Forthcoming

2016

**Title**: Identifying IT purchases anomalies in the Brazilian Government Procurement System using Deep Learning (paper)  
**Conference:** 15th IEEE International Conference on Machine Learning and Applications (ICMLA 2016)  
**URL:** <http://www.icmla-conference.org/icmla16/>   
**Abstract**: The Department of Research and Strategic Information (DIE), from the Brazilian Office of the Comptroller General (CGU), is responsible for investigating potential problems related to federal expenditures. To pursue this goal, DIE regularly has to analyze large volumes of data to search for anomalies that can reveal suspicious activities. With the growing demand from the citizens for transparency and corruption prevention, DIE is constantly looking for new methods to automate these processes. In this work, we investigate IT purchases anomalies in the Federal Government Procurement System by using a deep learning algorithm to generate a predictive model. This model will be used to prioritize actions carried out by the office in its pursuit of problems related to this kind of purchases. The data mining process followed the CRISP-DM methodology and the modeling phase tested the parallel resources of the H2O tool. We evaluated the performance of twelve deep learning with auto-encoder models, each one generated under a different set of parameters, in order to find the best input data reconstruction model. The best model achieved a mean squared error (MSE) of 0.0012775 and was used to predict the anomalies over the test file samples.

**Title**: Uncertainty modeling process for semantic technology (journal paper)  
**Journal:** PeerJ Computer Science  
**URL:** <https://peerj.com/articles/cs-77/>   
**Abstract**: The ubiquity of uncertainty across application domains generates a need for principled support for uncertainty management in semantically aware systems. A probabilistic ontology provides constructs for representing uncertainty in domain ontologies. While the literature has been growing on formalisms for representing uncertainty in ontologies, there remains little guidance in the knowledge engineering literature for how to design probabilistic ontologies. To address the gap, this paper presents the Uncertainty Modeling Process for Semantic Technology (UMP-ST), a new methodology for modeling probabilistic ontologies. To explain how the methodology works and to verify that it can be applied to different scenarios, this paper describes step-by-step the construction of a proof-of-concept probabilistic ontology. The resulting domain model is intended to support identification of fraud in public procurements in Brazil. While the case study illustrates the development of a probabilistic ontology in the PR-OWL probabilistic ontology language, the methodology is applicable to any ontology formalism that properly integrates uncertainty with domain semantics.

**Title**: Bayesian Models to Assess Risk of Corruption of Federal Management Units (paper)  
**Conference:** 13th Annual Bayesian Modeling Applications Workshop (BMAW 2016)  
**URL:** <http://c4i.gmu.edu/bmaw/2016/>   
**Abstract**: This paper presents a data mining project that generated Bayesian models to assess risk of corruption of federal management units. With thousands of extracted features related to corruptibility, the data were processed using techniques like correlation analysis and variance per class. We also compared two different discretization methods: Minimum Description Length Principle (MDLP) and Class-Attribute Contingency Coefficient (CACC). The feature selection process used Adaptive Lasso. To choose our final model we evaluated three different algorithms: Naïve Bayes, Tree Augmented Naïve Bayes, and Attribute Weighted Naïve Bayes. Finally, we analyzed the rules generated by the model in order to support knowledge discovery.

**Title**: Measuring the Risk of Public Contracts Using Bayesian Classifiers (paper)  
**Conference:** 13th Annual Bayesian Modeling Applications Workshop (BMAW 2016)  
**URL:** <http://c4i.gmu.edu/bmaw/2016/>  
**Abstract**: Bayesian Classifiers are widely used in machine learning supervised models where there is a reasonable reliability in the dependent variable. This work aims to create a risk measurement model of companies that negotiate with the government using indicators grouped into four risk dimensions: operational capacity, history of penalties and findings, bidding profile, and political ties. It is expected that this model contributes to the selection of contracts to be audited under the central unit of internal control of the Brazilian government, responsible for auditing more than 30,000 public contracts per year.

