Probabilistic Ontology and Knowledge Fusion for Procurement Fraud Detection in Brazil

Rommel N. Carvalho[1], Kathryn B. Laskey[1], Paulo C. G. Costa[1], Marcelo Ladeira[2], Laécio L. Santos[2], and Shou Matsumoto[2],

[1] George Mason University
4400 University Drive
Fairfax, VA 22030-4400 USA
rommel.carvalho@gmail.com, {klaskey, pcosta}@gmu.edu
[2] University of Brasilia
Campus Universitário Darcy Ribeiro
Brasilia – DF 70910-900 Brazil
mladeira@unb.br, {laecio, cardialfly}@gmail.com

Abstract. To cope with society's demand for transparency and corruption prevention, the Brazilian Office of the Comptroller General (CGU) has carried out a number of actions, including: awareness campaigns aimed at the private sector; campaigns to educate the public; research initiatives; and regular inspections and audits of municipalities and states. Although CGU has collected information from hundreds of different sources - Revenue Agency, Federal Police, and others - the process of fusing all this data has not been efficient enough to meet the needs of CGU's decision makers. Therefore, it is natural to change the focus from data fusion to knowledge fusion. As a consequence, traditional syntactic methods must be augmented with techniques that represent and reason with the semantics of databases. However, commonly used approaches fail to deal with uncertainty, a dominant characteristic in corruption prevention. This paper presents the use of Probabilistic OWL (PR-OWL) to design and test a model that performs information fusion to detect possible frauds in procurements involving Federal money. To design this model, a recently developed tool for creating PR-OWL ontologies was used with support from PR-OWL specialists and careful guidance from a fraud detection specialist from CGU.

Keywords: Probabilistic Ontology, PR-OWL, Ontology, Procurement Fraud Detection, Knowledge Fusion, MEBN, UnBBayes.

1 Introduction

A primary responsibility of the Brazilian Office of the Comptroller General (CGU) is to prevent and detect government corruption. To carry out this mission, CGU must gather information from a variety of sources and combine it to evaluate whether

further action, such as an investigation, is required. One of the most difficult challenges is the information explosion. Auditors must fuse vast quantities of information from a variety of sources in a way that highlights its relevance to decision makers and helps them focus their efforts on the most critical cases. This is no trivial duty. The Growing Acceleration Program (PAC) alone has a budget greater than 250 billion dollars with more than one thousand projects only on the state of Sao Paulo (http://www.brasil.gov.br/pac/). All of these have to be audited and inspected by CGU – and, in spite having only three thousand employees. Therefore, CGU must optimize its processes in order to carry out its mission.

The Semantic Web (SW), like the document web that preceded it, is based on radical notions of information sharing. These ideas [1] include: (i) the Anyone can say Anything about Any topic (AAA) slogan; (ii) the open world assumption, in which we assume there is always more information that could be known, and (iii) nonunique naming, which appreciates the reality that different speakers on the Web might use different names to define the same entity. In a fundamental departure from assumptions of traditional information systems architectures, the Semantic Web is intended to provide an environment in which information sharing can thrive and a network effect of knowledge synergy is possible. But this style of information gathering can generate a chaotic landscape rife with confusion, disagreement and conflict.

We call an environment characterized by the above assumptions a Radical Information Sharing (RIS) environment. The challenge facing SW architects is therefore to avoid the natural chaos to which RIS environments are prone, and move to a state characterized by information sharing, cooperation and collaboration. According to [1], one solution to this challenge lies in modeling, and this is where ontologies languages like Web Ontology Language (OWL) come in.

As it will be shown in Section 3, the domain of procurement fraud detection is a RIS environment. However, uncertainty is ubiquitous to knowledge fusion. Uncertainty is especially important to applications such as fraud detection, in which perpetrators seek to conceal illicit intentions and activities, making crisp assertions extremely hard and rare. In such environments, partial (not complete) or approximate (not exact) information is more the rule than the exception.

Bayesian networks (BNs) have been widely applied to draw inferences to information and knowledge fusion in the presence of uncertainty. However, according to [2] BNs are not expressive enough for many real-world applications. More specifically, BNs assume a simple attribute-value representation – that is, each problem instance involves reasoning about the same fixed number of attributes, with only the evidence values changing from problem instance to problem instance. Complex problems on the scale of the semantic web often involve intricate relationships among many variables, and the limited representational power of BNs is insufficient for building useful, detailed models.

