The Integral Dictionary: An Ontological Resource for the Semantic Web Integration of EuroWordNet, Balkanet, TID, and SUMO

Dominique Dutoit (1, 2), Pierre Nugues (3), Patrick de Torcy (1)

(1) Memodata, 17, rue Dumont d'Urville, F-14000 Caen, France
d.dutoit@memodata.com, p.detorcy@memodata.com
(2) Université de Caen, CRISCO, F-14032 Caen, France
(3) Lund University, LTH, Department of Computer science, Box 118, S-221 00 Lund, Sweden
Pierre.Nugues@cs.lth.se

Abstract

The semantic organization of the web is one the major challenges for the future of the Internet. This important task may be based on the development of new approaches, taking the risk of reinventing the wheel, or may consider the previous efforts and successes, offering the opportunity to move research to market. This paper is a technical study that examines issues related to the latter possibility. We will first consider the structure of the Suggested Upper Merged Ontology (SUMO), which is a general proposal on the semantic web. We will then outline the challenges and possible strategies to integrate two existing ontologies, Wordnet for the English language and the Integral Dictionary for French (TID), to SUMO. Then, we will discuss the motivation of the mappings.

1. Introduction

The semantic organization of the web is one the major challenges for the future of the Internet. This important task may be based on the development of new approaches, taking the risk of reinventing the wheel, or may consider the previous efforts and successes, offering the opportunity to move research to market. This paper is a technical study that examines issues related to the latter possibility.

We will first consider the structure of the Suggested Upper Merged Ontology (SUMO), which is a general proposal on the semantic web. We will then outline the challenges and possible strategies to integrate various existing ontologies, Wordnet for the English language (Fellbaum 1998), EuroWordNet (Vossen 1999), Balkanet (Stamou 2002) and the Integral Dictionary (TID) for French, and other languages (Dutoit 1992), to SUMO (Niles 2001). Then, we will discuss the motivation of the mappings.

2. Resources

In this section, we summarize the content of each resource. We give more details on TID because it is not as well known as the others.

2.1 SUMO

According to its authors (Niles and Pease 2003), The SUMO (Suggested Upper Merged Ontology) is

> "an ontology that was created at Teknowledge Corporation with extensive input from the SUO (IEEE standard upper ontology group) ontology mailing list, and it has been proposed as a starter document for the IEEE-sanctioned SUO Working Group. The SUMO was created by merging publicly available ontological content into a single, comprehensive, and cohesive structure. As of February 2003, the ontology contains 1000 terms and 4000 assertions."

The general organization of SUMO is an acyclic oriented graph. Table 1 shows details of this ontology. It mentions that there are 631 classes in SUMO and that 175 classes are linked by a Domain relation to one or more ObjectProperty.

Son	Parent	Class		DatatypeProperty		ObjectProperty		Description		Total number of relations where the node is a soon
Class	631	DISJOINT	34	DOMAIN	27	DOMAIN	175			236
		SUBCLASS	655	RANGE	25	RANGE	174			854
		TYPE	62							62
DatatypeProperty	28	TYPE	64							64
ObjectProperty	207	TYPE	384							384
Description	62	TYPE	52					SUBATTR	12	64
								CONTRARYATTR	13	13
Total number of relations where the node is a father			1251		52		349		25	1677

Table 1: Statistical summary of SUMO.

Figure 1 shows an example of Domain relation between the Agent class and an ObjectProperty called author.

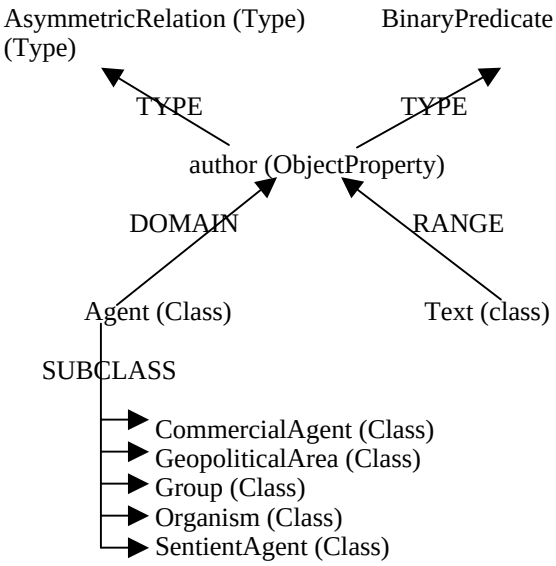

Figure 1: A part of SUMO.

