

PR-OWL: A Bayesian Ontology Language for the Semantic Web

Paulo Cesar G. da Costa, Kathryn B. Laskey
School of Information Technology and Engineering,
George Mason University
4400 University Drive
Fairfax, VA 22030-4444 USA
{pcosta, klaskey}@gmu.edu

Kenneth J. Laskey[#]
MITRE Corporation, M/S H305
7515 Colshire Drive
McLean, VA 22102-7508 USA
klaskey@mitre.org

Abstract. This paper addresses a major weakness of current technologies for the Semantic Web, namely the lack of a principled means to represent and reason about uncertainty. This not only hinders the realization of the original vision for the Semantic Web, but also creates a barrier to the development of new, powerful features for general knowledge applications that require proper treatment of uncertain phenomena. We propose to extend OWL, the ontology language recommended by the World Wide Web Consortium (W3C), to provide the ability to express probabilistic knowledge. The new language, PR-OWL, will allow legacy ontologies to interoperate with newly developed probabilistic ontologies. PR-OWL will move beyond the current limitations of deterministic classical logic to a full first-order probabilistic logic. By providing a principled means of modeling uncertainty in ontologies, PR-OWL will serve as a supporting tool for many applications that can benefit from probabilistic inference within an ontology language, thus representing an important step toward the W3C's vision for the Semantic Web.

1 A Deterministic View of a Probabilistic World

Uncertainty is ubiquitous. If the Semantic Web vision [1] is to be realized, a sound and principled means of representing and reasoning with uncertainty will be required. Existing Semantic Web technologies lack this capability. Our broad objective is to address this shortcoming by developing a Bayesian framework for probabilistic ontologies and plausible reasoning services. As an initial step toward our objective, we introduce PR-OWL, a probabilistic extension to the Web ontology language OWL.

Although our research is focused in the Semantic Web, we are tackling a problem that long predates the WWW: the quest for more efficient data exchange. Clearly, solving that problem requires precise semantics and flexible ways to convey information. While the WWW provides a new presentation medium, and technologies such as XML present new data exchange formats, neither addresses the semantics of data

[#] The author's affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions or viewpoints expressed by the author.

being exchanged. The SW is meant to fill this gap, and the realization of its goals will require major improvements in technologies for data exchange.

One of the main technical differences between the current World Wide Web and the Semantic Web is that while the first relies on syntactic-only protocols such as HTTP and HTML, the latter adds meta-data annotations as a means to convey shared, precisely defined terms. That is, semantic awareness is exploited to improve interoperability among Web resources. Semantic interoperability requires shared repositories of precisely defined concepts. Such repositories are called ontologies.

One can find many different definitions for the concept of ontology applied to information systems, each emphasizing a specific aspect its author judged most important. Our focus is on ontology's role as a structured form of knowledge representation. Thus, we define an ontology as an explicit, formal representation of knowledge about a domain of application. This includes: types of entities that exist in the domain, properties of those entities, relationships among entities, and processes and events that happen with those entities. In this definition, the term entity refers to any concept (real or fictitious, concrete or abstract) that can be described and reasoned about within the domain of application. Ontologies are used for the purpose of comprehensively describing knowledge about a domain in a structured and sharable way, ideally in a format that can be read and processed by a computer.

Semantically aware schemes must be able to represent and appropriately process semantic differences between syntactically identical terms (e.g., "Grape" as a fruit versus John Grape the person). This is not a trivial task. Semantic interoperability requires shared sources of precisely defined concepts, which is exactly where ontologies play a key role. Yet, a traditional ontology can at best list multiple possible senses for a word such as "Grape," with no ability to grade their relative plausibility in a given context. This is inadequate for an open world environment where incomplete information is the rule and plausible reasoning is required.

Current generation Semantic Web technology is based on classical logic, and is lacks adequate support for plausible reasoning. For example, OWL, a W3C Recommendation [2], has no built-in support for probabilistic information and reasoning. This is understandable, given that OWL is rooted in web language predecessors (i.e. XML, RDF) and traditional knowledge representation formalisms (e.g., Description Logics [3]). This historical background somewhat explains the lack of support for uncertainty in OWL. Nevertheless, it is a serious limitation for a language intended for environments where one cannot simply ignore incomplete information.