**Title**: Identifying the Main Problems in IT Auditing: a Comparison Between Unsupervised and

Supervised Learning  
**Conference:** 27th DEXA Conference/ EGOVIS  
**URL:** [http://link.springer.com/chapter/10.1007%2F978-3-319-44159-7\_17](http://link.springer.com/chapter/10.1007%2F978-3-319-44159-7_17%20)   
**Abstract**: One of the main challenges faced by the Brazilian Office of the Comptroller General (CGU) is applying consistent knowledge discovery tools and methodologies to learn from several years of auditing experience from hundreds of thousands of auditing reports with millions of pages it produced during these years. More specifically, we tackle the problem of identifying the most common topics in a context of Information Technology audits performed in Brazil since 2011. In order to tackle this problem, we compare two different approaches, supervised and unsupervised learning. On the one hand, the supervised learning approach generated a model that achieved around 73% accuracy for seven categories using random forest. On the other hand, the unsupervised learning approach using Latent Dirichlet Allocation (LDA) generated a model with five topics, which was considered the best model based

on the validation performed by the subject matter experts (SME) from CGU. Nevertheless, it is important to note that both approaches, although implemented independently, generated very similar topics. This also reinforces the success in identifying the main problems found during all these years of IT auditing at CGU using consistent and well-known knowledge discovery methods.

**Title**: A Comparison Between Supervised and Unsupervised Models for Identify a Large Number of Categories  
**Conference:** 17th IEEE International Conference on Information Reuse an Integration  
**URL:** <http://www.sis.pitt.edu/iri2016/>   
**Abstract**: Large amount of categories with skewed category distribution over documents still not a closed question in the state-of-the-art technologies in automated text classification.

In this paper we present a proof of concept for an automatic model of complaints screening, using text mining. Through a complaints link of the Office of the Comptroller General (CGU) site, citizens have access to a form to file their complaints. These complaints must be screened

and delivered to one of 64 CGU’s coordination by subject. Nowadays the complaints screening is done manually. Considering the complaints storage on the database now and the arrival of new complaints, combined with the time spent on manual sorting, the timely analysis of the reported occurrences it becomes increasingly difficult. We compare two approaches: supervised learning with classifiers algorithms and unsupervised leaning with topic modeling and text search. The best results were achieved using ranking based on the Huffman Tree algorithm. This proof of concept demonstrated the feasibility of automatic sorting without

the loss of quality compared to manual sorting.

2015

**Title**: Uso de Mineração de Dados e Textos para Cálculo de Preços de Referência em Compras do Governo Brasileiro (paper)  
**Award:** Segundo Lugar na Categoria Profissionais no 5º Prêmio Chico Ribeiro Sobre Informações de Custos e Qualidade do Gasto no Setor Público  
**URL:** <http://www.esaf.fazenda.gov.br/assuntos/premios/premios-1/premio-chico-ribeiro/resultado-do-5o-premio>   
**Abstract**: Uma das grandes responsabilidades da Controladoria-Geral da União (CGU) é identificar as compras do governo com valores diferentes dos praticados pelo mercado. Dessa forma, é possível mensurar o grau de eficiência das compras realizadas pelos órgãos governamentais. Essa informação é útil tanto para o auditor, que é responsável por fiscalizar o uso dos recursos públicos, como para o gestor, que pode melhorar seus processos observando as melhores práticas de outras unidades do governo. Dada a enorme quantidade e a diversidade das compras realizadas pelo Governo, essa análise se torna praticamente inviável sem a ajuda de algum mecanismo automatizado. No entanto, para que essa análise automatizada seja possível, é preciso ter antes de tudo, uma base de dados com os preços médios, ou de referência, para cada produto que se deseja analisar. Apesar de todas as compras do Governo Federal serem inseridas em um sistema único e centralizado, as informações armazenadas não são detalhadas e estruturadas o suficiente para se calcular esses preços de referência. Esse artigo apresenta a metodologia desenvolvida na CGU, baseada em técnicas de mineração de dados, para extrair as informações necessárias desse sistema centralizado de forma a possibilitar o cálculo de preços de referência para produtos comprados pelo Governo Federal. Além disso, são apresentadas também algumas análises feitas com base no banco de preços criado a partir dessa metodologia de forma a enfatizar sua importância para a melhoria da gestão dos recursos públicos.

