A Note on Methodology for Designing Ontology Management Systems

Francesco Colace and **Massimo De Santo** and **Paolo Napoletano**
DIIIE, University of Salerno
via Ponte don Melillo, 1
84084 Fisciano (SA), Italy

Abstract

In this note we propose a novel methodology for designing Ontology Management Systems architecture, which grounds on an ontology representation based on probabilistic Graphical Models.
By discussing about troubles with ontology as tool for managing knowledge, formal assumptions about semantics definition and representations rise, turning out an original architecture that will be presented and discussed.
In the paper we will further detail and discuss the OMS proposal, focusing on how methodologies and representation can fruitfully rely upon probabilistic GM.

Introduction

The Semantic Web (Berners-Lee, Hendler, & Lassila 2001) and Knowledge Engineering communities are both confronted with the endeavor to design and to build ontologies by means of different tools and languages, which in turn raises an "ontology management problem" related to the peculiar tasks of representing, maintaining, merging, mapping, versioning and translating. As a consequence, the search for a uniform framework to cope with such issues, in other terms, for an Ontology Management System (OMS), has become a central tenet of this research realm.

Despite the crucial importance of the problem, actual proposals hardly satisfy the compelling requirements posed by OMS (Gomez-Perez, Corcho-Garcia, & Fernandez-Lopez 2003). For instance, no existing OMS can cope with different ontology languages through a suitable parser, and meanwhile offer a uniform ontology graphical representation to exploit machine learning algorithms, while exploiting suitable interfaces for ontology validation and definition.

These mentioned above are well known concerns an rather technological, animating the debate in the ontology field. However in our opinion the utilization of different tools and languages is caused by a personal view of the problem of knowledge representation, which in turn raises a not uniform perspective.

Most important each ontology scientist may rely, deliberately or implicitly, on a different definition of the role of ontology as mean for semantics representation (Santini 2007).

Therefore we argue that a special effort should be devoted to better explain and clarify the theory of semantics knowledge and how we can correctly model the latter for being properly represented and used on a machine.

The main and novel contribution of this note is that we address a methodology for designing an OMS architecture, by taking into account such broader picture. Further this methodology grounds on an ontology representation based on probabilistic Graphical Models (GM) (Bishop 2006).

In section 1 we derive, a formal theory of semantic representation specifying what has to be computed. Once a computational model is available a formal language is needed for describing such knowledge, in section 2 a representation based on GM is provided. Further, in section 3, relying such computational model, architecture issues will be addressed, and finally in section 4 a probabilistic model for ontology building is presented and discussed through a case of study. Conclusions and future works are presented and discussed.

In search of semantics: troubles with ontology

Semantic representation is one of the main (never-ending) debates in cognitive psychology. In this section we highlight key issues of such debate pointing out the most important troubles with ontology as a way for representing semantic knowledge, we start from explaining what today is commonly acknowledged, among researchers and practitioners of information systems, as "ontology", and consequently, introducing a formal and complete theory of semantic representation.

In the field of computer science, the word ontology is used with two different connotations. The first one considers such word as a discipline, namely the discipline that studies conceptions of reality and the nature of being. The second one considers ontology for indicating artifacts that the discipline produces, in other terms as a name for such artifacts. Note that in the first case the word ontology is proper, and consistent with its meaning in philosophy:

[...] the study of Being as such, i.e. of the basic characteristics of all reality,

according to the Encyclopaedia Britannica.

However, in the second case the word is clearly improper (see (Fensel *et al.* 2001) for example), and as suggested by Santini (Santini 2007), a better name in this case would be "ontonomy":

> *[...] an ontonomy is, roughly, a set of terms V, a collection of relations over this set, and a collection of propositions (axioms) in some decidable logical system.*

Once the right meaning of the word ontology has been chosen, a question still goes unanswered: is such way of considering ontology sufficient to save inferential role semantics?

One important assumption in ontology—and in the representational theories of mind—is that meaning exists independently of the language in which a text is written, and of the act of interpretation. By recalling a classic model of a communication channel we derive a general scheme of an ontology communication scheme (in facts language) (Santini 2007):

$$\textbf{meaning} \rightarrow \textbf{encode} \rightarrow \textbf{language} \rightarrow \textbf{decode} \rightarrow \textbf{meaning} \quad (1)$$

In this model, differently from classic scheme of communication, where the noise corrupts the channel, the placement of noise bears a quite different role. In order to understand what is corrupted by noise, we must address the very actors of such communication process. A communication act through language could be compared to the act of reading a book. In this case the previous scheme can be reshaped as:

$$\textbf{author} \rightarrow \textbf{language} \rightarrow \textbf{reader} \quad (2)$$

In this model, the origin of the communicative act is a meaning that resides wholly with the author, and that the author wants to express in a permanent text. This meaning is a-historical, immutable, and pre-linguistic and is encoded on the left-hand side of process, it must be wholly dependent on an act of the author, without the possibility of participation of the reader in an exchange that creates, rather than simply register, meaning. The author translates such creation into the shared code of language, then he sends, opening a communication, it to the reader. It is well known that, due to the accidental imperfections of human languages, contingent imperfections may occurs, and consequently such translation process may be imperfect, which in turn means that such a process is corrupted by "noise". In a perfect translation process (ontology acknowledges that this might be an unattainable theoretical limit) we have a perfect reproduction of the essential meaning as it appears in the mind of the author. Once the translated meaning is delivered to reader, a process for decoding it starts. Such process (maybe also corrupted by some more noise) obtains a reasonable approximation of the original meaning as intended by the author. Meaning is never fully present in a sign, but it is scattered through the whole chain of signifiers, it is deferred, through the process that Derrida (Derrida 1997) indicates with the neologism differance, it is a dynamic process that takes plane on the syntagmatic plane of the text (Eco 1979).