Multi-Entity Bayesian Network (MEBN) logic can represent and reason with uncertainty about any propositions that can be expressed in first-order logic [3]. Probabilistic OWL (PR-OWL) uses MEBN's strengths to provide a framework for building probabilistic ontologies (PO), a major step towards semantically aware, probabilistic knowledge fusion systems [4]. This paper uses PR-OWL to design and

test a model for fusing information to detect possible frauds in procurements involving Federal funds.

The paper is organized as follows. Section 2 introduces Multi-Entity Bayesian Networks (MEBN), an expressive Bayesian logic, and PR-OWL, an extension of the OWL language that can represent probabilistic ontologies having MEBN as its underlying logic. Section 3 presents a case study from CGU to demonstrate the power of PR-OWL ontologies for knowledge representation and fusion. Finally, Section 4 presents some concluding remarks.

2 MEBN and PR-OWL

Multi-Entity Bayesian Networks (MEBN) [5 and 6] extend BNs (BN) to achieve first-order expressive power. MEBN represents knowledge as a collection of MEBN Fragments (MFrags), which are organized into MEBN Theories (MTheories).

An MFrag contains random variables (RVs) and a fragment graph representing dependencies among these RVs. An MFrag is a template for a fragment of a Bayesian network. It is instantiated by binding its arguments to domain entity identifiers to create instances of its RVs. There are three kinds of RV: context, resident and input. Context RVs represent conditions that must be satisfied for the distributions represented in the MFrag to apply. Input nodes represent RVs that may influence the distributions defined in the MFrag, but whose distributions are defined in other MFrags. Distributions for resident RV instances are defined in the MFrag. Distributions for resident RVs are defined by specifying local distributions conditioned on the values of the instances of their parents in the fragment graph.

A set of MFrags represents a joint distribution over instances of its random variables. MEBN provides a compact way to represent repeated structure in a BN. An important advantage of MEBN is that there is no fixed limit on the number of RV instances, and the random variable instances are dynamically instantiated as needed.

An MTheory is a set of MFrags that satisfies conditions of consistency ensuring the existence of a unique joint probability distribution over its random variable instances.

To apply an MTheory to reason about particular scenarios, one needs to provide the system with specific information about the individual entity instances involved in the scenario. On receipt of this information, Bayesian inference can be used both to answer specific questions of interest (e.g., how likely is it that a particular procurement is being directed to a specific enterprise?) and to refine the MTheory (e.g., each new tactical situation includes additional statistical data about the likelihood of a given attack for that set of circumstances). Bayesian inference is used to perform both problem specific inference and learning in a sound, logically coherent manner (for more details see [6 and 7]).

State-of-the-art systems are increasingly adopting ontologies as a means to ensure formal semantic support for knowledge sharing [8, 9, 10, 11, 12, and 13]. Representing and reasoning with uncertainty is becoming recognized as an essential capability in many domains. A common error is to provide support for uncertainty representation by just annotating ontologies with numerical probabilities. This

approach leads to brittleness, as too much information is lost due to the lack of a representational scheme that can capture structural nuances of the probabilistic information. More expressive representation formalisms are needed [4].

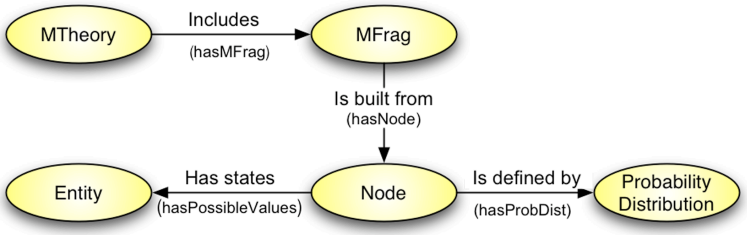

Fig. 1. PR-OWL main concepts.

Probabilistic Ontologies (PR-OWL) [14 and 15] was proposed as a more expressive formalism for representing knowledge in domains characterized by uncertainty. Figure 1 presents the main concepts needed to define an MTheory in PR-OWL. In the diagram, the ellipses represent the general classes, while the arcs represent the main relationships among the classes.

