# Uma análise da qualidade dos dados relativos aos boletins de ocorrências das rodovias federais brasileiras para o processo de Mineração de Dados

**Jefferson de J. Costa[1], Flavia Cristina Bernardini[1], Thiago J. B. de Lima[1], José Viterbo Filho[2]**

[1] Mestrado em Engenharia de Produção e Sistemas Computacionais - Instituto de Ciência e Tecnologia - Universidade Federal Fluminense – Campus Rio das Ostras
Rua Recife, s/n – Jardim Bela Vista – Rio das Ostras – RJ – Brasil

[2] Instituto de Computação - Universidade Federal Fluminense
Rua Passo da Pátria, 156 - Bloco E - 3º andar - São Domingos - Niterói – RJ

{jeffersoncosta,thiagojeffery,fcbernardini,jviterbo}@id.uff.br,

***Abstract.*** *This paper shows a study about the data mining process applied in federal highways occurrences data, generated by the Brazilian Federal Highway Police, in 2012. The purpose of this work is to analyze the viability of applying the process on this data to identify associations among variables related to traffic accidents in all brazilian federal highways. We show in this work the main issues found when applying the process, the results obtained using PART and Apriori learning algorithms, and future work to be conducted based on this work.*

***Resumo.*** *O artigo apresenta um estudo relativo à aplicação do processo de Mineração de Dados nos dados de ocorrências em rodovias federais, gerados pela Polícia Rodoviária Federal, em 2012. O objetivo desse estudo é analisar a viabilidade da aplicação do processo sobre os dados para identificar associações entre variáveis relacionadas aos acidentes de trânsito em todas as rodovias federais brasileiras. Apresentamos neste trabalho as principais dificuldades encontradas na aplicação do processo, os resultados obtidos utilizando os algoritmos de aprendizado PART e Apriori, e descrevemos os trabalhos futuros a serem realizados com base neste estudo.*

## 1. Introdução

Disponibilizar dados governamentais de forma pública, ou seja, acessível a qualquer cidadão, vem se tornando uma tendência em diversos países, visando aumentar a transparência nas ações governamentais e a participação popular. Esse movimento, denominado *Open Data*, teve início em 2009, com o objetivo de incentivar as autoridades em todo o mundo a disponibilizar informações governamentais de forma pública [1]. O Brasil aderiu a essa iniciativa e criou, em 2011, o portal *dados.gov.br*, onde disponibiliza diversos dados de interesse da população[1]. Dados de particular interesse, disponíveis nesse portal, são os do sistema desenvolvido pelo Departamento

[1] Fonte: http://dados.gov.br/sobre/. Acessado em 16 dez. 2013.

de Polícia Rodoviária Federal (DPRF), denominado BR-Brasil, "que visa suprir todas as deficiências operacionais em termos de informatização e controle, substituindo a grande maioria dos serviços burocráticos associados às atividades da PRF e disponibilizando seus registros on-line em todo o país"[2]. Nesse sistema, estão disponíveis os boletins de ocorrência em Rodovias Federais do país que ocorreram desde 2007.

Analisar os dados de acidentes rodoviários, tentando extrair algum padrão e encontrar os principais fatores que estejam causando estes acidentes, é uma tarefa importante que pode auxiliar o processo de tomada de decisão, assim como futuros planejamentos, para que haja uma redução de acidentes nas rodovias brasileiras. Para o processo de descoberta de padrões, técnicas de Mineração de Dados (MD) podem ser utilizadas [2,3]. Em [4] é apresentada uma proposta de dissertação para aplicar o processo de MD nos dados dos acidentes da BR-381, e em [5] foi proposto um método de análise multivariada dos acidentes da BR-277. Entretanto, não é de nosso conhecimento nenhum trabalho que explorou os dados de todas as rodovias federais brasileiras, bem como nenhum trabalho que descreve a aplicação do processo de MD em todo ou em parte dessa base.