This model of meaning is necessary for the ontological enterprise because it is the only one that allows meaning to be assigned to a text, and recorded in a formal language other than the natural language, from which it can be extracted through automatic means following a scheme like this:

$$\textbf{mean.} \rightarrow \textbf{formula} \rightarrow \textbf{formal system} \rightarrow \textbf{algorithm} \rightarrow \textbf{mean.} \quad (3)$$

where "mean." stands for "meaning".

In such framework, ontology is a static entity, contains fixed relations between words, relations that hold independently of the specific situations in which the word is used. It contains, in other words, paradigmatic relations. Ontology needs meaning to be fully present in a word, be it through some characteristic of the word itself or through the relation of the word with other words.

But Santini asserts that this is not the way in which meaning is constructed, consequently claims that ontology should abandon any velleity of defining meaning, or of dealing with semantics, and re-define itself as a purely syntactic discipline, much like the rest of computing activities. In this framework, there is a lot that the discipline can offer to help users discover the semantics of texts. It is clear, however, that this cant be done with the normativism of attaching a meaning to a text: ontology should simply be an instrument to facilitate the interaction of a user with the data, keeping in mind that the users situated, contextual presence is indispensible for the creation of meaning. For instance, it would be a good idea to partially formalize the syntactic part of the interaction process that goes into the creation of meaning.

As a consequence Santini reaches a conclusion claiming that meaning is the limit point of a temporal, situated process, in which the text acts as a boundary condition and in which the user is, ex necessitate, the protagonist. Such a view is not completely new to computing science, having been explored, e.g. by emergent semantics (Santini, Gupta, & Jain 2001); (Grosky, Sreenath, & Fotouhi 2002), a view of semantics in which the computer is just a syntactic instrument to aid the readers own discovery of meaning. Whatever technical path ontology will take, it can continue claiming to be involved with semantics only if it will become an interactive syntactic instrument for the user. In this way the computer could become an instrument to enrich our own immersion in meaning, much like a book is, rather than a factor of impoverishment and banalization of the process of signification and of our intellectual life, as too often is the case.

In the light of this discussion in the following subsection we will propose a view that taking into account the above remarks.

A viable road to semantics

As pointed out by Griffiths (T. L. Griffiths 2007) the semantic knowledge can be thought of as knowledge about relations among several types of elements, including *words*, *concepts*, and *percepts*. According to such definition the following relations must be taken into account:

1. *Concept – concept* relations: Knowledge that dogs are a kind of animal, that dogs have tails and can bark, or that animals have bodies and can move.
2. *Concept – action* relations: Knowledge about how to pet a dog or operate a toaster.
3. *Concept – percept* : Knowledge about what dogs look like, how a dog can be distinguished from a cat
4. *Word – concept* relations: Knowledge that the word dog refers to the concept dog, the word animal refers to the concept animal, or the word toaster refers to the concept toaster.
5. *Word – word* relations: Knowledge that the word dog tends to be associated with or co-occur with words such as tail, bone,

Obviously these different aspects of semantic knowledge are not necessarily independent rather those can influence behavior in different ways and seem to be best captured by different kinds of formal representations. As a consequence result, different approaches to modeling semantic knowledge tend to focus on different aspects of this knowledge, specifically we can distinguish two main traditions:

I One which has focused more on the structure of associative relations between words in natural language use and relations between words and concepts, along with the contextual dependence of these relations (Ericsson & Kintsch 1995; Kintsch 1988; Potter 1993). Such tradition is related to points 4 and 5, which can be defined as *light semantics*;

II One which emphasizes abstract conceptual structure, focusing on relations among concepts and relations between concepts and percepts or actions (Collins & Quillian 1969). Such tradition is related to points 1, 2 and 3, which can be defined as *deep semantics*.

Following this discussion one could decide that semantics representation could emerge through the interaction of both *light* and *deep semantics*. Thus, an an artificial system contending with semantics should necessary take into account both facets.

Light and *Deep semantics* as a computational problem

In order to provide a systematic account of deep and light semantics in a computational framework, while keeping in mind that such problems genuinely originate in the broader framework of cognitive science, we can resort to Marr's approach. Briefly, Marr (Marr 1982) distinguished between three levels at which any agent carrying out a cognitive task could be understood, the *what/why* level (computational theory), the *how* level (algorithm) and the *physical realization* level (implementation). In particular the first two levels are of interest here, the computational theory defining what is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out, and the representation and algorithm level accounting for how can the computational theory be implemented, what is the representation for the input and output, and what is the algorithm for the transformation.

More important for us, it has been recently argued (Chater, Tenenbaum, & Yuille 2006; Knill, Kersten, & Yuille 1996) that Marr's three-fold hierarchy could be reorganized into two levels: the computational theory level, which can be precisely formalized in terms of Bayesian theory, and the implementation theory level, embedding both Marr's algorithmic and physical realization levels. More precisely, it has been shown (Boccignone & Cordeschi 2007) that a formal statement of Marr's computational theory level can be given in terms of a theoretical model $\mathcal{M} = \langle P(\{X_k\}_{k=1}^{K}), \mathcal{GM}\rangle$, $P(\{X_k\}_{k=1}^{K})$ denoting a joint probabilistic distribution of random variables $\{X_k\}_{k=1}^{K}$ and $\mathcal{GM}$ a graphical model specifying the conditional dependencies among random variables, in other words what Marr called the *constraints* of the computational problem. Indeed, this novel view of Marr's proposal can be fruitfully exploited as we will see in the remainder of this note.