The procurement fraud detection probabilistic ontology was built in UnBBayes-MEBN, a tool for building and reasoning with PR-OWL probabilistic ontologies. UnBBayes-MEBN was the first software to implement PR-OWL/MEBN (see [16, 17, 18, 19] for more details). UnBBayes-MEBN supports Multi-Entity Bayesian Network (MEBN) and enables creation and editing of Probabilistic Ontologies in PR-OWL [18]. The MEBN/PR-OWL Graphical User Interface (GUI) [16] allows users to define MFrags and make probabilistic queries. UnBBayes-MEBN also implements an algorithm for generating a Situation Specific Bayesian Network (SSBN) [18, 19], which is an ordinary BN created by instantiating instances of the MFrags to respond to a probabilistic query. Once the SSBN is generated, the inference engine (Reasoning) is called to process findings and update beliefs. UnBBayes-MEBN uses the Protégé-OWL library to load and save PR-OWL files (IO) in a format compatible with OWL. It supports first order logic context node evaluation (FOL), through the use of the PowerLoom library. It also defines and implements a built-in mechanism for typing and recursion. Finally, it permits the definition of dynamic conditional probabilistic tables.

UnBBayes has proven to be a simple, yet powerful, tool for designing probabilistic ontologies and for uncertain reasoning in complex situations such as procurement fraud detection. It is straightforward to use and provides powerful features (e.g. dynamic table) not available in systems (e.g., Quiddity) previously employed to reason with PR-OWL/MEBN knowledge bases.

3 Procurement Fraud Detection

A major source of corruption is the procurement process. Although laws attempt to ensure a competitive and fair process, perpetrators find ways to turn the process to their advantage while appearing to be legitimate. This is why a specialist has didactically structured the different kinds of procurement frauds CGU has dealt with in past years.

These different fraud types are characterized by criteria, such as business owners who work as a front for the company, use of accounting indices that are not common practice, etc. Indicators have been established to help identify cases of each of these fraud types. For instance, one principle that must be followed in public procurement is that of competition. Every public procurement should establish minimum requisites necessary to guarantee the execution of the contract in order to maximize the number of participating bidders. Nevertheless, it is common to have a fake competition when different bidders are, in fact, owned by the same person. This is usually done by having someone as a front for the enterprise, which is often someone with little or no education.

The ultimate goal of this case study is to structure the specialist knowledge in a way that an automated system can reason with the evidence in a manner similar to the specialist. Such an automated system is intended to support specialists and to help train new specialists, but not to replace them. Initially, a few simple criteria were selected as a proof of concept. Nevertheless, it is shown that the model can be incrementally updated to incorporate new criteria. In this process, it becomes clear that a number of different sources must be consulted to come up with the necessary indicators to create new and useful knowledge for decision makers about the procurements.

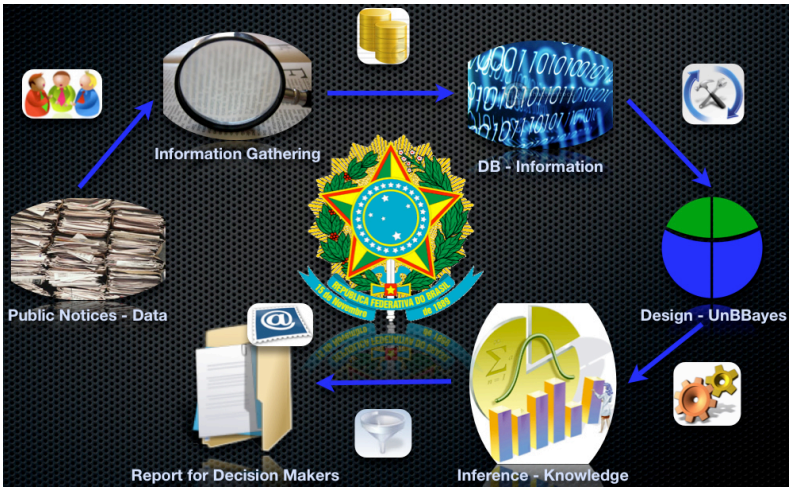

Fig. 2. Procurement fraud detection overview.

