

Brigham Young University BYU ScholarsArchive

International Congress on Environmental Modelling and Software 4th International Congress on Environmental Modelling and Software - Barcelona, Catalonia, Spain - July 2008

Jul 1st, 12:00 AM

An approach to formal modeling of environmental knowledge via concept maps and ontologies

A. Simón-Cuevas

L. Ceccaroni

A. Rosete-Suárez

Follow this and additional works at: https://scholarsarchive.byu.edu/iemssconference

Simón-Cuevas, A.; Ceccaroni, L.; and Rosete-Suárez, A., "An approach to formal modeling of environmental knowledge via concept maps and ontologies" (2008). *International Congress on Environmental Modelling and Software*. 142. https://scholarsarchive.byu.edu/iemssconference/2008/all/142

This Event is brought to you for free and open access by the Civil and Environmental Engineering at BYU ScholarsArchive. It has been accepted for inclusion in International Congress on Environmental Modelling and Software by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.

An approach to formal modeling of environmental knowledge via concept maps and ontologies

A. Simón-Cuevas¹, <u>L. Ceccaroni²</u> and A. Rosete-Suárez¹

 ¹ Centro de Estudios de Ingeniería de Sistemas (CEIS), Facultad de Ingeniería Informática, Instituto Superior Politécnico "José Antonio Echeverría", 114, No. 11901, Marianao, C. Habana, Cuba ({asimon, rosete}@ceis.cujae.edu.cu)
² Departament de Llenguatges i Sistemes Informàtics (LSI), Universitat Politècnica de Catalunya (UPC), Campus Nord, Edif. Omega, C. Jordi Girona, 1-3, 08034 Barcelona, Spain (luigi@lsi.upc.edu)

Abstract. In order to obtain semantic interoperability in open systems, the systems involved need to agree on commonly understandable knowledge representations. The case considered in this paper is one of a system needing to communicate and share knowledge using concept maps and ontologies, in the context of the environmental sciences and the semantic Web. We present how to formally obtain ontologies codified in the OWL language from concept maps. Concept maps are a flexible and informal form of knowledge representation, while OWL is a language oriented to processing carried out by machines. The mapping between concept maps and ontologies is a formal transformation, which semantically analyzes the relations linking the concepts in the map. The proposed method includes a concept-sense disambiguation procedure and uses the WordNet lexical knowledge base. It also includes automatic learning of the semantics of the relations between concepts. The proposed method has been applied to the environmental-knowledge domain, through tests of several concept maps with labels in Spanish.

Keywords: Semantic analysis; knowledge sharing; concept map; ontology; OWL.

1. INTRODUCTION

In the environmental sciences, as in most scientific domains, information needs sometimes to be analyzed and processed by machines. In the knowledge representation oriented to the semantic analysis and processing by machines, context in which a certain degree of formalization is required, the development and use of *ontologies* is increasingly common. However, the processes of designing and creating ontologies, the tools available, and the specification languages are still complex for non-experts in this subject. This complexity represents a difficulty in environments requiring the collaboration of humans for the development and processing of ontologies. This suggests that a form of representation that can be used naturally by humans and integrated with ontologies (in such a way that the latter can be automatically obtained) should be useful.

Concept maps (CMs) [Novak and Gowin, 1984] are one of these human-friendly knowledge-representations. The integration between CMs and ontologies, specifically in the case of OWL [Smith et al. 2004] ontologies, is pursued through the incorporation of more formalization in CMs and through the analysis of the relations among concepts. Parts of the proposed method are based on a concept-sense disambiguation (CSD) algorithm defined by Simón et al. [2007a], and on WordNet [Miller et al. 1990]. The CSD algorithm

tries to assign the most rational sense of a given concept in the CM, using WordNet, *contextual analysis* and *domains information*. The algorithm explores the context in which a given concept appear in the CM and try to determine a corresponding, similar context in WordNet, using the *synsets* of the concept at issue. A similar contextual analysis is carried out with the gloss.

Along the paper, to represent the English translation of the Spanish terms used, the following notation will be used: *español* ("Spanish").