**Title**: PR-OWL 2 RL-A Language for Scalable Uncertainty Reasoning on the Semantic Web (paper)  
**Conference:** 11th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2015)  
**URL:** <http://c4i.gmu.edu/ursw/2015/>   
**Abstract**: Probabilistic OWL (PR-OWL) improves the Web Ontology Language (OWL) with the ability to treat uncertainty using Multi-Entity Bayesian Networks (MEBN). PR-OWL 2 presents a better integration with OWL and its underlying logic, allowing the creation of ontologies with probabilistic and deterministic parts. However, there are scalability problems since PR-OWL 2 is built upon OWL 2 DL which is a version of OWL based on description logic SROIQ(D) and with high complexity. To address this issue, this paper proposes PR-OWL 2 RL, a scalable version of PR-OWL based on OWL 2 RL profile and triplestores (databases based on RDF triples). OWL 2 RL allows reasoning in polynomial time for the main reasoning tasks. This paper also presents First-Order expressions accepted by this new language and analyzes its expressive power. A comparison with the previous language presents which kinds of problems are more suitable for each version of PR-OWL.

2014

**Title**: Using Clustering and Text Mining to Create a Reference Price Database (journal paper)  
**Journal:** Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence  
**URL:** <http://www.lnlm-sbic.org/papers/vol12-no1-art3.pdf>   
**Abstract**: Since 2004, Brazil’s Office of the Comptroller General (CGU) has been publishing several data related to government expenditures in the Transparency Portal. In 2010, CGU started publishing daily every financial statement produced by the Federal Government. Nevertheless, inconsistencies which hinder accountability have been found in this data base. This paper presents how CGU uses clustering and text mining techniques to retrieve essential information for a good accountability, which includes what was bought, the price paid per item, a price reference per product, etc. This analysis has allowed CGU to draw some preliminary conclusions which are presented as a means to illustrate the research results. Finally, this information will eventually be incorporated in the Transparency Portal, allowing every citizen to understand how much the Government is really paying, in general, for products. Thus, improving social control and providing a solid accountability not only to CGU, as an internal control agency, but also to Brazil’s citizens who, in the end, are the ones paying the bill.

**Title**: Using Political Party Affiliation Data to Measure Civil Servants' Risk of Corruption (paper)  
**Conference:** Brazilian Conference on Intelligent Systems (BRACIS 2014)  
**URL:** <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6984825>   
**Abstract**: This paper presents a case study of machine learning applied to measure the risk of corruption of civil servants using political party affiliation data. Initially, a statistical hypothesis test verified the dependency between corruption and political party affiliation. Then, we constructed datasets with standardization and three different discretization techniques. Using Weka environment, this work shows the application and statistical evaluation of four classification algorithms to build models for predicting risk of corruption: Bayesian Networks, SVM, Random Forest, and Artificial Neural Networks with backpropagation. To evaluate the models we used data mining metrics such as precision, recall, kappa statistic and percent correct. Lastly, the case study compares the learned model with the best performance to the experts' model. The comparison not only confirms previous experts' affirmations, but also provides new assertions on the affiliation-corruptibility relation. CRISP-DM process model was the base reference for the data mining phases.

**Title**: Application of text mining techniques for classification of documents: a study of automation of complaints screening in a Brazilian Federal Agency (paper)  
**Conference:** Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2014)  
**URL:** <http://www.lbd.dcc.ufmg.br/bdbcomp/servlet/Trabalho?id=21567>  
**Abstract**: The Brazil's Office of the Comptroller General (CGU) is the agency of Federal Government in charge of assisting the President of Republic in matters related to internal control activities, public audits, corrective and disciplinary measures, corruption prevention and combat and coordinating ombudsman's activities. Through its website, citizens submit complains related to corruption and misuse of public sources to CGU. These complaints must be screened and delivered to the corresponding department based on its content. Nowadays the complaints screening is done manually and they are delivered to one of the 82 (eighty two) different units from various departments at CGU. This paper presents a proof of concept model to automatically classify the complaints in order to increase the speed of the current screening process. The proposed model was built using text mining and achieved an F-measure of 0.84 using Random Forest.