Additionally, SUMO adds a comment to the author ObjectProperty:

> *(authors ?AGENT ?TEXT) means that ?AGENT is creatively responsible for ?TEXT. For example, Agatha Christie is author of Murder_on_the_Orient_Express.*

We extracted statistics on classes, properties, and relations using the OWL version of SUMO, which is in XML format. The original format for SUMO is a variant of KIF, which has this structure:

```
(instance legalRelation BinaryPredicate)
(instance legalRelation SymmetricRelation)
(domain legalRelation 1 CognitiveAgent)
(domain legalRelation 2 CognitiveAgent)
(documentation legalRelation "(&%legalRelation
?AGENT1 ?AGENT2) means that ?AGENT1 and ?AGENT2
are relatives by virtue of a legal relationship.
Some examples include marriage, adoption, etc.")
```

Normally, the format type has no effect on the results, but unfortunately we have noticed some inconsistencies in OWL. So, our statistics refer to the OWL format and differ from those of the original SUMO in KIF.

2.2 WordNet

WordNet is a famous, comprehensive ontology available for English (Fellbaum 1998). Building on the WordNet popularity, the EC project EuroWordnet (Vossen 1999) has adapted its architecture to other languages like French. Many other similar projects like EuroWordNet exist today. So, WordNet was naturally the first choice to flesh out and validate SUMO's design (Niles and Pease 2003).

To date, all the nouns WordNet synsets have been mapped by the SUMO team to 1,000 terms of the SUMO ontology. WordNet 1.6 was used.

Although this integration is now complete, it leaves open some questions: is the mapping neutral or not? Was it possible to integrate without loss all the Wordnet knowledge in SUMO? Are the different relations of Wordnet 2.0 all well represented? How would it scale up to EuroWordNet or Balkanet, a similar EC project concerning the Balkan languages, in which we are also involved.

2.3 The Integral Dictionary

The Integral Dictionary, TID, (Dutoit 1992) is a semantic network associated to a lexicon. It's available mainly for French and being adapted to other languages notably English and German. Its size is comparable to that of WordNet. The Integral Dictionary organizes words into a variety of concepts and uses semantic lexical functions. Concept definitions are based on the componential semantic theory, the decomposition of the words into a set of smaller units of meaning, and the lexical functions are inspired by the Meaning-Text theory.

The basic component of TID is called a "concept". Each concept is annotated by a gloss written in mostly in French that describes intentionally its content. It consists of three main ontologies:

- A first ontology is based on the relations generic or specific. When a concept is entirely lexicalized, a particular relation between the concept and the literal is used: generic. When the word does not describe the concept entirely, the relation is said to be specific.
- A second one is based on a thesaurus, similar to the Roget's, but more linguistically restricted. It includes thousands of themes (domains or small conceptual worlds).
- The third ontology describes lexical-syntactic patterns.

The Integral Dictionary also contains a large number of lexical functions that generate word senses from another word sense given as an input.

One important property of the Integral Dictionary is its structure: merging several approaches (hence its name), the Integral Dictionary is fundamentally an acyclic oriented graph instead of a tree.

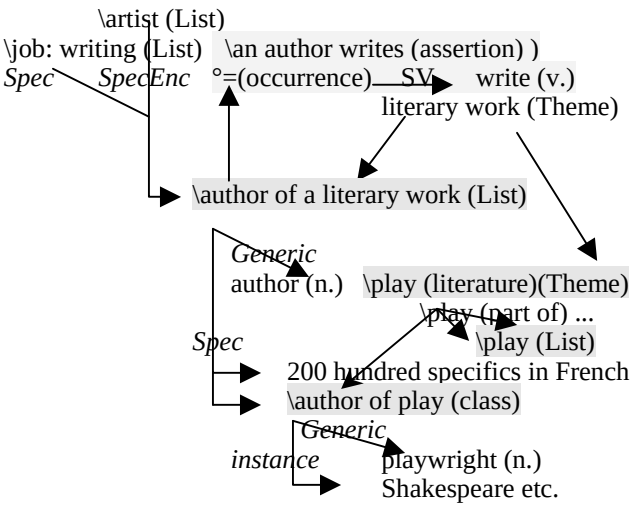

Figure 2: An excerpt of TID.

Figure 2 is an excerpt of TID, which shows that:

- The class \author (List) is possibly subsumed by the class: \artist (List). (*Enc* means potentiality, *Spec Enc* means "is a" potentially.
- In this class, the generic word in English is author.n.
- The class contains a subclass labeled "\author of play", which is a specific.
- Shakespeare is an instance of the previous class.
- The class \author (List) belongs to a theme, a possible topic called \literary work (theme).
- This theme contains the subtheme \play (literature) (Theme).
- Finally, the \author (List) is directly connected to a part of its preferred assertion: *write* (a literary text).