A similar historical progression occurred in Artificial Intelligence (AI). From its inception, AI has struggled with how to cope with incomplete information. Although probability theory was initially neglected due to tractability concerns, graphical probability languages changed things dramatically [4]. Probabilistic languages have evolved from propositional to full first-order expressivity (e.g., [5]), and have become the technology of choice for reasoning under uncertainty in an open world [6]. Clearly, the Semantic Web will pose similar uncertainty-related issues as those faced by AI. Thus, just as AI has moved from a deterministic paradigm to embrace probability, a similar path appears promising for ontology engineering.

This path is not yet being followed. The lack of support for representing and reasoning with uncertain, incomplete information seriously limits the ability of current Semantic Web technologies to meet the requirements of the Semantic Web. Our work

is an initial step toward changing this situation. We aim to establish a framework that enables full support for uncertainty in the field of ontology engineering and, as a consequence, for the Semantic Web. In order to achieve this goal, we choose to focus on extending OWL so it can represent uncertainty in a principled way.

2 Related Research

One of the main reasons why Semantic Web research is still focused on deterministic approaches has been the limited expressivity of traditional probabilistic languages. There is a current line of research focused on extending OWL so it can represent probabilistic information contained in a Bayesian Network (e.g. [7], [8]). The approach involves augmenting OWL semantics to allow probabilistic information to be represented via additional markups. The result would be a probabilistic annotated ontology that could then be translated to a Bayesian network (BN). Such a translation would be based on a set of translation rules that would rely on the probabilistic information attached to individual concepts and properties within the annotated ontology. BNs provide an elegant mathematical structure for modeling complex relationships among hypotheses while keeping a relatively simple visualization of these relationships. Yet, the limited attribute-value representation of BNs makes them unsuitable for problems requiring greater expressive power.

Another popular option for representing uncertainty in OWL has been to focus on OWL-DL, a decidable subset of OWL that is based on Description Logics [3]. Description Logics are a family of knowledge representation formalisms that represent the knowledge of an application domain (the “world”) 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 (the world description).

Description logics are highly effective and efficient for the classification and subsumption problems they were designed to address. However, their ability to represent and reason about other commonly occurring kinds of knowledge is limited. One restrictive aspect of DL languages is their limited ability to represent constraints on the instances that can participate in a relationship. As an example, suppose we want to express that for carnivore to be a threat to another carnivore in a specific type of situation it is mandatory that the two individuals of class Carnivore involved in the situation are not the same. Making sure the two carnivores are different in a specific situation is only possible in DL if we actually create/specify the tangible individuals involved in that situation. Indeed, stating that two “fillers” (i.e. the actual individuals of class Carnivore that will “fill the spaces” of concept carnivore in our statement) are not equal without specifying their respective values would require constructs such as *negation* and *equality role-value-maps*, which cannot be expressed in description logic. While equality role-value-maps provide useful means to specify structural properties of concepts, their inclusion makes the logic undecidable [9].

Although the above approaches are promising where applicable, a definitive solution for the Semantic Web requires a general-purpose formalism that gives ontology designers a range of options to balance tractability against expressiveness.

Pool and Aiken [10] developed an OWL-based interface for the relational probabilistic toolset Quiddity*Suite, developed by IET, Inc. Their constructs provide a very expressive method for representing uncertainty in OWL ontologies. Their work is similar in spirit to ours, but is specialized to the Quiddity*Suite toolset. We focus on the more general problem of enabling probabilistic ontologies for the SW. We employ Multi-Entity Bayesian Networks (MEBN) as our underlying logical basis, thus providing full first-order expressiveness.

