication;

Table 5: Examples of link additions

>: subtype; ~: similar; `l`: location; `p`: part; //: comment

#yellow > pm#blond_color;
#name > pm#previous_surname pm#middle_name;
#agency > pm#real_estate_agency;

#region > #dry_land (pm);   //"(pm)" explicits the creator
#mass > #mass_unit (pm);   // of the link; this is needed
#city > #capital_city (pm);   // between WordNet types
#male > #male_person (pm);
#Tasmania `l` #Tasmanian_Island (pm);
#Great_Britain `p` #England #Wales (pm);
#acceleration > #acceleration_unit (pm);
#length > #distance (pm) #distance.size (pm);
#Venus.Roman_deity ~ #Aphrodite (pm);

Table 6: Examples of value/artificial categories

#dark_red (^$(value)$ a red that reflects little light^) #gram__gramme__gm__g (^$(value)$ a metric unit of weight equal to one thousandth of a kilogram^) #west_by_south__WbS (^$(value)$ the compass point that is one point south of due west^) #andante (^$(value)$ a moderately slow tempo^) #Monday__Mon (^$(value)$ the second day of the week; the first working day^) #mealtime (^$(value)$ the time for eating a meal^)
#thing.action (^$(artificial)$ an action; "how could you do such a thing?"^) #thing.happening (^$(artificial)$ an event: "a funny thing happened on the way to the..."^) #tonight (^$(artificial)$ the present or immediately coming night^) #then (^$(artificial)$ that time; that moment; "we will arrive before then"^)

Table 7: Example of schema

[any #flight (^$(no inheritance)$^), pm#from_place: a pm#spatial_object, pm#to_place: a pm#spatial_object, #day_of_the_week: a #day_of_the_week, pm#via_place: a pm#spatial_object, pm#departure_time: a pm#time_measure, pm#arrival_time: a pm#time_measure, may have for pm#relation_from_situation (^$(explore)$^): a pm#thing, pm#agent: an #airplane_pilot, may have for pm#experiencer: several #passenger](pm);

would be detected. Negative constraints such as "if t2 specializes t1, then t2 cannot be linked by any other kind of link to t1" have not been exploited either but it does not seem that WordNet 1.7 has many problems of this kind.

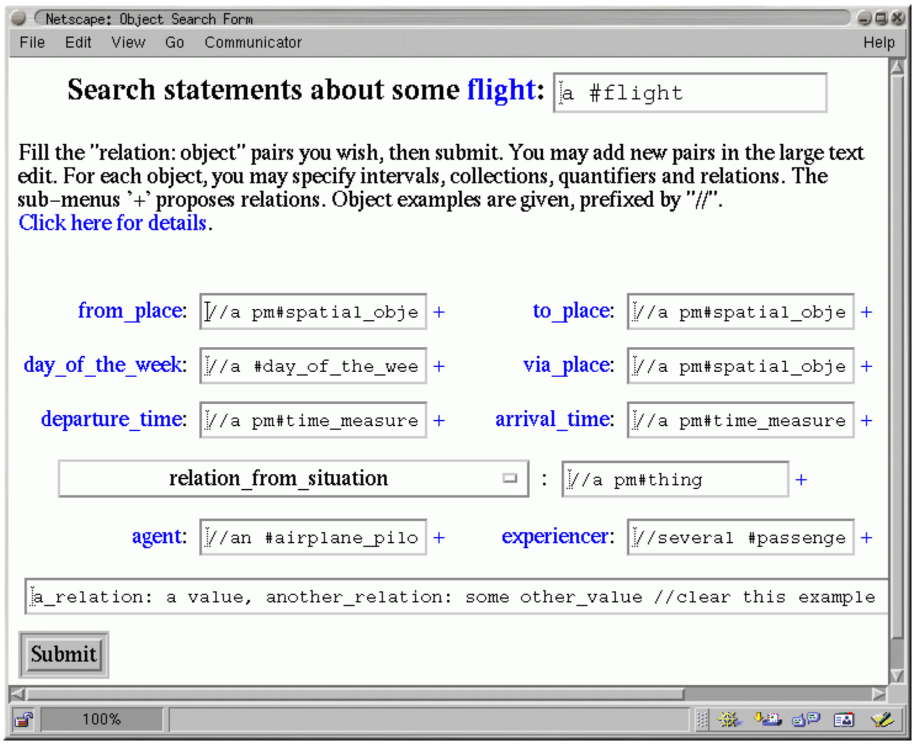
The corrections are documented [18]. Table 4 shows some examples in the WebKB-2 text backup format.