2. CONCEPT MAPS

CMs are a tool especially defined for application in the learning process; they are easy to be created, flexible and intuitive for people. They are a graphically rich technique for organizing and representing knowledge proposed by Novak and Gowin [1984]. They include *concepts, linking-words* (relationships between concepts) and *propositions*. CMs' propositions contain two or more concepts connected using linking-words and sometimes are called semantic units, or units of meaning. The concepts are represented in a hierarchical fashion with the most inclusive, most general concepts identified as *root-concept*, at the top of the CM and the more specific, less general concepts arranged hierarchically below. Figure 2 shows an example of a CM, about the nitrogen cycle as a part of the ecosystem in which *Nitrógeno* ("nitrogen") is the *root-concept*. The CM shown is an enriched interpretation of the picture in Figure 1. CMs, being easily created, flexible and intuitive for people, are especially useful for knowledge management.



Figure 1. Representation of nitrogen cycle obtained from Jones et al. [1992].

3. ONTOLOGIES AND OWL DL

In artificial intelligence, ontologies were introduced to share and reuse knowledge. They provide the common reference frame for communication languages in distributed environments (such as multi-agent systems or the semantic Web) and a formal description for automatic knowledge processing. Several languages have been defined to implement them; OWL [Smith et al. 2004] is the latest, standardized ontology language. OWL is based on XML, RDF and RDFS, and includes three specifications, with different expressiveness levels: OWL Lite, OWL DL and OWL Full. The code obtained by the method described in this paper is generated according to OWL DL specifications. OWL DL is so named due to its correspondence with *description logics*. *Description logic* (DL) is the name for a family of knowledge representation formalisms that represent the knowledge of a domain by first defining the relevant concepts of the domain (its terminology), and then using these concepts to specify properties of objects and individuals occurring in the domain [Baader and Nutt, 2003]. The terminology specifies the vocabulary of a domain, which consists of concepts and roles, where the concepts denote individuals while roles denote binary relationship between individuals.



Figure 2. Concept map of the nitrogen cycle (in Spanish).

4. BASIC ASPECTS OF CM-OWL MAPPING

The knowledge in OWL ontologies is expressed as *classes*, *properties* and *instances* [Smith et al. 2004], while in CMs much of this formal and explicit specification does not exist, and has to be inferred. Nonetheless, some initial structural mapping between CMs and OWL can be easily established [Simón et al. 2007b]:

- concepts correspond to: classes and instances;
- *linking-words* correspond to *properties* (A property is a binary relation between instances of classes in OWL [Smith et al. 2004].);
- propositions correspond to classes and properties' restrictions and other OWL constructs.

Some type of semantic relation, such as *class-subclass*, *class-property*, *class-property-value*, *class-instance*, can be inferred from certain *linking-words* used in CMs, in accordance with Brilhante et al. [2006].

In addition to the CM to be formalized, two external knowledge sources will be used in this work: WordNet [Miller et al. 1990] and a CM repository. WordNet is a lexical knowledge base, whose basic structure is the *synset*. *Synsets* form a semantic network and are interconnected among themselves by several types of relations, some of which are used in the proposed algorithm, such as *hypernymy-hyponymy* (class/subclass) and *meronymy-holonymy* (part/whole). The *synset* define the meanings of a word, which, in the case of polysemy, can be found in various *synsets*. WordNet can be used as an ontology if its links are associated to a formal semantics. The CM repository used here is ServiMap [Simón et al. 2006], which stores several CMs of different domains (including the environmental domain).

The mapping and semantic inference leading to OWL coding in this paper, is carried out combining the analysis of:

- the syntax of the propositions;
- the occurrence of similar relations in WordNet and the external CM repository.

Initially, some frequently used *linking-words* are defined and organized in four **categories**, according to the semantics that can be associated to them and their correspondence with the semantic relations in WordNet. They are:

- Classification (CC), for *linking-words* that may indicate (*super class-class*)relations between concepts in a *proposition* (e.g., *comprende a* ("comprise") in the proposition (*Plantas, comprende a, Algas*) in Figure 2); corresponding to *hypernymy* and *hyponymy* relations in WordNet;
- Instance (IC), for *linking-words* that may indicate (*class-instance*) relations between concepts in a *proposition* (e.g., *por ejemplo* ("for example") in the proposition (*Bacteria, por ejemplo, Bacteria Nitrificante*) in Figure 2);
- Property (PC), for *linking-words* that may indicate (*class-property*) relations between concepts in a *proposition* (e.g., *tiene* ("has")); corresponding to *has_meronym* and *has_holonym* relations in WordNet;
- Property-Value (PVC) for *linking-words* that may indicate (*class-property-value*) relations between concepts in a *proposition*, such as nouns (e.g., *lugar* ("place") in the proposition (*Nitrógeno en Océano*, *lugar*, *Océano*) in Figure 2); corresponding to basic *meronymy* and *holonymy* WordNet's relations, and different from the more specific *has_meronym* and *has_holonym* relations (e.g. *has_mero_partOf* and *has_holo_madeOf*).