3 Multi-Entity Bayesian Networks

The acknowledged standard for logically coherent reasoning under uncertainty is Bayesian probability theory. Bayesian theory provides a principled representation of uncertainty, a logic for combining prior knowledge with observations, and a learning theory for refining the ontology as evidence accrues. The logical basis for PR-OWL is MEBN logic [5], which combines Bayesian probability theory with classical First Order Logic. Probabilistic knowledge is expressed as a set of MEBN fragments (MFrag) organized into MEBN Theories. An MFrag is a knowledge structure that represents probabilistic knowledge about a collection of related hypotheses. Hypotheses in an MFrag may be *context* (must be satisfied for the probability definitions to apply), *input* (probabilities are defined in other MFrag), or *resident* (probabilities defined in the MFrag itself). An MFrag can be instantiated to create as many instances of the hypotheses as needed (e.g., an instance of the “Disease” hypothesis for each patient at a clinic). Instances of different MFrag may be combined to form complex probability models for specific situations. A MEBN theory is a collection of MFrag that satisfies consistency constraints ensuring the existence of a unique joint probability distribution over instances of the hypotheses in its MFrag.

MEBN inference begins when a query is posed to assess the degree of belief in a target random variable given a set of evidence random variables. We start with a generative MTheory, add a set of finding MFrag representing problem-specific information, and specify the target nodes for our query. The first step in MEBN inference is to construct a situation-specific Bayesian network (SSBN), which is a Bayesian network constructed by creating and combining instances of the MFrag in the generative MTheory. When each MFrag is instantiated, instances of its random variables are created to represent known background information, observed evidence, and queries of interest to the decision maker. If there are any random variables with undefined distributions, then the algorithm proceeds by instantiating their respective home MFrag. The process of retrieving and instantiating MFrag continues until there are no remaining random variables having either undefined distributions or unknown values. A SSBN may contain any number of instances of each MFrag, depending on the number of entities and their interrelationships. Next, a standard Bayesian network inference algorithm is applied. Finally, the answer to the query is obtained by inspecting the posterior probabilities of the target nodes.

MEBN logic overcomes the limitations of the attribute-value representation of standard BNs. To understand this limitation, consider a relational database in which some entries are uncertain. A BN can represent only probabilities for a single table,

and treats the rows of the table independently of each other. For example, in a medical system, the “Patient” table might include information such as age, smoking history, family history, and whether the patient has emphysema. A BN might represent the probability of emphysema as a function of smoking history, age, and family history. If a patient’s family history were unknown, the BN could estimate the probability of emphysema using probabilities for the family history. However, a BN cannot represent relational information such as the increase in the probability of emphysema for all siblings upon learning that one of their parents had emphysema. To incorporate this kind of knowledge in a coherent manner, we need to combine *relational* knowledge (e.g., siblings have the same family history) with attribute-value knowledge (e.g., family history of emphysema increases the likelihood of emphysema).

To draw generalizations about individuals related in various ways, we need first-order expressive power. Description logics are attractive because they provide limited first-order expressivity, yet certain reasoning problems such as classification and subsumption are decidable. Many researchers have worked to identify decidable classes of problems for which efficient probabilistic algorithms exist (e.g., Naïve Bayes classification, in which features are modeled as conditionally independent given an object’s class). The ontology language P-SHOQ(D) [11], based on description logics, falls into this class.

We have chosen to base PR-OWL on MEBN logic because of its expressiveness: MEBN can express a probability distribution over models of any finitely axiomatizable first-order theory. As a consequence, there are no guarantees that exact reasoning with a PR-OWL ontology will be efficient or even decidable. On the other hand, a future objective is to identify restricted sub-languages of PR-OWL specialized to classes of problems for which efficient exact or approximate reasoning algorithms exist. It is our view that a general-purpose language for the Semantic Web should be as expressive as possible, while providing a means for ontology engineers to stay within a tractable subset of the language when warranted by the application.

4 Probabilistic Ontologies

Before presenting our probabilistic ontology language, we begin by defining a probabilistic ontology. Intuitively, an ontology that has probabilities attached to some of its elements would qualify for this label, but such a limited definition is inadequate for our purposes. Merely adding probabilities to concepts does not guarantee interoperability with other ontologies that also carry probabilities. More is needed than syntax for including probabilities if we are to justify a new category of ontologies.

A *probabilistic ontology* is an explicit, formal knowledge representation that expresses knowledge about a domain of application. This includes: (i) Types of entities that exist in the domain; (ii) Properties of those entities; (iii) Relationships among entities; (iv) Processes and events that happen with those entities; (v) Statistical regularities that characterize the domain; (vi) Inconclusive, ambiguous, incomplete, unreliable, and dissonant knowledge related to entities of the domain; and (vii) Uncertainty about all the above forms of knowledge. In this definition, the term entity refers

to any concept (real or fictitious, concrete or abstract) that can be described and reasoned about within the domain.