Light semantics As described above such semantics emphasize relatively light representations that can be learned from large text corpora, and on explaining the structure of word–word and word–concept associations, rooted in the contexts of actual language use.

Although the interpretation of sentences requires semantic knowledge that goes beyond these contextual associative relationships, many theories confirm the fact that, though the interpretation of sentences requires semantic knowledge that goes beyond these contextual associative relationship, it still identify this level of knowledge as playing an important role in the early stages of language processing (Ericsson & Kintsch 1995; Kintsch 1988; Potter 1993)

Following Marr' theory of computation (Marr 1982) we obtain the light semantics and then the building of an ontology that can be called static ontology, can be obtained as:

a Word patching: define relations among words;

b Prediction: predict the next word or concept, facilitating retrieval;

c Disambiguation: identify the senses or meanings of words;

d Gist extraction: pick out the gist of a set of words.

Here the adjective "static" stands for a property that describes the fact that the meaning can be extracted only from the text and without the help of user that could be introduce a sort of variability.

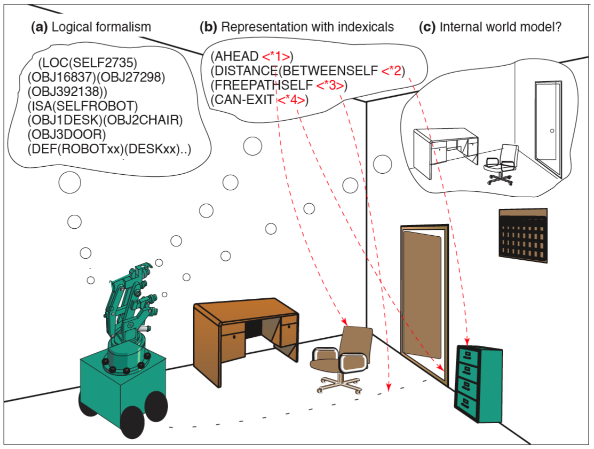


Figure 1: Three different ways in which a robot might represent its world: (a) Logical formalism, (b) Representation with indexicals, (c) internal world of model. (Pylyshyn 2000).

Once a computational model is available a formal language could describe such knowledge that can be stored in a information system, as we will see in more details in next sections.

Deep semantics This knowledge is traditionally represented in terms of systems of abstract proposition (Collins & Quillian 1969). Models in this tradition have focused on explaining phenomena such as the development of conceptual hierarchies that support propositional knowledge, reaction time to verify conceptual propositions in normal adults, and the decay of propositional knowledge with aging or brain damage.

While *Concept – concept* relations could be modeled using the prototype theory that plays a central role in Linguistics, as part of the mapping from phonological structure to semantics (Gärdenfors 2004) and in this note we don't spend time in illustrating the way to do that, then the *Concept – action* relations can be revealed using the theory of *emergent semantics* pointed out by Santini and Grosky (Santini, Gupta, & Jain 2001; Grosky, Sreenath, & Fotouhi 2002).

In details, the semantics of a web page is defined by its content and context. Understanding of textual documents is beyond the capability of todays artificial intelligence techniques, and the many multimedia features of a web page make the extraction and representation of its semantics even more difficult. Modern search engines rely on keyword matching and link structure, but the semantic gap is still not bridged. Previous studies have shown that users surfing on the web exhibit coherent intentions (or browsing semantics) and that these intentions can be learned and used for the prefetching of related web pages (Ibrahim & Xu 2000).

In our approach, the semantics of a web page can be derived statistically through analyzing the browsing paths of users toward this page. For this reason, we also refer to these emergent semantics of a page as *dynamic semantics*.

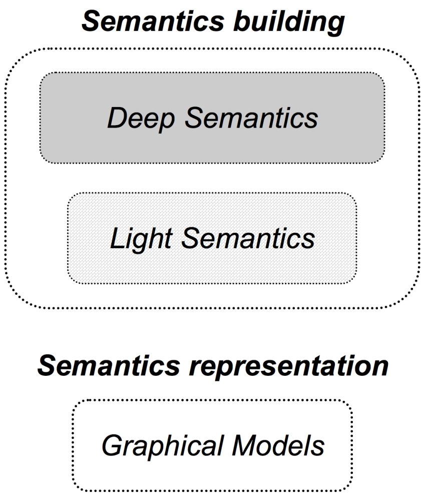


Figure 2: Semantics knowledge could conceived as the interaction of both *light* and *deep semantics*. The result of such interaction generates what we call computational theory of semantics (semantics building), which must be constrained through a GM (semantics representation).

Building semantics by using perception (vision, etc.), that is the modeling of *Concept – percept* relations, is yet a problem that can be understood by considering the Marr's computational theory (Marr 1982). Here could be investigated the mechanism describing how the human make use of perception (in broad sense) for encoding knowledge representation. In such perspective studies in the field of Computer Vision could be useful (Ballard & Brown 1982; Ballard 1997). One of the approach that seems to be suitable for our purpose is that proposed by Pylishyn (Pylyshyn 2000). He asserts a theory for situating vision in the world by differentiating three different ways in which an artificial agent (namely a robot) might represent its world in order to carry out actions in real world, Figure 1.

Ontology representation

Once a semantics computational theory has been delivered, defining a joint probabilistic distribution of random variables (in the sense of Boccignone (Boccignone & Cordeschi 2007)), we have to introduce the GM which specifies the conditional dependencies among random variables (the Marr's constraints).

It is our conviction that one of the major limitations of languages for representing ontologies - and in this respect OWL is no exception - stems from the static assignment of relations between concepts, e.g. "Man is a subclass of Human".