**Title**: 'Who is their mother?': A classification work to get answers over registration people databases (paper)  
**Conference:** Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2014)  
**URL:**  <http://www.lbd.dcc.ufmg.br/colecoes/eniac/2014/0069.pdf>  
**Abstract**: Discovering how people are related allows for a better understanding of several scenarios within an criminal investigation. If two people have family ties, such as, brothers or cousins and one of them has been involved in criminal activities, there is a high probability that the other has also been involved in these same or other similar activities. Unfortunately, in some of the cases investigated, the information we have about the people involved is not complete, and tracking relevant facts on large databases to entertain several possibilities, and then crosschecking these for accuracy, is cumbersome, often precarious and time consuming. To facilitate this effort, we have explored this problem of identifying relationships from registration forms in a database, starting with the mother because this role is the relational base of several others. We summarize the data in a proposal that collects information of two people attributing a score to them using two well-known models to treat the datasets: the Support Vector Machines, based on hyperplanes, and the Na\'{i}ve-Bayes, a model which uses conditional probability to classify the input. At the end of the process, we present a minimum set of derived attributes which contribute to get answers to this problem with the selected model. We anticipate that this model will help the specialists of the Brazilian Office of the Comptroller General (CGU) to identify collusion in the government.

**Title**: Price Database: methodology for computing reference prices in the Brazilian Government purchases (paper)  
**Conference:** XIX Congresso Internacional del CLAD - Sobre la Reforma del Estado y de la Administración Pública

**Panel:** Innovative mechanisms for management using technology and data analysis   
**URL:** <http://www.clad.org/images/pdf/Congreso/19/AT2.pdf>   
**Abstract**: One of the major responsibilities of the Brazilian Office of the Comptroller General (CGU) is to identify government purchases that deviate from market prices. Thus, it is possible to measure the degree of efficiency of the purchases made by government agencies. This information is useful both to the auditor, who is responsible for overseeing the use of public resources, as for the manager, who can improve his processes following best practices of other units in the government. Given the huge amount and diversity of the purchases made by the government, this analysis becomes almost impossible without the help of some automated mechanism. However, in order to make this automated analysis possible, it is necessary, first of all, to have a database of average prices, or benchmark, for each product that you want to analyze. Despite all purchases of the Federal Government being inserted into a single and centralized system, the stored information is not detailed and structured enough to calculate these reference prices. This paper presents a methodology developed at CGU, based on data mining techniques to extract the necessary information from the centralized system in order to enable the calculation of reference prices for products purchased by the Federal Government. Moreover, they also present some analyzes based on the database created using this methodology in order to emphasize its importance for improving the management of public resources.

**Title**: Strategic cost management: modification of the value chain to increase the efficiency of public spending.  
**Conference:** XIX Congresso Internacional del CLAD - Sobre la Reforma del Estado y de la Administración Pública  
**Panel:** Innovative mechanisms for management using technology and data analysis   
**URL:** <http://www.clad.org/images/pdf/Congreso/19/AT2.pdf>  
**Abstract**: The challenges of governments in developed and developing countries to increase the efficiency, effectiveness and quality of public spending increasingly require the use of appropriate management tools. The "Strategic Cost Management - SCM" is presented in this context as a management tool to support companies or entities to increase competitiveness. This tool can be used in the public sector to improve the management, decision making and the quality of public goods and services provided to society. This study seeks to practical application of the concepts of strategic cost management and value chain in the Comptroller General of the Union - CGU, to analyze how the components of the value chain can be adjusted to improve the quality of their products and services, reduce costs and increase the efficiency of public spending. The reference period used for the study were the financial years 2010 and 2011. The study provides a mapping of core competencies, the institutional structure, the strategic objectives and the products and services resulting from the activities of the CGU. The cost of each unit of CGU were determined and allocated through cost drivers, the products arising from the legal duties of the CGU, which allows the approximate calculation of the cost of the value chain of the CGU. From these data, we analyze the main cost driver of CGU: personnel expenses. Based on SCM literature evaluate the benefits of these costs to comply with the legal duties of CGU and availability of products and services to society. The study identifies some factors that limit the costs of obtaining information from the CGU and processes and routines that can be optimized to increase the efficiency and quality of spending the CGU.