Probabilistic Ontologies are used for the purpose of comprehensively describing knowledge about a domain and the uncertainty regarding that knowledge in a principled, structured and sharable way, ideally in a format that can be read and processed by a computer. They also expand the possibilities of standard ontologies by introducing the requirement of a proper representation of the statistical regularities and the uncertain evidence about entities in a domain of application.

5 PR-OWL

PR-OWL is an extension that enables OWL ontologies to represent complex Bayesian probabilistic models in a way that is flexible enough to be used by diverse Bayesian probabilistic tools based on different probabilistic technologies. That level of flexibility can only be achieved using the underlying semantics of first-order Bayesian logic, which is not a part of the standard OWL semantics and abstract syntax. Therefore, it seems clear that PR-OWL can only be realized via extending the semantics and abstract syntax of OWL. However, in order to make use of those extensions, it is necessary to develop new tools supporting the extended syntax and implied semantics of each extension. Such an effort would require commitment from diverse developers and workgroups, which falls outside our present scope.

Therefore, in this initial work our intention is to create an upper ontology to guide the development of probabilistic ontologies. Daconta *et al.* define an upper ontology as a set of integrated ontologies that characterizes a set of basic commonsense knowledge notions [12]. In this preliminary work on PR-OWL as an upper ontology, these basic commonsense notions are related to representing uncertainty in a principled way using OWL syntax. If PR-OWL were to become a W3C Recommendation, this collection of notions would be formally incorporated into the OWL language as a set of constructs that can be employed to build probabilistic ontologies.

The PR-OWL upper ontology for probabilistic systems consists of a set of classes, subclasses and properties that collectively form a framework for building probabilistic ontologies. The first step toward building a probabilistic ontology in compliance with our definition is to import into any OWL editor an OWL file containing the PR-OWL classes, subclasses, and properties.

From our definition, it is clear that nothing prevents a probabilistic ontology from being “partially probabilistic”. That is, a knowledge engineer can choose the concepts he/she wants to include in the “probabilistic part” of the ontology, while writing the other concepts in standard OWL. In this case, the “probabilistic part” refers to the concepts written using PR-OWL definitions and that collectively form a MEBN Theory. There is no need for all the concepts in a probabilistic ontology to be probabilistic, but at least some have to form a valid MEBN Theory. Of course, only the concepts that are part of the MEBN Theory will be subject to the advantages of the probabilistic ontology over a deterministic one.

The subtlety here is that legacy OWL ontologies can be upgraded to probabilistic ontologies only with respect to concepts for which the modeler wants to have uncer-

tainty represented in a principled manner, make plausible inferences from that uncertain evidence, or to learn its parameters from incoming data via Bayesian learning. While the first two are direct consequences of using a probabilistic knowledge representation, the latter is a specific advantage of the Bayesian paradigm, where learning falls into the same conceptual framework as knowledge representation.

The ability to perform probabilistic reasoning with incomplete or uncertain information conveyed through an ontology is a major advantage of PR-OWL. However, it should be noted that in some cases solving a probabilistic query might be intractable or even undecidable. In fact, providing the means to ensure decidability was the reason why the W3C defined three different version of the OWL language. While OWL Full is more expressive, it enables an ontology to represent knowledge that can lead to undecidable queries. OWL-DL imposes some restrictions to OWL in order to eliminate these cases. Similarly, restrictions of PR-OWL could be developed that limit expressivity to avoid undecidable queries or guarantee tractability. Possible restrictions to be considered for an eventual PR-OWL Lite include (i) constraining the language to classes of problems for which tractable exact or approximate algorithms exist; (ii) restrict the representation of the conditional probability tables (CPT) to express a tractable and expressive subset of first-order logic; and/or (iii) to employ a standard semantic web language syntax to represent the CPTs (e.g. RDF). As an initial step, we chose to focus on the most expressive version of PR-OWL, which does not have expressivity restrictions and provides the ability to represent CPTs in multiple formats.