On the one hand, ontology languages in the semantic web, such as OWL and RDF, are based on crisp logic (Guarino 1998) and thus cannot handle incomplete, partial knowledge for any domain of interest (Antoniou & van Harmelen 2004). On the other hand, it has been shown (see, for instance (Ding, Peng, & Pan 2004)) how uncertainty exists

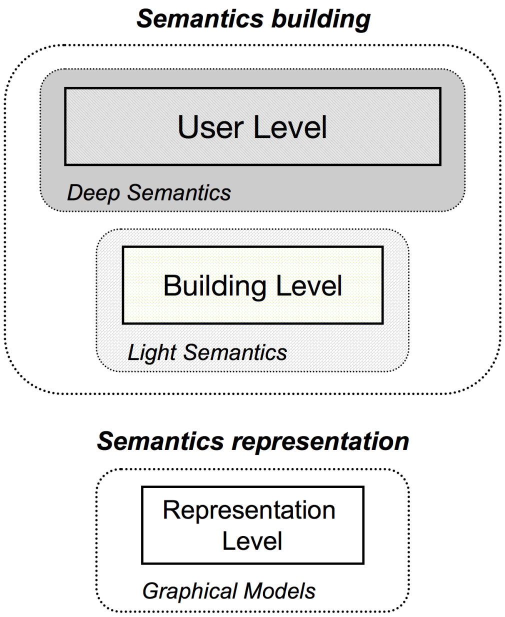


Figure 3: Semantics knowledge could conceived as the interaction of both light and deep semantics constrained through a GM. This corresponds, from an architectural standpoint, to identification of two levels for semantics building, *User* and *Building level*, and a level for semantic representation, called *Representation level*.

in almost every aspects of ontology engineering, and probabilistic directed GMs such as Bayesian Nets (BN) can provide a suitable tool for coping with uncertainty. Yet, in our view, the main drawback of BNs as a representation tool, is in the reliance on class/subclass relationships subsumed under the directed links of their structure. We argue that an ontology is not just the product of deliberate reflection on what the world is like, but is the realization of semantic interconnections among concepts, where each of them could belong to different domains.

Indeed, since the seminal and outstanding work by Anderson on probabilistic foundations of memory and categorization, concepts/classes and relations among concepts arise in terms of their prediction capabilities with respect to a given context (Anderson 1991). Further, the availability of a category grants the individual the ability to recall patterns of behavior (stereotypes, (Roland G. Fryer & Jackson 2003)) as built on past interactions with objects in a given category. In these terms, an object is not simply a physical object but a view of an interaction.

Thus, even without entering the fierce dispute whether ontologies should or should not be shaped in terms of categories (Eco 1997), it is clear that to endow ontologies with predictive capabilities together with properties of re-configurability, what we name *ontology plasticity*, one should relax constraints on the GM structure and allow the use of cyclic graphs. A further advantage of an effort in this direction is the availability of a large number of conceptual and algorithmic tools that have been produced by the Machine Learning community in most recent years (Bishop 2006). For instance, one could model ontology evolution in time as a Dynamic Bayesian Network (Bishop 2006).

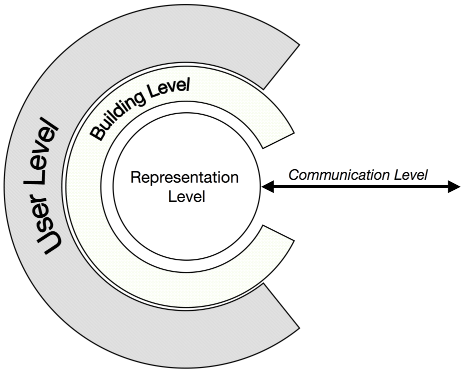


Figure 4: OMS Core Architecture. The Representation level could communicate to everyone knows its language, for instance other levels of our architecture or other OMS.

What we propose here is to use both the tradition/level of ontology for building semantic knowledge and such representation stage for its representation, as illustrated in Figure 2. In order to do that we introduce the notion of architecture for capturing those different levels discussed above.

Putting things together: architectural issues in designing OMS

Recalling that the aim of this work is to provide a methodology for designing OMS as a sound basis to address architectural issues. Before introducing our proposal we point out the reason why designing OMS could be useful.

Since ontology developers were engaging in building ontologies (to be more precise "ontonomies", in the vein of Santini) they have been committed to different tools and different languages, inevitably causing an "ontology management problem": representing, maintaining, merging, mapping, versioning, translating, etc. As a consequence, a uniform framework to jointly maintain and manage ontology, in a word an OMS, is required. Although IT community is deeply involved in providing a unified account of such systems, no result currently satisfies jointly all the above requirements (Gomez-Perez, Corcho-Garcia, & Fernandez-Lopez 2003; Noy & Musen 2004). For instance, none existing OMS can jointly manage different ontology language providing a suitable parser, provide a uniform ontology graphical representation for machine learning algorithms and allow a kind of human aid for both ontology building and validation.