Figure 2 presents an overview of the procurement fraud detection process. The data for our case study represent several requests for proposal and auctions that are issued

by the Federal, State and Municipal Offices (Public Notices – Data). As the focus of the work is in representing the specialist knowledge and reasoning through probabilistic ontologies and not in the collection of information, the idea is that the analysts that work at CGU, already making audits and inspections, accomplish the collection of information through questionnaires that can specifically be created for the collecting of indicators for the selected criteria (Information Gathering). These questionnaires can be created using a system that is already in production at CGU. Once they are answered the necessary information is going to be available (DB – Information). Hence, UnBBayes, using the probabilistic ontology designed by experts (Design – UnBBayes), will be able to collect these millions of items of information and transform them into dozens or hundreds of items of knowledge, through logic and probabilistic inference, e.g. procurement announcements, contracts, reports, etc - a huge amount of data - are analyzed allowing the gathering of relevant relations and properties - a large amount of information - which in turn are used to draw some conclusions about possible irregularities - a smaller number of items of knowledge (Inference – Knowledge). This knowledge can be filtered so that only the procurements that show a probability higher than a threshold, e.g. 20%, are automatically forwarded to the responsible department along with the inferences about potential fraud and the supporting evidence (Report for Decision Makers).

The criteria selected by the specialist were the use of accounting indices and the demand of experience in just one contract. There are four common types of indices that are usually used as requirements in procurements (ILC, ILG, ISG, and IE). Any other type could indicate a made-up index specifically designed to direct the procurement to some specific company. The greater the numbers of uncommon accounting indices used by the procurement the more suspicious it is, i.e. the higher the chance of having fraud. In addition, a procurement specifies a minimum value for these accounting indices. The minimum value that is usually required is 1.0. The higher this minimum value, the more the competition is narrowed, and therefore the higher the chance the procurement is being directed to some company.

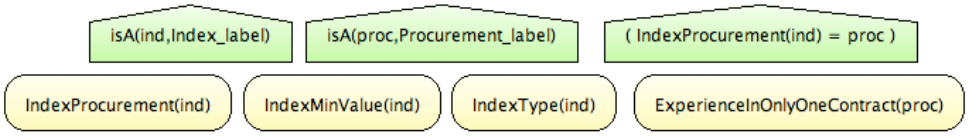

Fig. 3. ProcurementRequirement MFrag.

The other criterion, demanding proof of experience in only one contract, is suspect because in almost every case, the experience is not gained only by a particular contract, but also by doing it over and over again in different contracts. It does not matter if you have built 1,000 ft2 of wall in just one contract or 100 ft2 in 10 different contracts. The experience gained will be basically the same.

The procurement fraud detection model was developed as a probabilistic ontology (using PR-OWL) to define its semantics and uncertain characteristics. The MTheory created for the model, using UnBBayes-MEBN, was divided into three different MFrags.

The first, Figure 3, presents the criteria required from a company to participate in the procurement, containing information about the type of accounting index (ILC, ILG, ISG, IE, and Other) and the minimum value for it (between 0 and 1, between 1 and 2, between 2 and 3, and greater than 3). This MFrag also contains information about where a specific index is used (which procurement), and if the procurement demands experience in only one contract.

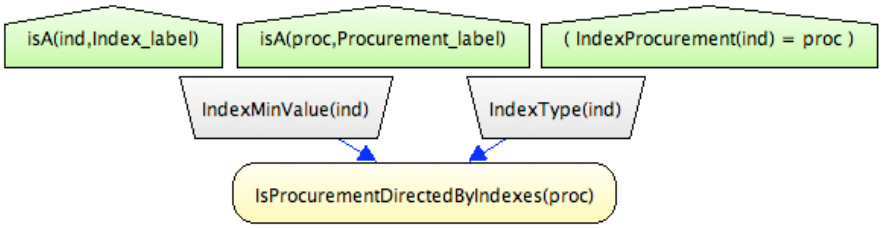

Fig. 4. DirectingProcurementByIndexes MFrag.