We call relation a link from a node to another node and we never count the symmetrical links. For French, TID contains around 220,000 relations similar to that of the example in Figure 2. Concerning the lexical function borrowed from the Meaning-text theory, we have also 150,000 occurrences of relations for French. A part of them, 15,000, is not validated yet.

The multilingual part, English, Italian, Spanish, German, Dutch, and Portuguese, represents 300,000 relations to add to the previous number.

3. Integration

The task of integrating many various resources is a priority for several reasons:

- TID was designed originally for French and when we ported it other languages, we observed discrepancies and even contradictions in concepts. We tried to reconcile them and we believe that a multilingual viewpoint is crucial to clarify some questions in semantics.
- Nothing certifies that our approach is the best one. On the contrary, most models bring us new ideas and content. It is the case of SUMO, as we will see it.
- One meaning of the title of our work (The Integral Dictionary) addresses precisely this task.

SUMO is appealing because of the development of the semantic web technologies. Merging many independent ontologies should grow in importance in the near future. But, we have to solve problems concerning the nature of the content of the ontology. Can we duplicate information? How to solve inconsistencies? etc.

3.1 General questions concerning integration

The SUMO team merged its ontology with WordNet:

> *Once we decided to restrict our attention to noun synsets, we had to settle on the relations to be used to map these synsets to SUMO concepts. There are three possible relations of interest: synonymy, hypernymy, and instantiation.*

Let's focus on the *synonymy* relation and how it is dealt with in the case of ontology merging. Notably, if the synonymy relation still corresponds to its standard definition: "Two different words that express the same meaning in at least one context."

In WordNet, we consider one English word. On the other hand, in SUMO, we have a formal term that exists only in the formal ontology. The question is about the existence of at least one context where these two data can be substituted. And such context may not exist. We can dispute that SUMO and WordNet terms are synonyms, strictly speaking, because they will never share a common context.

Now the problem is to determine if this result is a major problem or, alternatively, a good opportunity. Let's make the assumption that the elements are synonyms if, and only if, we can find exactly the same information in the two models. In this case, we can observe that the data are redundant and have not to be duplicated. But if we consider that the goals of WordNet and SUMO are different, we can try to identify what is the available knowledge from WordNet and what is available from SUMO. For a notion like the predicate *author*, we have in WordNet 1.6:

- hypernym(author)=communicator
- hypernym(communicator)=human being
- hypernym(human being)=living thing (and other word in the synset)
- hypernym(human being)=causal agent
- meronym(human being)=people

...

If we compare these data to the data written out from SUMO in Figure 2, we realize that the two records do not match. They merely offer two different viewpoints on a same word.

More precisely, the comparison between data of SUMO and WordNet shows crucial differences. In WordNet, as in a dictionary, the synset describing *author* is a definition and contains hyponyms like *{a speaker or writer who makes use of alliteration}*, *{a writer whose work is published in a newspaper or magazine or as part of a book}*, etc.

In SUMO, instead of these data, we find an assertion about the formal *author* (*ObjectProperty*), which defines objects able to be authors (*CommercialAgent, GeopoliticalArea, etc.)* and other axioms in the logical part of SUMO. The conclusion of this paragraph is very simple: the two ontologies do not deal with the same thing. So to our initial questions:

- Is the mapping neutral or not? The answer is no. Things are different.
- Is it possible to integrate all the Wordnet knowledge in SUMO without loss? Again, no. The goals are different.
- Are the different relations of Wordnet 2.0 all well represented? No. SUMO is not designed to register this information.

The three negative answers do not give any information about why and how to deal with SUMO with regard to linguistic ontology.

We consider the problem very similar to the modeling of syntactic relations when we related them to the paradigmatic dimension. We solved it using two relational models that we integrated in a same database. The first one described syntactic patterns and the second one, hierarchical data.

5. Technical Solution

Basically, TID, WordNet and SUMO are acyclic oriented graphs. Let's consider the relations in Figure 2 again. Figure 3 shows the initial data format that TID used to represent it.

Child	*Parent*	*KindOfRel*
Author (n)	*\author of a lit…*	*Generic*
\author of play	*\author of a lit…*	*Specific*
etc.		

Figure 3: A general record in the table RELATION in TID.

Although this format was satisfactory for hierarchical data, it reached its limits when we introduced syntactical relations. Let's consider the syntactic definition in Figure 2:

\author of a literary work (List)	SV	\write
	VO	\texts

Figure 4 shows the table in TID using the same formalism.