6. ADDITIONS

Up to March 1993 (and apart from the connections of WordNet upper categories to my top-level concept types), I have added 161 links, 17 during the integration of WordNet to WebKB-2 and 143 later when using the ontology for representing knowledge (thus, this excludes the 3000 specializations of WordNet categories that I have created for specific applications or domains, e.g. information technology). About 65 of these links connect WordNet categories, while 90 connect a WordNet category to a new specialization. Table 5 shows some examples in the FO notation.

I also added sub-annotations in some category annotations to guide or check knowledge representation. For example, since I do not want to distinguish between qualities and quales *using subtype links*, the subtypes of `pm#attribute_or_measure` representing values need to be distinguished in another way to prevent them being used within relations (or proposed in menus generated by WebKB-2 for relations). Hence, I checked all the subtypes of `pm#thing_that_can_be_seen_as_a_relation`) and added the string `$(value)$` in the annotations of about 1300 of them. (It should be noted that individuals are representing values and hence such sub-annotations are not required for them). I also added the string `$(artificial)$` in the annotations of WordNet categories that I found unfit for knowledge representation purposes, generally because they had a lexical rather than semantic character. Table 6 gives some examples.

Finally, I entered statements representing the most common relations that are or may be associated to certain categories. I call them schemas. Table 7 shows an example in FCG. WebKB-2 exploits such schemas to generate menus helping users to search or represent knowledge. Figure 1 shows an example based on the schema in Table 7 and where only this schema is exploited because of its sub-annotation `$(no inheritance)$`. The sub-annotation `$(explore)$` in a relation annotation directs WebKB-2 to present the subtypes of the type used for the relation, in a select menu (except for the subtypes marked as "value" or "artificial"). The '+' symbols in the menus permit the user to access sub-menus to detail relations from/to any destination object s/he has entered; in other words, menus can be cascaded to guide query/statement entering.


Netscape: Object Search Form
File Edit View Go Communicator Help
Submit
100%

Search statements about some flight: a #flight

Fill the "relation: object" pairs you wish, then submit. You may add new pairs in the large text edit. For each object, you may specify intervals, collections, quantifiers and relations. The sub-menus '+' proposes relations. Object examples are given, prefixed by "//".
Click here for details.

from_place: //a pm#spatial_obje + to_place: //a pm#spatial_obje +

day_of_the_week: //a #day_of_the_wee + via_place: //a pm#spatial_obje +

departure_time: //a pm#time_measure + arrival_time: //a pm#time_measure +

relation_from_situation : //a pm#thing +

agent: //an #airplane_pilo + experiencer: //several #passenge +

a_relation: a value, another_relation: some other_value //clear this example

Figure 1: A generated menu to help searching flights in the knowledge base.

7. CONCLUSIONS

The noun-related part of WordNet has been transformed into a genuine "lexical ontology" usable as a component in various knowledge-based applications: resource indexation, metadata registries, Yellow-pages like catalogs, query expansion, Semantic Web, etc. The focus was to guide and ease the representation, retrieval and sharing of general knowledge. This involved the generation of readable and unambiguous identifiers, the extraction of individuals, the merge with various top-level ontologies, and the correction of lexical and semantic problems. The result ontology is downloadable, browsable and extendible by anyone at http://www.webkb.org/.

Although I structured the top-level of WordNet and added a few links in other parts, the direct specializations of nearly all WordNet categories remain quite heterogeneous, with few exclusion links, and without distinction between role types and natural types. This lack of structure may be a problem for certain applications but fixing it might be as difficult as creating a better WordNet from scratch.

Another problem is that distinctions in WordNet seem to have often been made not simply on semantic grounds but also on lexical grounds, thus leading to a multiplicity of "artificial" categories or categories that should be connected but are not. A few categories have been marked as "artificial" but many more would need to be similarly marked, or connected by specialization links, to improve knowledge normalization and retrieval.