An Ontology Management System (OMS) can be conceived as a uniform framework that helps users in managing multiple ontologies by leveraging data and algorithms developed for one tool in another. "Uniform" means that we propose a system based on a specific language for

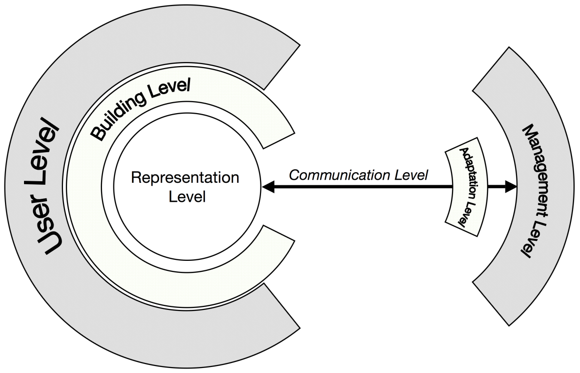


Figure 5: Complete OMS Architecture. The Adaptation and Management levels are "plug in" components.

semantics representation that could be shared by all the OMSs. Moreover, this framework allows users to validate the current ontologies, or to load ontologies from files or via the Internet and to locally create, to modify, to query, and to store ontologies.

Starting from the previously discussed assumptions about existing ways for representing semantics (and consequently for representing ontologies), the first steps are strictly connected to concepts there introduced.

The semantics knowledge can be properly described by putting in connection *deep* and *light semantics*.

While light semantics can be throughly instantiated, from an architectural point of view, in an artificial agent, deep semantics must necessary involve a human agent in the building group. This corresponds, from an architectural standpoint, to identification of two levels: *building level* (for the artificial agent) and *user level* (for the human one), Figure 3.

As previously discussed in order to design consistent semantic relations a unique representation of ontologies can be shaped in the form of probabilistic Graphical Model. This is the core business of our proposal, which we name the "Representation Level". An illustration of the three levels is in Figure 3 and an illustration of their relations is Figure 4.

Note that through the communication level the OMS could establish connection to others helpful levels (shortly we discuss them) or directly to others homogeneous OMS.

Furthermore, our core architecture proposal, for satisfying the previously accounted requirements, needs to be provided of other helpful levels. Specifically we introduce an adaptation level which acts as a language parser and a management level, which, interacting directly with the Representation level allows the user to handling ontologies (versioning, merging, etc.). In Figure 5 we illustrate the complete proposal where the last levels has been introduced as "plug in" components.

In Figure 6 we represent how the endeavor to design a uniform framework for managing ontology could be realized for a network of OMS which may also be seen as a sort of social semantics.

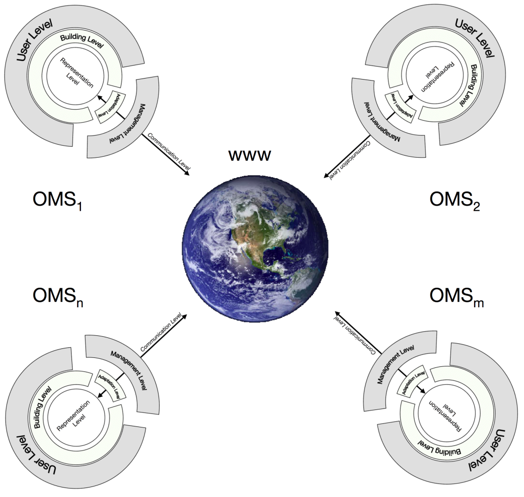


Figure 6: OMSs Networks: a distributed representation of social semantics.

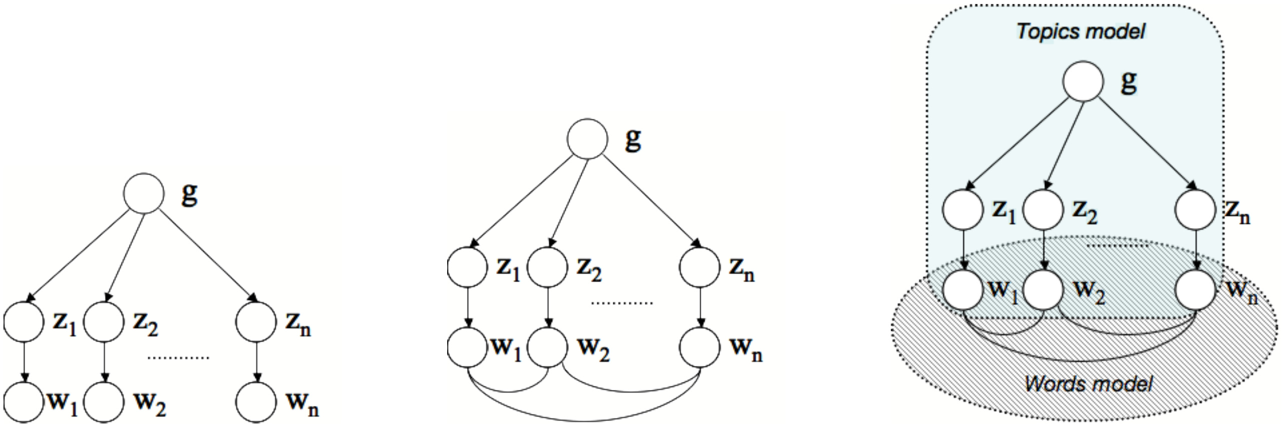
Summing up:

1. *Human Ontology Interaction (HOI): human-in-the-loop aid and validation*. Human aid is useful in order to build knowledge representation based tools (Deep semantics). We propose the "User Level " for both accomplishing ontology definition and validation.
2. *Ontology Adaptation: unifying languages for the ontology*. Here, the main idea is to set up an "Adaptation Level" as a parser for converting different ontologies into one, by using a suitable (W3C) language, e.g. the Ontology Web Language (OWL).
3. *Ontology representation: designing consistent semantic relations between words*. It is the core business of our proposal, which we name the Representation Level. Here a unique representation of ontologies based on probabilistic Graphical Model is provided.
4. *Ontology building: identifying, defining and entering concept definition*. At this level the GM representation can be fully exploited for providing a "Building Level" relying on machine learning techniques.
5. *Ontology management: versioning, merging and mapping*. This "Management Level" deals with general management of the ontology. Some basic ontology inference techniques have to be embedded here in order to perform consistency checking, versioning, merging and mapping management.