An overview of the general concepts involved in the definition of a MEBN Theory in PR-OWL is depicted in Figure 1. In this diagram, the ovals represent general classes; and arrows represent major relationships between classes. A probabilistic ontology must have at least one individual of class MTheory, which is a label linking a group of MFrag that collectively form a valid MEBN Theory. In actual PR-OWL syntax, that link is expressed via the object property hasMFrag (which is the inverse of object property isMFragIn).

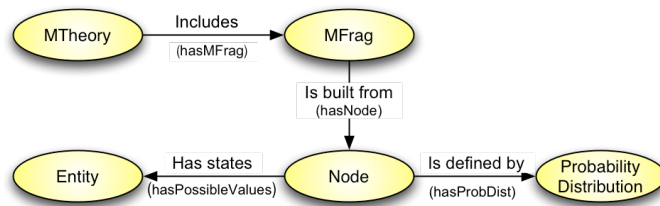


Fig. 1. Overview of a PR-OWL MEBN Theory Concepts

Individuals of class MFrag are comprised of nodes, which can be resident, input, or context nodes (not shown in the picture). Each individual of class Node is a random variable and thus has a mutually exclusive and collectively exhaustive set of possible states. In PR-OWL, the object property hasPossibleValues links each node with its possible states, which are individuals of class Entity. Finally, random variables (represented by the class Nodes in PR-OWL) have unconditional or conditional probabil-

ity distributions, which are represented by class Probability Distribution and linked to its respective nodes via the object property hasProbDist.

The scheme in Figure 1 is intended to present just a general view and thus fails to show many of the intricacies of an actual PR-OWL representation of a MEBN Theory. Figure 2 shows an expanded version conveying the main elements in Figure 1, their subclasses, the secondary elements that are needed for representing a MEBN Theory and the reified relationships that were necessary for expressing the complex structure of a Bayesian probabilistic model using OWL syntax.

Reification of relationships in PR-OWL is necessary because of the fact that properties in OWL are binary relations (i.e. link two individuals or an individual and a value), while many of the relations in a probabilistic model include more than one individual (i.e. N-ary relations). The use of reification for representing N-ary relations on the Semantic Web is covered by a working draft from the W3C's Semantic Web Best Practices Working Group [13].

Although the scheme in Figure 2 shows all the elements needed to represent a complete MEBN Theory, it is clear that any attempt at a complete description would render the diagram cluttered and incomprehensible. A complete account of the classes, properties and the code of PR-OWL that define an upper ontology for probabilistic systems is given in [14]. These definitions can be used to represent any MEBN Theory.

In its current stage, PR-OWL contains only the basic elements needed to represent any MEBN theory. Such a representation could be used by a Bayesian tool (acting as a probabilistic ontology reasoner) to perform inferences to answer queries and/or to learn from newly incoming evidence via Bayesian learning.

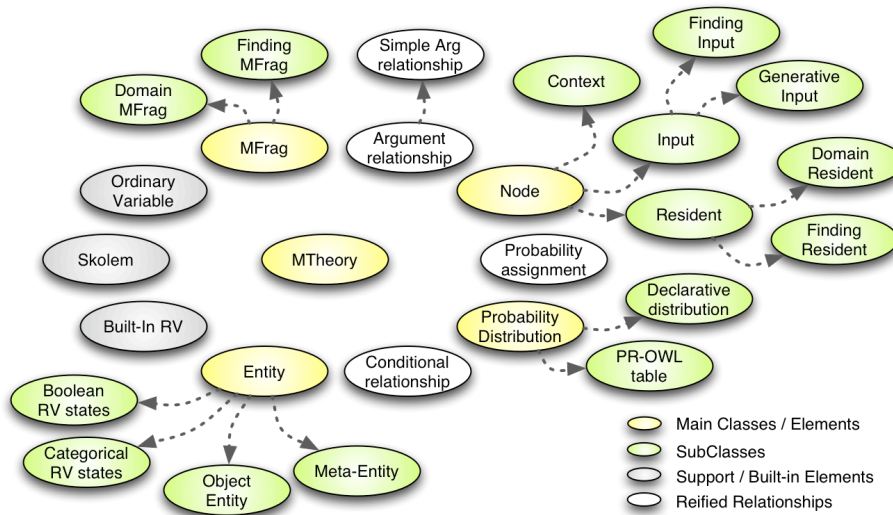


Fig. 2. Elements of a PR-OWL Probabilistic Ontology

However, building MFrags and all their elements in a probabilistic ontology is a manual, error prone, and tedious process. Avoiding errors or inconsistencies requires