**Title**: UMP-ST Plug-in: Documenting, Maintaining and Evolving Probabilistic Ontologies Using UnBBayes Framework (book chapter – extended version)  
**Book:** Uncertainty Reasoning for the Semantic Web III  
**URL:** <http://link.springer.com/chapter/10.1007/978-3-319-13413-0_1>   
**Abstract**: Several approaches have been proposed for dealing with uncertainty in the Semantic Web (SW). Although probabilistic ontologies (PO) is one of the most promising approach to model uncertainty in ontologies, no support has been offered to ontological engineers on how to create this more complex type of ontologies. This task has proven to be extremely difficult and hard, which motivated the creation of the Uncertainty Modeling Process for Semantic Technologies (UMP-ST), a process that guides users in modeling POs. This paper presents the UMP-ST plug-in, a tool that implements this process and shows how the plug-in, implemented in UnBBayes Framework, overcomes the main problems on modeling probabilistic ontologies: the complexity in creating; the difficulty in maintaining and evolving; and the lack of a centralized tool for documenting these ontologies. The probabilistic ontology for Procurement Fraud Detection and Prevention in Brazil is used to show how the UMP-ST plug-in overcomes these problems. This probabilistic ontology is a proof-of-concept use case created as part of a research project at the Brazilian Office of the Comptroller General (CGU).

**Title**: Improving Social Accountability of Government Purchases Through Financial Statements Data Mining (demo/abstract)  
**Conference:** Unleash Data: Accelerate Impact – KDD at Bloomberg   
**URL:** <http://www.kdd.org/kdd2014/bloombergpress.html>   
**Abstract**: Since 2004, Brazil's Office of the Comptroller General (CGU) has been publishing several data related to government expenditures as open government data in the Transparency Portal (http://transparencia.gov.br/), which was considered ``one of the five best practice in strategies for transparency and fight on corruption in the world'' by the United Nations in 2008. It has more than 1.5 billion transactions summing up to about 6 trillion US Dollars. In 2010, CGU started publishing daily every financial statement produced by the Federal Government. Nevertheless, inconsistencies which hinder accountability have been found in this data set. Among these inconsistencies, there is diesel fuel being bought for less than a cent per liter as well as for far beyond the price paid by a regular citizen. Moreover, the structured data for classifying a product is not well organized and used, thus we rely on unstructured data to identify the product we want to compute the reference price of. These difficulties and the sheer volume of data we have to analyze make this project a non-trivial endeavor. The price reference tool we developed at CGU uses statistics, data mining, and text mining techniques to retrieve essential information for a better accountability, which includes among others: what was bought, the price paid per item, a price reference per product, which agencies are buying below this price reference (in order to identify best practices), which agencies are buying above this price reference (in order to find where the Government can improve), and which agencies are deviating too much from this price reference (in order to identify possible causes of inconsistencies). Just by analyzing the last three years of purchases from a sample of 52 different products varying from fuel, food, and office supply, CGU was able to identify a potential in savings around 30 million US Dollars, which represents about 20% of the total spent. This analysis has allowed CGU to learn guidelines on purchasing best practices in different product categories, forwarding them to the agencies in need, as well as to identify and solve problems related to the use of a centralized financial system, improving accountability. Finally, this information will be incorporated in the Transparency Portal, allowing every citizen to understand how much the Government is really paying, in general, for products. In fact, this project is one of the commitments Brazil has made in its second action plan within the Open Government Partnership (OGP): “to develop a database containing reference prices for the most purchased items by the Federal Government, from data published on the Transparency Portal. The interface shall provide for the identification of items average prices, thus constituting an efficient strategy for formulating budgets and procurements, disseminating best practices in public purchases, as well as for supporting actions aimed at fighting corruption, especially in circumstances where overprice purchases are identified.” Therefore, with the publication of the results obtained from this endeavor, it will be possible to improve social control and provide a solid accountability not only to CGU, as an internal control agency, but also to Brazil's citizens who, in the end, are the ones paying the bill.