The second, Figure 4, represents whether procurement is being directed to a specific company by the use of unusual accounting indices. As explained before, this analysis is based on the type of the index and the minimum value it requires. This evaluation takes into consideration every index used in a specific procurement, hence it is dynamic.

The last MFrag, Figure 5, represents the overall possibility that procurement is being directed to a specific company based on the result of its being directed by the use of unusual indices and by the requirement of experience in only one contract, as explained before.

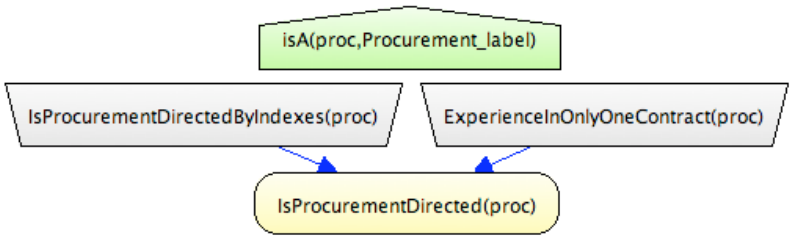

Fig. 5. DirectingProcurement MFrag.

To test the model, two scenarios, that represent the two groups of suspect and non suspect procurements, were chosen from a set of real cases, as shown:

- Suspect procurement (*proc1*):
  - ind1 = ILC >= 2.0;
  - ind2 = ILG >= 1.5;
  - ind3 = Other >= 3.0.
  - It demands experience in only one contract.
- Non suspect procurement (*proc2*):
  - ind4 = IE >= 1.0;

- ind5 = ILG >= 1.0;
- ind6 = ILC >= 1.0;
- It does not demand experience in only one contract.

The information above was introduced in our model as known entities and findings. After that we queried the system to give us information about the node *IsProcurementDirected(proc)* for both *proc1* and *proc2*. UnBBayes-MEBN than executed the SSBN algorithm and generated the same node structure as shown in Figure 6, because both procurements have three accounting indices and information about the demanding experience in only one contract. However, as expected, the parameters and findings are different giving different results to the query, as shown below:

- Non suspect procurement:
  - 0.01% that the procurement was directed to a specific company by using accounting indices;
  - 0.10% that the procurement was directed to a specific company.
- Suspect procurement:
  - 55.00% that the procurement was directed to a specific company by using accounting indices;
  - 29.77%, when the information about demanding experience in only one contract was omitted, and 72.00%, when it was given, that the procurement was directed to a specific company.

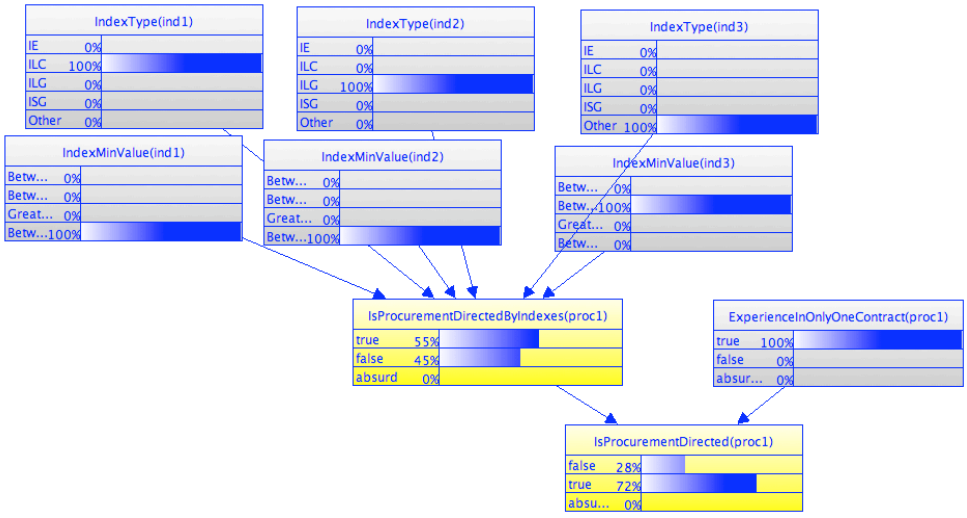

Fig. 6. Generated SSBN for query IsProcurementDirected(proc1).