This method allows everyday natural language to be used at CM construction time. Lexemes are used to avoid duplications due to verb forms' variability, e.g.: *ten-* (from *tener*) instead of *tiene* ("has") and *tienen* ("have"). The *linking-words* are defined in Spanish, which is the natural language used for this study, and they are continually and automatically enriched.

In the mapping method, the CM under consideration is analyzed as a structured text. A concept sense-disambiguation algorithm [Simón et al., 2007a] is used to infer the most rational sense (in terms of WordNet's *synsets*) for all concepts in the CM. Once inferred a *synset* for each concept in a *proposition*, the semantics of the CM relation among them can be inferred from the relation in WordNet (if one exists).

5. OBTAINING OWL-DL ONTOLOGIES

In this section, the process of obtaining OWL-DL ontologies form CMs is presented. It begins with a CM expressed as an XML file, in which the elements are ordered using a breadth-first strategy (beginning from the *root-concept*) to explore the content of the CM. The content structure (concepts, *linking-words* and propositions) of the CM is maintained in the XML file. The process is organized in three phases: preprocess, mapping and codification. Four modules are defined for the implementation of the system: parser, disambiguator, semantic interpreter and OWL codifier.

5.1 Preprocess phase

The *parser* analyzes the CM to be translated to OWL, identifying propositions and their parts (*concepts* and *linking-words*):

- It extracts all propositions from the CM's XML file and creates a proposition set (PS) with (C_o , *linking-word*, C_d) as basic structure, where C_o is the origin concept, and C_d is the destination concept.
- It creates the concepts set (CS) which includes all concepts in PS.
- It extracts from the CM repository all propositions with at least one concept included in CS and creates, with these, a second proposition set (PS-CMR), which is used to infer symmetric properties.

The *disambiguator* infers the most rational *sense* (in terms of WordNet *synsets*), using the algorithm defined by Simón et al. [2007a], and analyzes this sense according to the relations between *synsets* found in WordNet:

- A **PS_WN**_{hype-hypo} set is created with the pair (C, C'), if the synsets of two concepts C and C' are directly related by a hyperonymy or hyponymy relation in WordNet.
- A **PS_WN**_{mero-holo} set is created with:
 - the pair (C, C') if the *synsets* of two concepts *C* and *C*' are directly related by a *has meronym* WordNet's relation and;
 - the pair (C', C) if the *synsets* of two concepts *C* and *C*' are directly related by a *has_holonym* WordNet's relation.
 - A **PS_WN**_{mero-holo-type} set is created with:
 - the triple (C, C', type of relation) if the *synsets* of two concepts C and C' are directly related by some type of *meronymy* WordNet's relation (e.g. *has_mero_madeOf*), different from *has_meronym* and;
 - the triple (C', C, type of relation) if the *synsets* of two concepts C and C' are directly related by some type of *holonymy* WordNet's relation (e.g. *has_holo_madeOf*), different to *has_holonym*.

5.2 Mapping phase

Several heuristic rules (not presented here) are defined for mapping between the propositions included in PS to OWL constructs. These rules use the set of *linking-words* (*lw*) included in the categories presented in section 4, the sets generated in the *preprocess phase* (see section 5.1) and the sets generated as follows:

- **S_CS** is a set of pairs of concepts (C, C') which are part of a proposition of type (*superclass, lw, subclass*), where C is the superclass and C' is the subclass; the *lw* included in CC and the PS_WN_{hype-hypo} set are used.
- **S_CI** is a set of pairs of concepts (C, I) which are part of a proposition of type (*class, lw, instance*), where C is the class and I is an instance of C; the *lw* included in *IC* are used.
- **S_CP** is a set of pairs of concepts (C, Pr) which are part of a proposition of type (*class, lw, property*), where C is the class and Pr is a property of C; the *lw* included in PC and the PS_WN_{mero-holo} set are used.
- S_CPV is a set of propositions of type (*class, property, range of value*); the *lw* included in *PVC* and the PS_WN_{mero-holo-type} set are used.
- **S_CPV**_{hasValue} is a set of propositions of type (*class*, *property*, *instance*).
- **Pr**_{symmetric} is a set of symmetric properties.
- **Pr**_{functional} is a set of functional properties.
- Intersection_{classes} is a set of pairs $(C, \{..., C'_{i}...\})$, where C is subclass of each C'_i;
- **Intersection**_{classes-property} is a set of triples (*C*, *C*', *Pr*), where *C* is subclass of *C*' and *Pr* is one property of *C*;
- Union is a set of pairs $(C, \{..., C'_{i},...\})$, where C is superclass of each C'_{i} .

The *semantic interpreter* applies the set of heuristic rules to the propositions obtained by the *parser*.

5.3 Codification phase

The *OWL codifier* uses the sets generated by the *semantic interpreter* and writes out the corresponding OWL constructs according to W3C Recommendation [Smith et al. 2004], considering the mapping conventions shown in Table 1.

6. APPLICATION IN THE ENVIRONMENTAL DOMAIN

The theoretical modeling method presented has been applied to several CMs about the environmental domain. These CMs have been constructed from different source (texts and

figures) with the assistance of environment experts (e.g., the Figure 2 is constructed using the Figure 1 as source). In summary, the modeling method starts with the CM constructed using Macosoft CM editor [Simón et al. 2006] and expressed as an XML file. In the preprocess phase, all propositions are extracted by the *parser* and stored in *PS*; propositions in the CM repository whose concepts appear in *PS* are also obtained by the *parser*.

Inferred sets	Basic structure	OWL constructs
S_CS	(C, C')	C is coded as owl:class. C' as owl:class and
		also as <i>rdf:subClassOf</i> (of <i>C</i>).
S_CI	(C, I)	C is coded as <i>owl:class</i> . I as an instance of C:
		.
S_CP	(C, Pr)	An <i>owl:ObjectProperty</i> with the name formed
		by the concatenation of labels <i>tiene</i> ("has") and
		the one used for Pr is coded. C is coded as
		owl:class and also as rdfs:domain (of
		owl:ObjectProperty). Pr is coded as owl:class
		and also as <i>rdfs:range</i> (of <i>owl:ObjectProperty</i>).
S_CPV	(C, Pr, C')	C is coded as owl:class. Pr as owl:onProperty
		(of C). C' as restriction owl:someValueFrom
		(of Pr) in the code of C .
S_CPV _{hasValue}	(C, Pr, C')	C and C' are coded as <i>owl:class</i> . Pr is code as
		owl:onProperty (of C) with restriction of
		owl:hasValue being C' the value.
Pr _{symmetric}	$Pr_1 \dots Pr_n$	The <i>owl:SymmetricProperty</i> construct is
		incorporated to the code of each Pr_i .
Pr _{functional}	$Pr_1 \dots Pr_n$	The owl:FunctionalProperty construct is
*		incorporated to the code of each Pr_i .
Intersection _{classes}	$(C, \{C'_{i}\})$	C and C'_i are coded as <i>owl:class</i> . The collection
		of C'_i is codes as <i>owl:intersectionOf</i> .
Intersection _{class-property}	(C, C', Pr)	C is coded as owl:class. C' is coded as
		owl:class. Pr as owl:onProperty (of C), with
		restriction <i>owl:hasValue</i> of C'. The collection
		of C' and Pr is coded as owl:intersectionOf.
Union	$(C, \{C'_{i}\})$	C is coded as owl:class. C'_i as owl:class.
		Collection of C' _i as <i>owl:unionOf</i> .

Table 1. Conventions for mapping between inferred sets and OWL constructs.