Child	*Parent*	*KindOfRel*
\author of a literary work (List)	\write	SV
\write	\texts	VO
etc.		

Figure 4: A part of TID.

However, it not possible to consider that *\author of a literary work (List)* is the child of *\write* and the grandchild of *\text* in Figure 4 in the same way it is the child of *\author of a lit…* in Figure 3. In addition, in terms of graph, Figure 4 cannot record the syntactic paths without ambiguity, for example if *write* exists in many different assertions.

Syntactic patterns and lexical ontology represent two different viewpoints that are not necessarily related. To represent them with a relational database, we must take into account that these two dimensions (syntactic/paradigmatic) are different. Figure 5 shows the integration results where

- OntoTID means ontology of TID and SyntTID means Syntactical Pattern of TID.
- The index (1) is the key of the complete pattern.
- The two last records indicate that OntoTID and SyntTID are parts of TID.

This format is more flexible and provides rich new possibilities. First, the format can record any kind of hypergraph in a relational database. Second, it enables us to extend the group theory approach to a more general mereology.

Child	*Parent*	*KindOfRel*	*Location*
Author (n)	*\author of a lit…*	*Generic*	*OntoTID*
\author of play etc.	*\author of a lit…*	*Specific*	*OntoTID*
\author …(List)	*\write*	*SV (1)*	*SyntTID*
\write etc.	*\texts*	*VO (1)*	*SyntTID*
OntoTID	*PartOfTID*	*part of*	*TID*
SyntTID	*PartOfTID*	*part of*	*TID*

Figure 5: A part of TID.

We have used this format to integrate a set of ontological resources. Concerning EuroWordNet and Balkanet, the format allows us to upload data from xml files to a relational database. Figure 6 shows an excerpt of records where (1) is a key identifying a synset.

Since a synset has its gloss and literal, we have the English gloss {writes (books or stories or articles or the like) professionally (for pay)…} and the English literal *author* located in the *English WordNet.* We notice that in this case, *auteur (n)* is placed in the synset (1) in the *French WordNet.* In the end, it's also possible to generate the complete list of InterLingua index (ILI).

Child	*Parent*	*KindOfRel*	*Place*
Author (n)	*(ILI 1)*	*Literal*	*EnWordNet*
{writes (books or stories or articles or the like) professionally (for pay)…}	*(ILI 1)*	*Gloss*	*EnWordNet*
auteur (n)	*(ILI 1)*	*Litteral*	*FrWordNet*
(ILI 1)	*Interlingua*	*Elementof*	*ILIs*

Figure 6: The WordNets.

Figure 7 shows the integration of WordNet(s) data from Figure 2 to TID and Figure 8, the integration of SUMO data.

Child	*Parent*	*KindOfRel*	*Location*
(1)	*\author of a literary work (List)*	*Generic*	*TID*

Figure 7: Integration of WordNet into TID.

Child	*Parent*	*KindOfRel*	*Location*
\Agent(Class)	*Author(Object Property)*	*Domain (X)*	*SyntSUMO*
Author(Obj….)	*Text(Class)*	*Range*	*SyntSUMO*
Author(Obj….)	*(ILI 1)*	*SUMOItem*	*SUMO*

Figure 8: SUMO in TID.

6. Conclusion

In this paper, we have described a strategy to support a variety of semantic initiatives. It underlines the complementary nature of the views concerning the linguistic sign.

We have also showed how to group these different ontologies in a single 'mereological' database. The tool that manages the database is called LEXIDIOM (Java and Firebird).

7. References

Dominique Dutoit. 1992. A set-theoretic approach to lexical semantics, *Proceedings of Coling-92***.**

Christiane Fellbaum (ed). 1998. WordNet: An Electronic Lexical Database, MIT Press.

Ian Niles, Adam Pease. 2001. Towards a Standard Upper Ontology. In *Proceedings of FOIS-2001*.

Ian Niles, Adam Pease. 2003. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of IKE '03*.

Piek Vossen. 1999. *EuroWordNet, Building a multilingual database with wordnets for several European languages*, University of Amsterdam.

Sofia Stamou, Kemal Oflazer, Karel Pala, Dimitris Christoudoulakis, Dan Cristea, Dan Tufis, Svetla Koeva, George Totkov, Dominique Dutoit, Maria Grigoriadou. 2002. Balkanet: A multilingual Semantic Network for Balkan Languages, In *Proceedings of the First International WordNet Conference*, Mysore, India.