The specialist analyzed and agreed with the knowledge generated by the probabilistic ontology reasoned developed using PR-OWL/MEBN in UnBBayes. He stated that the probabilities represent, semantically (i.e. high, medium, and low chance), what he would think when analyzing the same entities and findings.

Although the SSBNs generated for this proof of concept present the same structure, it is common to have a different one as the context varies from procurement to

procurement. For instance, we have come across several procurements that have all four common indices and some other different ones. In this case, if there are two additional indices (*ind5* and *ind6*), then the resulting SSBN would have two more copies for nodes *IndexType(index)* and*IndexMinValue(index)*. This would make the use of BN not applicable. The ability to make multiple copies of nodes based on a context is only available in a more expressive formalism, as MEBN.

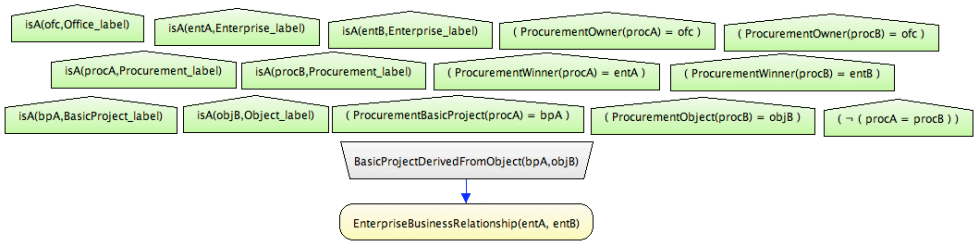

Fig. 7. EnterpriseBusinessNetwork MFrag.

An additional capability not available with BN is to specify constraints on applicability of knowledge. Such constraints can only be implemented in a more expressive language. As we are dealing with BN formalism it is only natural to think of a formalism that extends BN. MEBN, as a Bayesian first-order logic, makes it possible to define these constraints using FOL.

Figure 7 presents the constraints (context nodes) necessary to model the fraud detection scenarios considered here. In this MFrag, the criterion is to identify if there is a suspicious business relationship between enterprises *entA* and *entB*. The more cases where enterprise B wins a procurement that the basic project was developed by enterprise A, the higher the chance they have some kind of personal business relationship, which means that it is more likely that enterprise B is developing the basic projects in such a way that will favor enterprise A, inhibiting the desired competition.

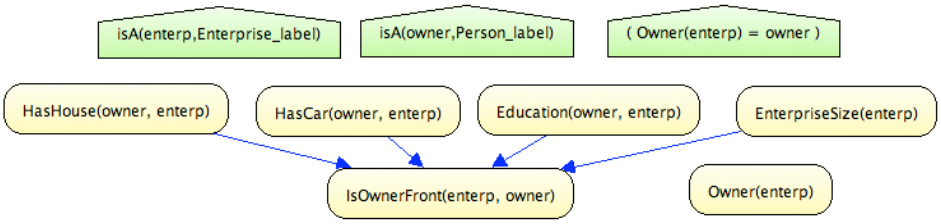

Fig. 8. OwnerFront MFrag.

Since the designed model is restricted to just two criteria, the team started to think about other criteria that could be incorporated and tested further. Figure 8 presents the suggested MFrag for detecting owners that act as a front to the real owner of the company (the person who really has the power to make decisions and that gets all the money), by looking up their socio-economic attributes and checking the size of the

company. In other words, if a company is highly profitable, yet has an owner with little education, low income, no car, no house, etc, then the company is probably a front.

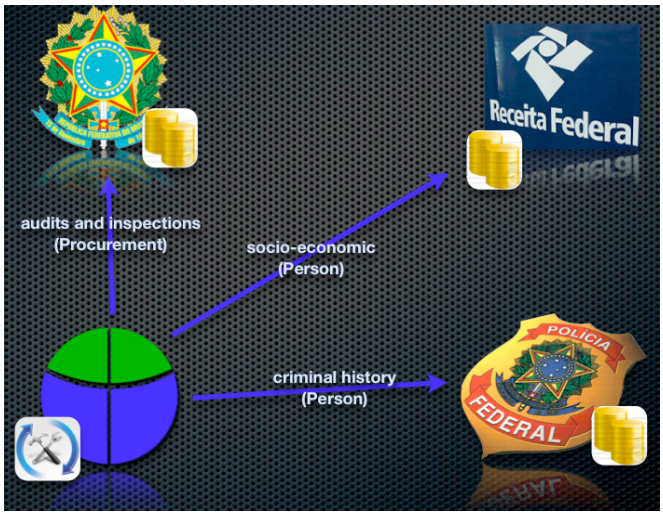

Fig. 9. Knowledge fusion from different Government Offices DBs.