**Title**: Using Bayesian Networks to Identify and Prevent Split Purchases in Brazil (paper)  
**Conference:** 11th Annual Bayesian Modeling Applications Workshop (BMAW) at the 30th Conference on Uncertainty in Artificial Intelligence  
**Presentation**: <https://www.youtube.com/watch?v=UVOsztdSQ3A>   
**Slides**: <http://pt.slideshare.net/rommelnc/bmaw-2014-using-bayesian-networks-to-identify-and-prevent-split-purchases-in-brazil>   
**URL:** <http://seor.gmu.edu/~klaskey/BMAW2014/>   
**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 various different sources - Revenue Agency, Federal Police, and others -, going through all the data in order to find suspicious transactions has proven to be really challenging. In this paper, we present a Data Mining study applied on real data - government purchases - for finding transactions that might become irregular before they are considered as such in order to act proactively. Moreover, we compare the performance of various Bayesian Network (BN) learning algorithms with different parameters in order to fine tune the learned models and improve their performance. The best result was obtained using the Tree Augmented Network (TAN) algorithm and oversampling the minority class in order to balance the data set. Using a 10-fold cross-validation, the model correctly classified all split purchases, it obtained a ROC area of .999, and its accuracy was 99.197%.

**Title**: Risk Prevention In Brazilian Government Contracts Using Credit Scoring (book chapter – extended version)  
URL: <https://ideas.repec.org/p/por/obegef/019.html>   
**Abstract**: Credit Scoring models are statistical applications used by financial institutions to classify would-be customers as to the possibility of becoming defaulters. This research aims to take this acknowledged experience from the private sector to a governmental context, seeking to adapt it and test its performance in the identification of bidders that will most likely not fulfill their obligations in government contracts. The results of methodologies based on different statistical techniques are compared. We hope to contribute to the preventive control of contractual risks that both public administrators and government control agencies may face.

2013

**Title**: Methodology for Creating the Brazilian Government Reference Price Database (paper)  
**Conference:** X Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)  
**URL:** <http://www.lbd.dcc.ufmg.br/colecoes/eniac/2013/0033.pdf>   
**Abstract**: One of the main responsibilities of the Brazilian Office of the Comptroller General (CGU) is to identify purchases that deviate from normality. A key requirement of this process is to create and maintain a reference price database. Even though the purchases are recorded daily in a centralized system, product classification is not detailed enough to produce consistent statistics of prices per product. This paper introduces a methodology developed at CGU that relies on Data Mining techniques to address the problem of ensuring a reliable repository from a wealth of mostly unreliable data. The generation of key statistical parameters and some preliminary conclusions are presented as a means to illustrate the research results.

**Title**: Risk prevention of public procurement in the Brazilian government using credit scoring (paper)  
**Conference:** First OBEGEF International Conference: Interdisciplinary Insights on Fraud and Corruption  
**URL:** <http://www.obegef.pt/i2fc/>  
**Abstract**: Credit Scoring models are statistical applications used by financial institutions to classify applicants as to the possibility of becoming defaulters. This work aims to bring that good experience from the private sector to the governmental context, seeking to adapt it and test its performance in identifying bidders likely to fail in the fulfillment of obligations under contracts with the government. The results of methods based on different statistical techniques are compared. We hope to contribute to the preventive control of the contractual risks, both by the public manager as by the agencies of government control.

**Title**: Probabilistic Ontology and Knowledge Fusion for Procurement Fraud Detection in Brazil (book chapter – extended version)  
**Book:** Uncertainty Reasoning for the Semantic Web II  
**URL:** <http://www.springer.com/computer/ai/book/978-3-642-35974-3>  
**Abstract**: To cope with citizens’ 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 should be augmented with techniques that represent and reason with the semantics of databases. However, commonly used approaches, such as Semantic Web technologies, fail to deal with uncertainty, a dominant characteristic in corruption prevention. This paper presents the use of probabilistic ontologies built with Probabilistic OWL (PR-OWL) to design and test a model that performs information fusion to detect possible frauds in procurements involving Federal money in Brazil. 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. At present, the task of procurement fraud detection is done manually by an auditor. The number of suspicious cases that can be analyzed by a single person is small. The experimental results obtained with the presented approach are preliminary, but show the viability of developing a tool based on PR-OWL ontologies to automatize this task. This paper also exemplifies how to use PR-OWL 2.0 to provide a link between the deterministic and probabilistic parts of the ontology.