The next step is to integrate other ontologies from the IEEE Standard Upper Ontology library [25], in particular the Suggested Upper Merged Ontology (SUMO), and the DAML Ontology Library [26], in particular the CIA World Factbook. In the mapping that has been done between the SUMO and WordNet, one SUMO categories is often linked to several WordNet categories. That will give us cues to find and mark many WordNet categories as "artificial".

8. REFERENCES

[1] G. Miller. Wordnet: a Lexical Database for English. *Communications of the ACM*, 11:39–41, 1995. *http://www.cogsci.princeton.edu/˜wn/*

[2] A. Gangemi, N. Guarino and A. Oltramari. Restructuring Wordnet's Top-Level. *AI Magazine*, 40(5):235–244, fall 2002.

[3] N. Guarino and G. Vetere. Ontoseek: Content-based Access to the Web. *IEEE Intelligent Systems*, 14(3):70–80, October 1999.

[4] A. Puder and K. Romer. Generic Trading Service in Telecommunication Platforms. In *Proceedings of ICCS'97*. LNAI 1257:551–565, August 1997.

[5] A. Smeaton, F. Kelledy, R. O'Donnell, I. Quigley and E. Townsend. Expansion with WordNet and POS tagging of Spanish. In *Proceedings of IA'95*, Montpellier, France, 1995.

[6] C. Kwok, O. Etzioni and D. Weld. Scaling Question Answering to the Web. *ACM Transactions on Information Systems*, 19(3), July 2001.

[7] J. Allen. Different Kinds of Controlled Languages. *TC-Forum magazine*, 1(99):4–5, 1999.

[8] R. Colomb. Impact of Semantic Heterogeneity on Federating Databases. *The Computer Journal*, 40(5):235–244, 1997.

[9] J. Hendler. Agents and the Semantic Web. *IEEE Intelligent Systems*, 16(2):30–37, 2001.

[10] K. Knight and S. Luk. Building a Large-Scale Knowledge Base for Machine Translation. In *Proceedings of AAAI'94*, 773–778, Seattle, USA, July 1994.

[11] P. Martin. Using the Wordnet Concept Catalog and a Relation Hierarchy for Knowledge Acquisition. In *Proceedings of Peirce'95*, Santa Cruz, CA, August 1995. *http://www.webkb.org/doc/papers/peirce95/*

[12] P. Martin. Knowledge Representation in CGLF, CGIF, KIF, Frame-CG and Formalized-English. In *Proceedings of ICCS'02*, LNAI 2393:77–91. *http://www.webkb.org/doc/papers/iccs02/*

[13] P. Martin. *Knowledge Representation, Sharing and Retrieval on the Web*. Web Intelligence (Eds.: N. Zhong, J. Liu, Y. Yao). Springer-Verlag, January 2003. *http://www.webkb.org/doc/papers/wi02/*

[14] R. Mihalcea and D. Moldovan. Semantic Indexing using WordNet Senses. In *Proceedings of ACL Workshop on IR and NLP*, October 1990.

[15] J. Sowa. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, Reading, MA, 1984.

[16] J. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing Co., Pacific Grove, CA, 2000. See also *http://users.bestweb.net/ sowa/ontology/*

[17] R. Guha and R. McCool. TAP: a system for integrating web services into a global knowledge base. 2002. *http://tap.stanford.edu/*

[18] The WordNet 1.7 integration documentation. *http://www.webkb.org/doc/wn/*

[19] The Semantic Web. *http://www.semantic-web.org/*

[20] The DAML+OIL Ontology. *http://www.daml.org/2001/03/daml+oil.daml*

[21] The DOLCE Ontology. *http://wonderweb.semanticweb.org/*

[22] The Suggested Upper Merged Ontology. *http://ontology.teknowledge.com/*

[23] The HPKB Ontology. *http://WWW-KSL-SVC.stanford.edu:5915/*

[24] The CYC Top-level Ontology. *http://www.cyc.com/cyc-2-1/cover.html*

[25] The SUO Library. *http://suo.ieee.org/refs.html*

[26] The DAML Library. *http://www.daml.org/ontologies/*