From the criteria presented and modeled in this Section, we can clearly see the need for a principled way of dealing with uncertainty. But what is the role of Semantic Web in this domain? Well, it is easy to see that our domain of fraud detection is a RIS environment. The data CGU has available does not come only from its audits and inspections. In fact, much complementary information can be retrieved from other Federal Agencies, including Federal Revenue Agency, Federal Police, and others. Imagine we have information about the enterprise that won the procurement, and we want to know information about its owners, such as their personal data and annual income. This type of information is not available at CGU's Data Base (DB), but must be retrieved from the Federal Revenue Agency's DB. Once the information about the owners is available, it might be useful to check their criminal history. For that (see Figure 9), information from the Federal Police must be used. In this example, we have different sources saying different things about the same person: thus, the AAA slogan applies. Moreover, there might be other Agencies with crucial information related to our person of interest; in other words, we are operating in an open world. Finally, to make this sharing and integration process possible, we have to make sure we are talking about the same person, who may (especially in case of fraud) be known by different names in different contexts.

5 Conclusion

The problem that CGU and many other Agencies have faced of processing all the available data into useful knowledge is starting to be solved with the use of

probabilistic ontologies, as the procurement fraud detection model showed. Besides fusing the information available, the designed model was able to represent the specialist knowledge for the two real cases we evaluated. UnBBayes reasoning given the evidence and using the designed model were accurate both in suspicious and non suspicious scenarios. These results are encouraging, suggesting that a fuller development of our proof of concept system is promising.

In addition, it is fairly easy to introduce new criteria and indicators in the model in an incremental way. Thus, new rules for identifying fraud can be added without rework. After a new rule is incorporated into the model, a set of new tests can be added to the previous one with the objective of always validating the new model proposed, without doing everything from scratch.

Furthermore, the use of this formalism through UnBBayes allows advantages such as impartiality in the judgment of irregularities in procurements (given the same conditions the system will always deliver the same result), scalability (capacity to analyze thousands of procurements in a short time when compared to human capacity) and a joint analysis of large volumes of indicators (the higher the number of indicators to examine jointly the more difficult it is for the specialist analysis to be objective and consistent).

As a next step, CGU is choosing new criteria to be incorporated into the designed probabilistic ontology. This next set of criteria will require information from different Brazilian Agencies' databases. Therefore, the semantic power of ontologies with the uncertainty handling capability of PR-OWL will be extremely useful for fusing information from multiple databases.

Acknowledgments. Rommel Carvalho gratefully acknowledges full support from the Brazilian Office of the Comptroller General (CGU) for the research reported in this paper, and its employees involved in this research, especially Mário Vinícius Claussen Spinelli, the domain expert.

References

1. Allemang, D. & Hendler, J. A. 2008. Semantic web for the working ontologist modeling in RDF, RDFS and OWL. Elsevier, ISBN 978-0-12-373556-0, United States.
2. Costa, P. C. G., Laskey, K. B., Takikawa, M., Pool, M., Fung, F., and Wright, E. J. 2005. MEBN logic: A Key Enabler for Network Centric Warfare. *In Proceedings of the 10th International Command and Control Research and Technology Symposium (10th ICCRTS)*. McLean, Virginia, USA, CCRP publications.
3. Laskey, K. B., Mahoney, S. M., and Wright, E. 2001. Hypothesis Management in Situation-Specific Network Construction. *Uncertainty in Artificial Intelligence: Proceedings of the Seventeenth Conference*, San Mateo, CA, Morgan Kaufman.
4. Laskey, K. B., Costa, P. C. G., and Janssen, T. 2008. Probabilistic Ontologies for Knowledge Fusion. In *Proceedings of the 11th International Conference on Information Fusion*.
5. Laskey, K. B. and Costa, P. .C. G. 2005. Of Klingons and Starships: Bayesian Logic for the 23rd Century. *In Uncertainty in Artificial Intelligence: Proceedings of the Twenty-first Conference (UAI 2005)*. AUAI Press: Edinburgh, Scotland.
6. Laskey, K. B. 2008. MEBN: A language for first-order Bayesian knowledge bases. *In Artificial Intelligence*, Volume 172, Issues 2-3, February 2008, Pages 140-178