**Title**: UMP-ST plug-in: a tool for documenting, maintaining, and evolving probabilistic ontologies (paper)  
**Conference:** 9th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW)  
**URL:** <http://c4i.gmu.edu/ursw/2013/agenda>   
**Presentation**: <http://www.youtube.com/watch?v=r-ygmpf217U>  
**Slides**: <http://www.slideshare.net/rommelnc/ursw-2013-umpst-plugin>   
**Abstract**: Although several languages have been proposed for dealing with uncertainty in the Semantic Web (SW), almost no support has been given to ontological engineers on how to create such probabilistic ontologies (PO). This task of modeling POs has proven to be extremely difficult and hard to replicate. This paper presents the first tool in the world to implement a process which guides users in modeling POs, the Uncertainty Modeling Process for Semantic Technologies (UMP-ST). The tool solves three main problems: the complexity in creating POs; the difficulty in maintaining and evolving existing POs; and the lack of a centralized tool for documenting POs. Besides presenting the tool, which is implemented as a plug-in for UnBBayes, this paper also presents how the UMP-ST plug-in could have been used to build the Probabilistic Ontology for Procurement Fraud Detection and Prevention in Brazil, a proof-of-concept use case created as part of a research project at the Brazilian Office of the Comptroller General (CGU).

2012

**Title**: Transparency and Control of Government Spending in Brazil: The Role of the Public Expenditure Observatory (book chapter)  
**Book:** Open Government and Targeted Transparency: Trends and Challenges for Latin America and the Caribbean  
**URL:** <http://www.iadb.org/en/topics/transparency/support-for-countries/publication-open-government-and-targeted-transparency,7365.html>  
**Abstract**: This chapter discusses one initiative that is helping to expand government transparency and public spending monitoring. This initiative, led by the Office of the Comptroller-General (Controladoria-Geral da União, or CGU), is known as the Public Expenditure Observatory (Observatório da Despesa Pública, or ODP). One of the ODP’s main objectives is to identify risks of fraud in the use of public funds. Established as part of a CGU effort to increase the effectiveness of existing controls on the use of public funds, the ODP has helped to increase the transparency of the state. Through information technology and skilled personnel, the ODP has gradually increased the number of analyses performed. Currently, it is conducting a project known as ODP.nano, which is designed to transfer the lessons learned from the implementation of the ODP to the subnational level, for example, in the states of Santa Catarina and Bahia.

**Title**: A multi-agent data mining system for cartel detection in Brazilian government procurement (journal paper)  
**Journal:** Expert Systems with Applications: An International Journal  
**URL:** <http://www.sciencedirect.com/science/article/pii/S0957417412006343>  
**Abstract**: The main focus of this research project is the problem of extracting useful information from the Brazilian federal procurement process databases used by government auditors in the process of corruption detection and prevention to identify cartel formation among applicants. Extracting useful information to enhance cartel detection is a complex problem from many perspectives due to the large volume of data used to correlate information and the dynamic and diversified strategies companies use to hide their fraudulent operations. To attack the problem of data volume, we have used two data mining model functions, clustering and association rules, and a multi-agent approach to address the dynamic strategies of companies that are involved in cartel formation. To integrate both solutions, we have developed AGMI, an agent-mining tool that was validated using real data from the Brazilian Office of the Comptroller General, an institution of government auditing, where several measures are currently used to prevent and fight corruption. Our approach resulted in explicit knowledge discovery because AGMI presented many association rules that provided a 90% correct identification of cartel formation, according to expert assessment. According to auditing specialists, the extracted knowledge could help in the detection, prevention and monitoring of cartels that act in public procurement processes.