Next, the sense of each concept included in *PS* is inferred, and the *hyperonymy* and *meronymy* relations between *synsets* in a Spanish version of WordNet (the version developed by the Natural Language Processing Group (TALP) of the Software Department (LSI) of the Technical University of Catalonia (UPC)) are identified by the *disambiguator*. We use a portion of the CM shown in Figure 2 as the *experimental set* (ES), which includes the propositions for the following modeling example. All propositions included in ES are sequentially analyzed by the *semantic interpreter*:

- In the case of proposition (*Nitrógeno, agrupa, Nitrógeno en Tierra*), the rule *if* $P = (C_o, lw, C_d) \in PS \land lw \in CC$ then $PS = PS \{P\}$, $S_CS = S_CS \cup \{(C_o, C_d)\}$ is executed. The *lw agrupa* ("groups") is included in Classification category.
- In the case of proposition (Organismos, una de sus clasificaciones, Microorganismos), the rule if P = (C_o, lw, C_d) ∈ PS ∧ lw ∉ CC ∧ (C_o, C_d) ∈ PS_WN_{hype-hypo} then PS = PS {P}, S_CS = S_CS ∪ {(C_o, C_d)}, CC = CC ∪ {lw} is executed. The synsets {ser_vivo#1, ser#1, organismo#1} of concept Organismo ("organism") and {microorganismo#1} of concept Microorganismo ("microorganism") are inferred by the disambiguator and the has_hyponym relation between them is found in WordNet; consequently, the lw una de sus clasificaciones ("one of their classification") is added to Classification category.

- In the case of proposition (Aire, compuesto de, Oxígeno), the rule if $P = (C_o, lw, C_d) \in PS \land (C_o, C_d, type) \in PS_WN_{mero-holo-type}$ then $PS = PS \{P\}$, $S_CPV = S_CPV \cup \{(C_o, property, C_d)\}$, $PVC = PVC \cup \{lw\}$ is executed. In this rule, property, is the property corresponding to type in PVC, for example, hecho de ("made of") in case of has_mero_madeof relation. The synsets {aire#1} of concept Aire ("air") and {oxígeno#1, O#1, número_atómico_8#1} of concept Oxígeno ("oxygen") are inferred by the disambiguator and the has_mero_madeof relation between them is found in WordNet; consequently, the lw compuesto de ("compose of") is added to the Property-Value category.
- In the case of proposition (*Líquido, instancia de, Estado*), the rule if P = (C_o, lw, C_d) ∈ PS ∧ lw ∈ IC then PS = PS {P}, S_CI = S_CI ∪ {(C_o, C_d)} is executed. The *lw instancia de* ("instance of") is included in Instance category.
- In the case of proposition (Agua, estado, Líquido), the rule if (C, Pr, I) ∈ S_CPV ∧ ∃(C', I) ∈ S_CI then S_CPV_{hasValue} = S_CPV_{hasValue} ∪ {(C, Pr, I)} is executed. The *lw estado* ("state") is a noun according to WordNet, therefore (Agua, estado, Líquido) is included in S_CPV as result of the rule if P = (C_o, *lw*, C_d) ∈ PS ∧ *lw* ∈ PVC then PS = PS {P}, S_CPV = S_CPV ∪ {(C_o, *lw*, C_d)} previously executed, and the pair (Líquido, Estado) is included in S_CI.

Finally, the code obtained by OWL codifier is:

<owl:Class rdf:ID="Nitrógeno"/> <owl:Class rdf:ID="Oxígeno"/> <owl:Class rdf:ID="Estado"/> <owl:Class rdf:ID="Organismos"/> <owl:Class rdf:ID="Nitrógeno en Tierra"> <rdfs:subClassOf rdf:resource="#Nitrógeno"/> </owl:Class> <owl:Class rdf:ID="Microorganismos"> <rdfs:subClassOf rdf:resource="#Organismos"/> </owl:Class> <owl:Class rdf:ID="Aire"> <rdfs:subClassOf> <owl:Restriction> <owl:onProperty rdf:resource="#hecho de" /> <owl:someValuesFrom rdf:resource="#Oxígeno" /> </owl:Restriction> </rdfs:subClassOf> </owl:Class> <Líquido rdf:ID="Estado"/> <owl:Class rdf:ID="Agua"> <rdfs:subClassOf> <owl:Restriction> <owl:onProperty rdf:resource="#estado" /> <owl:hasValue rdf:resource="#Líquido" /> </owl:Restriction> </rdfs:subClassOf> </owl:Class>