In the following section we will address the problem of ontology building, then a case study of building a wine ontology in the light of this framework is considered.

Ontology building in a probabilistic framework

The description of both *Word – Word* and *Word – Concept* relations is based on an extension of the computational



(a)

(b)

(c)

Figure 7: Graphical Models of *light semantics*. 7(a): Griffiths' model (T. L. Griffiths 2007) relying on Latent Dirichlet allocation (Blei, Ng, & Jordan 2003). Such GM don't allow relations among words by assuming statistical independence among variables. 7(b): Our GM proposal. The statistical dependence among words is modeled by suitable connections. 7(c): Our GM proposal. As a consequence of statistical dependence assumption, our model is the union of Topics model (Blei, Ng, & Jordan 2003) and Words model.

model depicted above and discussed in (T. L. Griffiths 2007), where statistic dependence among words is assumed. As previously discussed, 4 problems have to be solved: *word patching*, *prediction*, *disambiguation* and *gist extraction*. The original theory of Griffiths mainly asserts a semantic representation in which word meanings are represented in terms of a set of probabilistic topics resulting in the GM reported in 7(a), where the assumption of statistically independence among words was made. On the contrary, our extension provide word–word relations, which are represented as a set of probabilistic connections, as a result we can draw the GM of Figure 7(b). Summing up, we propose a probabilistic model that, together with the *topics model* (which models Word–Concept, (T. L. Griffiths 2007)), considers what we call the *words model*, in order to performs well in predicting word association and the effects of semantic association and ambiguity on a variety of language-processing and memory tasks, Figure 7(c).

Assume we have seen a sequence of words $\mathbf{w} = (w_1, \ldots, w_n)$. These n words manifest some latent semantic structure l. We will assume that l consists of the gist of that sequence of words g and the sense or meaning of each word, $\mathbf{z} = (z_1, \ldots, z_n)$, so $l = (\mathbf{z}, \mathbf{g})$. We can now formalize the four problems identified in the previous section:

- Word patching: Compute (w_i, w_j) from $\mathbf{w}$.
- Prediction: Predict $w_{n+1}$from $\mathbf{w}$.
- Disambiguation: Infer $\mathbf{z}$ from $\mathbf{w}$.
- Gist extraction: Infer $\mathbf{g}$ from $\mathbf{w}$.

Each of these problems can be formulated as a statistical problem. In this model, latent structure generates an observed sequence of words $\mathbf{w} = (w_1, \ldots, w_n)$. This relationship is illustrated using graphical model notation (Pearl 1988; Jordan 1998; Bishop 2006). Graphical models provide an efficient and intuitive method of illustrating structured probability distributions. In a graphical model, a distribution is associated with a graph in which nodes are random variables and edges indicate dependence. Unlike artificial neural networks, in which a node typically indicates a single unidimensional variable, the variables associated with nodes can be arbitrarily complex. The graphical model shown in Figure 7(a) is a directed graphical model, with arrows indicating the direction of the relationship among the variables. The graphical model shown in the figure indicates that words are generated by first sampling a latent structure, l, from a distribution over latent structures, $P(l)$, and then sampling a sequence of words, $\mathbf{w}$, conditioned on that structure from a distribution $P(\mathbf{w}|l)$. The process of choosing each variable from a distribution conditioned on its parents defines a joint distribution over observed data and latent structures. In the generative model shown in Figure 7(a), this joint distribution is $P(\mathbf{w}, l) = P(\mathbf{w}|l)P(l)$. With an appropriate choice of l, this joint distribution can be used to solve the problems of word patching, prediction, disambiguation, and gist extraction identified above. In particular, the probability of the latent structure l given the sequence of words $\mathbf{w}$ can be computed by applying Bayes's rule:

$$P(l|\mathbf{w}) = \frac{P(\mathbf{w}|l)P(l)}{P(\mathbf{w})} \tag{4}$$

where

$$P(\mathbf{w}) = \sum_l P(\mathbf{w}, l)P(l) \tag{5}$$

This Bayesian inference involves computing a probability that goes against the direction of the arrows in the graphical model, inverting the generative process.

Equation 5 provides the foundation for solving the problems of word patching, prediction, disambiguation, and gist extraction.

Summing up:

- Word patching

$$P(w_i, w_j) = \sum_{\mathbf{w}-(w_i,w_j)} \sum_{l} P(\mathbf{w}, l)P(l) \tag{6}$$

- Prediction

$$P(w_{n+1}, \mathbf{w}) = \sum_{l} P(w_{n+1}|l, \mathbf{w})P(l|\mathbf{w}) \tag{7}$$

- Disambiguation

$$P(\mathbf{z}|\mathbf{w}) = \sum_{g} P(l|\mathbf{w}) \tag{8}$$

- Gist extraction

$$P(\mathbf{g}|\mathbf{w}) = \sum_{z} P(l|\mathbf{w}) \tag{9}$$

A multidocument corpus can be expressed as a vector of words $\mathbf{w} = (w_1, \ldots, w_n)$, where each w_i belongs to some document d_i, as in a word–document co–occurrence matrix, cfr. Figure 8. We will use a generative model introduced by Blei et al. (Blei, Ng, & Jordan 2003) called latent Dirichlet allocation. In this model, the multinomial distribution representing the gist is drawn from a Dirichlet distribution, a standard probability distribution over multinomials (e.g., (Gelman *et al.* 1995)). In through the *words model* we can build consistent relations between words measuring their degree of dependence, formally by computing mutual information:

$$I(w_i, w_j) = \sum_{W_i} \sum_{W_j} P(w_i, w_j) \log \left(\frac{P(w_i, w_j)}{P(w_i)P(w_j)} \right) \tag{10}$$

Such measure establishes how much two variables (words) are statistically dependent, in facts the hardness of such statistical dependence increases as mutual information measure increases. By selecting hard connections among existing all, for instance choosing a threshold for the mutual information measure, a GM for the words can be delivered, (cfr. Figure 9).