2011

**Title**: Probabilistic ontology: Representation and modeling methodology (PhD thesis – chapter 5 has ODP’s use case in fraud detection)  
**Degree/School:** PhD in Systems Engineering and Operations Research / George Mason University  
**URL:** <http://hdl.handle.net/1920/6616> / <http://seor.gmu.edu/phd/phd-seor.html>  
**Presentation**: <http://www.youtube.com/watch?v=Zl5rmag6BqY>  
**Slides**: <http://www.slideshare.net/rommelnc/probabilistic-ontology-representation-and-modeling-methodology-8647132>  
**Abstract**: The past few years have witnessed an increasingly mature body of research on the Semantic Web (SW), with new standards being developed and more complex problems being addressed. As complexity increases in SW applications, so does the need for principled means to cope with uncertainty in SW applications. Several approaches addressing uncertainty representation and reasoning in the SW have emerged. Among these is Probabilistic Web Ontology Language (PR-OWL), which provides WEB Ontology Language(OWL) constructs for representing Multi-Entity Bayesian network (MEBN) theories. However, there are several important ways in which the initial version PR-OWL 1 fails to achieve full compatibility with OWL. Furthermore, although there is an emerging literature on ontology engineering, little guidance is available on the construction of probabilistic ontologies. This research proposes a new syntax and semantics, defined as PR-OWL 2, which improves compatibility between PR-OWL and OWL in two important respects. First, PR-OWL 2 follows the approach suggested by Poole et al. to formalizing the association between random variables from probabilistic theories with the individuals, classes and properties from ontological languages such as OWL. Second, PR-OWL 2 allows values of random variables to range over OWL datatypes. To address the lack of support for probabilistic ontology engineering, this research describes a new methodology for modeling probabilistic ontologies called the Uncertainty Modeling Process for Semantic Technologies (UMP-ST). To better explain the methodology and to verify that it can be applied to different scenarios, this dissertation presents step-by-step constructions of two different probabilistic ontologies. One is used for identifying frauds in public procurements in Brazil and the other is used for identifying terrorist threats in the maritime domain. Both use cases demonstrate the advantages of PR-OWL 2 over its predecessor.

**Title**: AGMI-An agent-mining tool and its application to Brazilian government auditing (poster)  
**Conference:** 7th International Conference on Web Information Systems and Technologies (WEBIST)  
**URL:** <http://www.informatik.uni-trier.de/~ley/db/conf/webist/webist2011.html>  
**Abstract**: This paper presents research combining two originally separated areas increasingly interrelated: distributed multi-agent systems and data mining. In our approach, we prove the interaction features in a bilateral and complementary way, since we have defined an integrated architecture and developed a prototype, which has been used in a government auditing study case. In Brazil, government auditing is performed by the Office of the Comptroller General (CGU), where several approaches are being used to prevent and fight corruption. However, some activities such as government purchasing fraud detection are limited by the difficulty in finding effective ways to implement. Considering data mining perspective, we have used different model functions, such as clusterization and link analysis with association rules. Our approach integrating multi-agent and data mining techniques resulted in expressive discovered knowledge, which would help detection of cartels acting in public bidding processes at CGU.

2009

**Title**: Probabilistic Ontology and Knowledge Fusion for Procurement Fraud Detection in Brazil (paper)  
**Conference:** 5th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW)  
**URL:** <http://c4i.gmu.edu/ursw/2009/papers.html#agenda>  
**Presentation**:   
-Part 1/4: <http://www.youtube.com/watch?v=mItXL1nh0Rs>  
-Part 2/4: <http://www.youtube.com/watch?v=LnyzcsvHupk>  
-Part 3/4: <http://www.youtube.com/watch?v=cOueScpF8vg>  
-Part 4/4: <http://www.youtube.com/watch?v=HrcFgGsheso>  
**Slides**: <http://www.slideshare.net/rommelnc/ursw-2009-probabilistic-ontology-and-knowledge-fusion-for-procurement-fraud-detection-in-brazil>  
**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.