very deep knowledge of the logic and of the data structure of PR-OWL. Without considering the future paths to be followed by research on PR-OWL (i.e. whether it will be kept as an upper ontology or transformed into an actual extension to the OWL language), the framework discussed here and in greater detail in [14] makes it already possible to facilitate probabilistic ontology usage and editing by developing plugins to current OWL editors. Figure 3 illustrates a plugin concept for the OWL Protégé editor (which is itself a Protégé plugin). The figure illustrates how graphical construction of an MFrag can be performed in a similar fashion to how a BN is constructed in one of the many graphical editors for BNs. In this proposed scheme, in order to build an MFrag a user would select the icon for the type of node he/she wants to create (e.g. resident, input, context, etc.), connect that node with its parents and children, and enter its basic characteristics (i.e. name, probability distribution, etc.) either by double-clicking on it or via another GUI-related facility. Such a plugin would hide from users the complex constructs required to convey the many details of a probabilistic ontology, providing a more intuitive and less error-prone means of constructing and maintaining probabilistic ontologies.

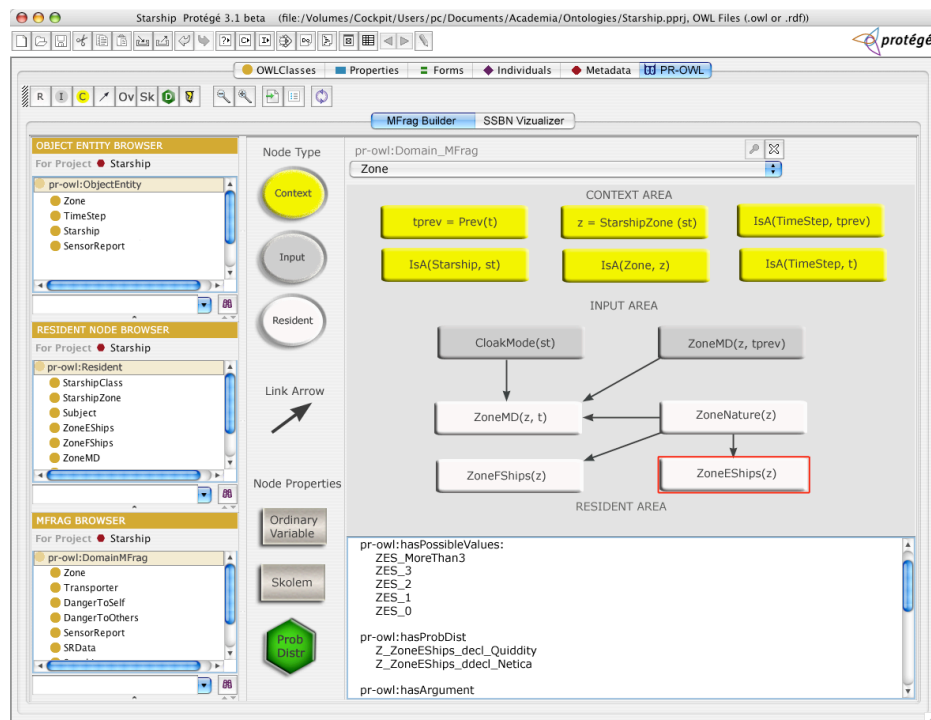


Fig. 3. Elements of a PR-OWL Probabilistic Ontology

This brief idea of an operational concept barely scratches the surface of the many possibilities for the technology presented here. Implementing a plugin such as the one envisioned here is a development task that is a topic for future research. Nonetheless,

the PR-OWL upper ontology definitions take an important first step toward making probabilistic ontologies a reality. By opening the door to wide use of PR-OWL probabilistic ontologies, the present research makes a significant contribution to realizing the Semantic Web vision.

6 Conclusion

This paper describes a coherent, comprehensive probabilistic framework for the Semantic Web, that provides a means of representing probabilistic knowledge and providing web services such as plausible inference and Bayesian learning. The proposed framework is an initial step towards a more comprehensive effort focused on representing uncertainty in the Semantic Web.