7. Mahoney, S. and Laskey, K. B. 1998. Constructing Situation Specific Belief Networks, In Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI-98). San Francisco, CA: Morgan Kaufmann.
8. Chen, H., and Wu, Z.. 2003. On Case-Based Knowledge Sharing in Semantic Web. *In Tools with Artificial Intelligence*, IEEE International Conference on, 0:200. Vol. 0. Los Alamitos, CA, USA, IEEE Computer Society.
9. Chen, H., Wu, Z., and Xu, J. 2003. KB-Grid: Enabling Knowledge Sharing on the Semantic Web. *In Challenges of Large Applications in Distributed Environments*, International Workshop on, 0:70. Vol. 0. Los Alamitos, CA, USA: IEEE Computer Society.
10. Costa, Paulo C. G., Chang, KC., Laskey, K. B., and Carvalho, R. N. 2009. A Multi-Disciplinary Approach to High Level Fusion in Predictive Situational Awareness. *In Proceedings of the 12th International Conference on Information Fusion*. Seattle, WA, USA.
11. Dadzie, AS., Bhagdev, R., Chakravarthy, A., Chapman, S., Iria, J., Lanfranchi, V., Magalhães, J., Petrelli, D., and Ciravegna. F. 2008. Applying semantic web technologies to knowledge sharing in aerospace engineering. *Journal of Intelligent Manufacturing* 20, no. 5 (6): 611-623. doi:10.1007/s10845-008-0141-1.
12. Kings, N. J. and Davies, J. 2009. Semantic Web for Knowledge Sharing. *In Semantic Knowledge Management*, 103-111. http://dx.doi.org/10.1007/978-3-540-88845-1_8.
13. Veres, G. V., Huynh, T. D., Nixon, M. S., Smart, P. R. and Shadbolt, N. R. 2006. The Military Knowledge Information Fusion Via Semantic Web Technologies. http://eprints.ecs.soton.ac.uk/14278/.
14. Costa, P. C. G. 2005. Bayesian Semantics for the Semantic Web. PhD Diss. Department of Systems Engineering and Operations Research, George Mason University. 315p, Fairfax, VA, USA.
15. Costa, P. C. G., Laskey, K. B., and Laskey, K. J. 2005. PR-OWL: A Bayesian Ontology Language for the Semantic Web. In *Proceedings of the ISWC Workshop on Uncertainty Reasoning for the Semantic Web*, Galway, Ireland.
16. Carvalho, R. N., Ladeira, M., Santos, L. L., and Costa, P. C. 2007. A GUI Tool for Plausible Reasoning in the Semantic Web using MEBN. In *Proceedings of the Seventh international Conference on intelligent Systems Design and Applications,* 381-386. ISDA. IEEE Computer Society, Washington, DC, USA.
17. Carvalho, R. N., Ladeira, M., Santos, L. L., Matsumoto, S., and Costa, P. C. G. 2008. UnBBayes-MEBN: Comments on Implementing a Probabilistic Ontology Tool. In *Proceedings of the IADIS International Conference on Applied Computing,* 211-218.
18. Costa, P. C. G., Ladeira, M, Carvalho, R. N., Santos, L. L., Matsumoto, S., and Laskey, K. B. 2008. A First-Order Bayesian Tool for Probabilistic Ontologies. In *Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference,* 631-636. Menlo Park, California, USA, The AAAI Press.
19. Carvalho, R. N., Ladeira, M., Santos, L. L., Matsumoto, S., and Costa, P. C. G. 2009. A GUI Tool for Plausible Reasoning in the Semantic Web Using MEBN. In *Book Innovative Applications in Data Mining,* 17-45. DOI: 10.1007/978-3-540-88045-5_2. Springer Berlin / Heidelberg.