7. CONCLUSIONS

To realize a semantic Web, tools are required that allow users with little technical background to generate their own ontologies and collaborate in the construction of distributed knowledge bases. The work presented here is a contribution to the creation of these tools: a method to formally obtain ontologies codified in the OWL-DL language from an informal knowledge representation, such as concept maps. In this paper, we combine the semantic analysis of some predefined *linking-words* (continuously automatically enriched), mechanisms of natural language processing based on a concept-sense-disambiguation

algorithm, and two knowledge bases (WordNet and a concept maps repository) in a novel way, which significantly distinguishes from the existing literature. The mapping between concept maps and OWL ontologies creates the bases for the collaborative development of ontologies in a more intuitive, friendlier manner for humans. In its current state, the method presented advances the state of the art through the use of tools and techniques from natural language processing, such as the integration of WordNet and the definition of a concept sense disambiguation algorithm. This, combined with the topological analysis of concept maps, allows maintaining a greater flexibility and more independence during concept map construction. These aspects are important for less-expert users in ontology construction, and are not considered by Gómez et al. [2004] and Hayes et al. [2005]. These aspects, combined with the increase in the formalizations level of the linking-words and the use of external knowledge representations, allow to augment the semantic inference in concept maps and to obtain more expressiveness in the resulting OWL than the ones reported by Gómez et al. [2004] and Brilhante et al. [2006]. The automatic enrichment of the linkingwords repository is another contribution of the work presented in this paper. The method presented is mainly applicable to shallow domains, due to the fact that the terms in WordNet are about general knowledge; and this is an important restriction to be taken into account. Finally, the use of domain ontologies as alternative knowledge sources (to be used in a way similar to WordNet) and the use of concept-map repositories will increase the applicability of the method proposed; we are considering both issues as future work.

REFERENCES

- Baader, F., and W. Nutt, Basic Description Logics, The Description Logic Handbook: Theory, Implementation, and Applications, Cambridge University Press, 47-100, 2003.
- Brilhante, V., G. Macedo, and S. Macedo, Heuristic Transformation of Well-Constructed Conceptual Maps into OWL Preliminary Domain Ontologies, paper presented at Workshop on Ontologies and their Applications, WONTO'2006, Brazil, 2006.
- Gómez, H., B. Díaz, and A. González, Two layered approach to knowledge representation using conceptual maps description logic, paper presented at the first International Conference on Concept Mapping, Pamplona, Spain, 2004.
- Hayes, P. J., T. C. Eskridge, R. Saavedra, T. Reichherzer, M. Mehrotra, and D. Bobrovnikoff, Collaborative knowledge capture in ontologies, paper presented at the third International Conference on Knowledge Capture, Canada, 99-106, 2005.
- Jones, G., A. Robertson, J. Forbes, and G. Hollier, The HarperCollins Dictionary of Environmental Science, HarperCollins, 453 pp., 1992.
- Miller, G. A., R. Beckwidth, C. Fellbaum, D. Gross, and K. J. Miller, Introduction to WordNet: An On-line Lexical Database, *International Journal of Lexicography*, 3(4), 235-244, 1990.
- Novak, J. D., and D. B. Gowin, Learning how to learn, Cambridge University Press, 1984.
- Simón, A., V. Estrada, A. Rosete, and V. Lara, GECOSOFT: Un Entrono Colaborativo para la Gestión del Conocimiento con Mapas Conceptuales, paper presented at the second International Conference on Concept Mapping, Vol. 2, San José, Costa Rica, 114-117, 2006.
- Simón, A., L. Ceccaroni, and A. Rosete, Concept sense disambiguation in concept maps using WordNet, Technical Report LSI-07-44-R, Departament de Llenguatges i Sistemes Informàtics (LSI), Universitat Politècnica de Catalunya (UPC), Campus Nord, Edif. Omega, c. Jordi Girona, 1-3, 08034 Barcelona, Spain, 2007a.
- Simón, A., L. Ceccaroni, and A. Rosete, Generation of OWL Ontologies from Concept Maps in Shallow Domains, LNAI, Vol. 4788, Springer-Verlag, 259-267, 2007b.
- Smith, M., Ch. Welty, and D. McGuinness, OWL Web Ontology Language Guide, W3C, 2004.