Building Wine Ontology: a case of study of light semantic

Here we present a case of study of light semantics representation. Once topic is chosen, the words connections, namely *words model*, are learned from large text corpora, and consequently a Graphical model representing wine ontology is builded.

The multidocument corpus, extracted from a web repository, is represented in Figure 8 through the word–document co–occurrence matrix, where the black color indicates the highest word's frequency, and white indicates zero. The number of documents are 50 and the chosen topic is "wine", which in italian language is "vino".

As a result, we show the GM representing light semantics relations for the "vino" ("wine") topic. Here the threshold for connections selection is set to $0,06$.

Conclusions and future works

The main and novel contribution of this note is that we address a methodology for designing an OMS architecture, by taking into account a broader picture of the animated debate about ontology as a way for semantic representation. The discussed ontology facets have allowed to propose a formal computational theory of semantics, which in turn has inspired the designing of an original architectural of systems for managing knowledge, namely OMS.

As a result, the semantic representation could emerge through the interaction of two aspects which we discussed above and which we called: *light* and *deep semantics*.

Once a semantics computational theory has been delivered, which has defined a joint probabilistic distribution of random variables, we introduced the GM which specifies the conditional dependencies among random variables. Finally we focused on how building ontology in a probabilistic framework, by providing a probabilistic model relying on an extension of Griffiths' theory of topics in semantic representation, in which the words are assumed to be statistically dependent. The proposed model of semantics representation is experienced on case of study: "wine" ontology building. The produced GM represents, recalling previously definition, a static ontology, therefore it contains fixed relations between words, relations that hold independently of the specific situations in which the word is used, in other terms the meaning is extracted only from the text and without the help of user that could be introduce a sort of variability.

As future work we propose of providing the described system of a model for computing what we called *deep semantics*, which would introduce a sort of dynamism in building ontology.

Acknowledgements

The authors wish to thank Prof. Giuseppe Boccignone for his contribution to the ideas contained in this work.

References

Anderson, J. R. 1991. The adaptive nature of human categorization. *Psychological Review* 98(3):409–429.

Antoniou, G., and van Harmelen, F. 2004. *A semantic web primer*. Cambridge:MIT Press.

Ballard, D., and Brown, C. 1982. *Computer Vision*. New York, N.Y.: Prentice Hall.

Ballard, D. 1997. *An Introduction to Natural Computation*. Cambridge, MA: The MIT Press.

Berners-Lee, T.; Hendler, J.; and Lassila, O. 2001. The semantic web. *Scientific American* May.

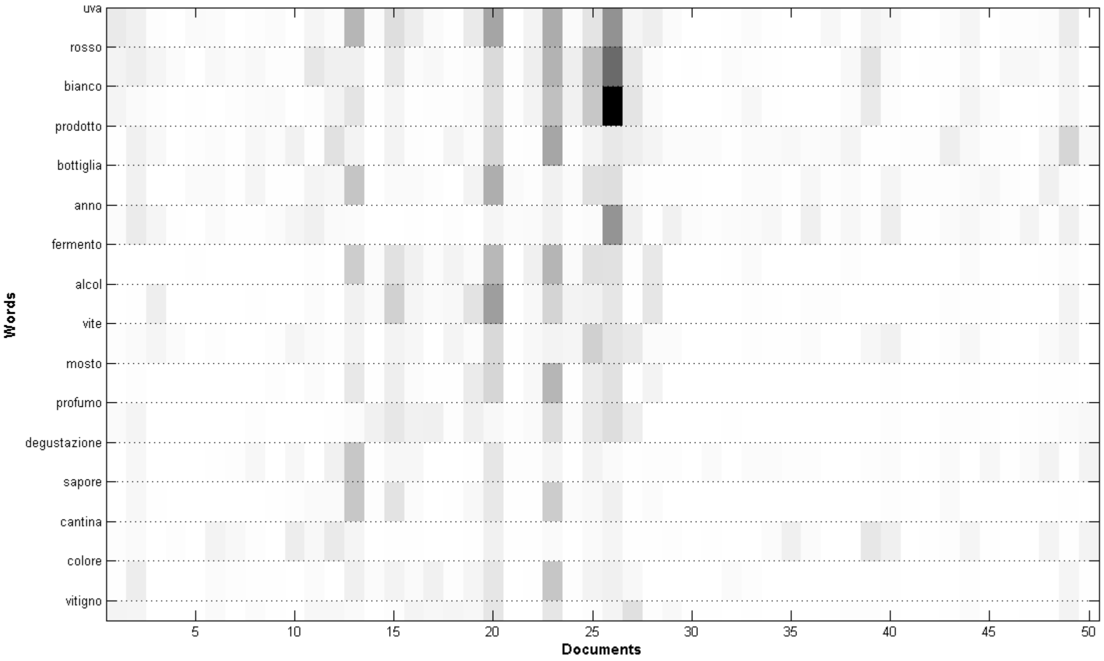


Figure 8: A word–document co-occurrence matrix for "vino" topic ("wine"), indicating the frequencies of 16 words across 50 documents extracted from a Web repository. A total of 41 documents use the word "uva", 37 use the word "bianco", and 16 use the word "alcol". Each row corresponds to a word in the vocabulary, and each column corresponds to a document in the corpus. Grayscale indicates the frequency with which the 4223 tokens of those words appeared in the 50 documents, with black being the highest frequency and white being zero. In the following the translation of each word: "uva=grapes", "rosso=red", "bianco=white","prodotto=product", "bottiglia=bottle", "anno=year", "fermento=ferment", "alcol=alcohol", "vite= grapes' tree", "mosto=must", "profumo=fragrance", "degustazione=tasting", "sapore=flavour", "cantina=cellar", "colore=colour", "vitigno=tendrill".