A PR-OWL plugin for current OWL ontology editors is a priority for future efforts. The process of writing probabilistic ontologies can be greatly improved via automation of most of the steps in the ontology building, not only for defining MFragments to represent sets of related hypotheses, but also for consistency checking, reified relations and other tasks that demand unnecessary awareness of the inner workings of the present solution. Once implemented, such a plugin has the potential to make probabilistic ontologies a natural, powerful tool for the Semantic Web.

Finally, the most important requirement for adoption of a language is the standardization process. This process goes significantly beyond academic research and thus falls outside the scope of the present work. Nonetheless, we are confident of its feasibility, which we believe we have demonstrated in this effort, and of its desirability, given its potential to help solve many of the obstacles that stand in the way of realizing the W3C's vision for the Semantic Web.

References

1. Berners-Lee, T. and M. Fischetti, *Weaving the Web: the original design and ultimate destiny of the World Wide Web by its inventor*. 1st pbk. ed. 2000, New York: Harper-Collins Publishers. ix, 246 p.
2. Patel-Schneider, P.F., P. Hayes, and I. Horrocks, *OWL Web ontology language - Semantics and abstract syntax*, in *W3C Recommendation*. 2004, World Wide Web Consortium: Boston, MA. p. W3C Recommendation.
3. Baader, F., et al., eds. *The Description Logic Handbook: Theory, Implementation and Applications*. First edition ed. 2003, Cambridge University Press: Cambridge, UK. 574.
4. Korb, K.B. and A.E. Nicholson, *Bayesian Artificial Intelligence*. Series in Computer Science and Data. 2003: Chapman & Hall/CRC. 392.
5. Laskey, K.B. and P.C.G. Costa, *Of Klingons and Starships: Bayesian Logic for the 23rd Century*, in *Uncertainty in Artificial Intelligence: Proceedings of the Twenty-first Conference*. 2005, AUAI Press: Edinburgh, Scotland.
6. Heckerman, D., A. Mamdani, and M.P. Wellman, *Real-world applications of Bayesian networks*. Communications of the ACM, 1995. **38**(3): p. 24-68.

7. Ding, Z. and Y. Peng. *A probabilistic extension to ontology language OWL*. in *37th Annual Hawaii International Conference on System Sciences (HICSS'04)*. 2004. Big Island, Hawaii.
8. Gu, T., P.H. Keng, and Z.D. Qing. *A Bayesian approach for dealing with uncertainty contexts*. in *Second International Conference on Pervasive Computing*. 2004. Vienna, Austria: Austrian Computer Society.
9. Calvanese, D. and G. De Giacomo, *Expressive Description Logics*, in *The Description Logic Handbook: Theory, Implementations and Applications*, F. Baader, et al., Editors. 2003, Cambridge University Press: Cambridge, UK. p. 184-225.
10. Pool, M. and J. Aikin, *KEEPER: and Protégé: An elicitation environment for Bayesian inference tools*, in *Workshop on Protégé and Reasoning held at the Seventh International Protégé Conference*. 2004: Bethesda, MD, USA.
11. Giugno, R. and T. Lukasiewicz. *P-SHOQ(D): A probabilistic extension of SHOQ(D) for probabilistic ontologies in the Semantic Web*. in *European Conference on Logics in Artificial Intelligence (JELIA 2002)*. 2002. Cosenza, Italy: Springer.
12. Daconta, M.C., L.J. Obrst, and K.T. Smith, *The Semantic Web: A guide to the future of XML, Web services, and knowledge management*. 2003, Indianapolis, IN: Wiley Publishing, Inc. 312.
13. Noy, N.F. and A. Rector, *Defining N-ary relations on the Semantic Web: Use with individuals*, in *W3C Working Draft*. 2004, World Wide Web Consortium: Boston, MA. p. W3C Working Draft.
14. Costa, P.C.G., *Bayesian Semantics for the Semantic Web*, in *Department of Systems Engineering and Operations Research*. 2005, George Mason University: Fairfax, VA, USA. p. 312. Available at www.pr-owl.org. Available at <http://www.pr-owl.org>.