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3(993–1022).

Boccignone, G., and Cordeschi, R. 2007. Bayesian models and simulations in cognitive science. In *Models and Simulations 2*. Tilburg, NL: PhilSci Archive.

Chater, N.; Tenenbaum, J.; and Yuille, A. 2006. Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences* 10(7):287–291.

Collins, A. M., and Quillian, M. R. 1969. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior* (8):240–247.

Derrida, J. 1997. De la grammatologie. *Paris:Minuit*.

Ding, Z.; Peng, Y.; and Pan, R. 2004. A bayesian approach to uncertainty modeling in owl ontology. In *Proceedings of the International Conference on Advances in Intelligent Systems - Theory and Applications*.

Eco, U. 1979. A theory of semiotics. *Bloomington:Undiana University Press.*

Eco, U. 1997. *Kant and the Platypus: Essays on Language and Cognition*. First Harvest edition.

Ericsson, K. A., and Kintsch, W. 1995. Long-term working memory. *Psychological Review.* 102:211–245.

Fensel, D.; van Harmelen, F.; Horrocks, I.; McGuinness, D.; and Patel-Schneider, P. 2001. Oil: an ontology infrastructure for the semantic web. *IEEE intelligent systems*.

Gärdenfors, P. 2004. *Conceptual Spaces: The Geometry of Thought.* MIT Press.

Gelman, A.; Carlin, J. B.; Stern, H. S.; and Rubin, D. B. 1995. *Bayesian data analysis*. New York: Chapman & Hall.

Gomez-Perez, A.; Corcho-Garcia, O.; and Fernandez-Lopez, M. 2003. *Ontological Engineering*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Grosky, W. I.; Sreenath, D. V.; and Fotouhi, F. 2002. Emergent semantics and the multimedia semantic web. In *SIGMOD Record*, volume 31, 54–58.

Guarino, N. 1998. Formal ontology and information systems. In Press., A., ed., *In Proceedings of FOIS 98, Trento, Italy*, 3–15.

Ibrahim, T. I., and Xu, C.-Z. 2000. Neural net based prefetching to tolerate www latency. In *20th International Conference on Distributed Computing Systems*.

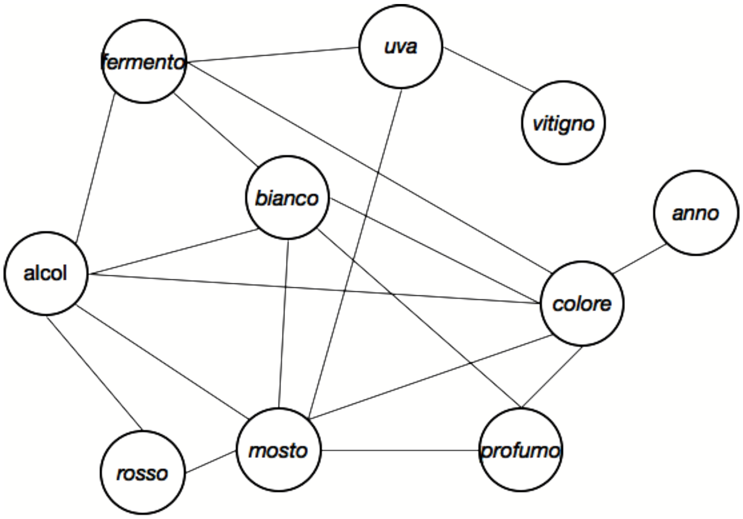


Figure 9: Graphical model representing wine ontology.

Jordan, M. I. 1998. *Learning in graphical models*. Cambridge, MA: MIT Press.

Kintsch, W. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review* 95:163–182.

Knill, D.; Kersten, D.; and Yuille, A. 1996. Introduction: A bayesian formulation of visual perception. In Knill, D., and Richards, W., eds., *Perception as Bayesian Inference*, 1–21. Cambridge University Press.

Marr, D. 1982. *Vision*. S. Francisco,CA: Freeman.

Noy, N. F., and Musen, M. A. 2004. Ontology versioning in an ontology management framework. *IEEE Intelligent Systems* 19(4):6–13.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Potter, M. C. 1993. Very short term conceptual memory. *Memory & Cognition* (21):156–161.

Pylyshyn, Z. 2000. Situating vision in the world. *Trends in Cognitive Sciences* 4(5):197–207.

Roland G. Fryer, J., and Jackson, M. O. 2003. Categorical cognition: A psychological model of categories and identification in decision making. *Working Paper Series* National Bureau of Economic Research.

Santini, S.; Gupta, A.; and Jain, R. 2001. Emergent semantics through interaction in image databases. *IEEE Transactions on Knowledge and Data engineering* 13(3):337–51.

Santini, S. 2007. Summa contra ontologiam. *International journal on data semantics* submitted.

T. L. Griffiths, M. Steyvers, J. B. T. 2007. Topics in semantic representation. *Psychological Review* 114(2):211–244.