Uma das etapas do processo de MD é a de pré-processamento, que engloba tratamento e preparação dos dados. Para que sejam descobertos padrões de qualidade no processo de MD, é importante que essa etapa seja cuidadosamente executada. Estão incluídas nessa fase: limpeza dos dados, tratamento de ruídos, tratamento de dados faltantes, dentre outros [2,3]. Para nosso estudo, utilizamos uma porção dos dados relativos às ocorrências do ano de 2012. Observamos diversos problemas que dificultaram o processo de descoberta de novos conhecimentos.

O objetivo deste trabalho é apresentar as dificuldades encontradas na aplicação do processo de MD utilizando a base de dados, da Polícia Rodoviária Federal (PRF), das ocorrências ocorridas em 2012 em todas as rodovias federais do Brasil. Também, apresentamos alguns resultados obtidos, e realizamos uma discussão sobre alguns tratamentos de dados que foram necessários na base de dados. Para o desenvolvimento deste trabalho, utilizamos a ferramenta Weka [3] para aplicação do processo de MD.

O artigo está organizado da seguinte forma: a Seção 2 apresenta uma breve fundamentação teórica sobre o processo de MD, bem como das técnicas utilizadas no presente trabalho; na Seção 3 a descrição do domínio do problema será feita; a Seção 4 descreve como foi realizado o processo de MD nos dados do sistema BR-Brasil, além de apresentar os resultados e os problemas encontrados durante a realização do estudo de caso. Por fim, na Seção 5 são feitas as conclusões e a apresentação dos trabalhos futuros.

## 2. Mineração de Dados (MD)

O processo de MD é dividido, basicamente, em três partes [2,3]: (i) pré-processamento dos dados; (ii) extração de modelos e padrões, e (iii) avaliação dos modelos e padrões extraídos. A fase de pré-processamento dos dados envolve tarefas de limpeza dos dados,

[2] Fonte: http://dados.gov.br/dataset/acidentes-rodovias-federais. Acessado em 26 fev. 2014.

tais como aplicação de filtros, seleção e construção de atributos, preenchimento de valores faltantes, tratamento de ruídos, dentre outras tarefas, com o objetivo de tornar os dados estatisticamente de melhor qualidade para extração de padrões.

Na fase de extração de modelos e padrões, podem ser utilizados diferentes métodos e técnicas de aprendizado de máquina [2,3]. No problema padrão de aprendizado de máquina supervisionado, a entrada do algoritmo consiste de um conjunto de exemplos $S$, com $N$ exemplos $T_i$, $i = 1, ..., N$, escolhidos de um domínio $X$ com uma distribuição $D$ fixa, desconhecida e arbitrária, da forma $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$ para alguma função desconhecida $y = f(\mathrm{x})$. Os $\mathbf{x}_i$ são tipicamente vetores da forma $(x_{i1}, x_{i2}, ..., x_{iM})$ com valores discretos ou numéricos. $x_{ij}$ refere-se valor atributo $j$, denominado $X_j$, do exemplo $T_i$. Os valores de $y_i$ referem-se ao valor do atributo $Y$, frequentemente denominado classe. Os valores de y em problemas de classificação, como é o caso neste trabalho, são tipicamente pertencentes a um conjunto discreto de classes $C_v$, $v = 1, ..., N_{Ci}$, ***i.e*** $y \in \{C_1, ..., C_{NCl}\}$ [6].

Neste trabalho, utilizamos o algoritmo de aprendizado de máquina PART [3] e o algoritmo de construção de regras de associação *Apriori* [7]. Ambos os algoritmos oferecem como saída conjuntos de regras facilmente interpretáveis por seres humanos. O objetivo do algoritmo PART é induzir um classificador composto por regras de decisão, e do algoritmo *Apriori*, é construir um conjunto de regras de associação. Neste trabalho, representamos uma regra $R$ construída como $R = B \rightarrow H$, onde $B$ é o corpo, ou condição da regra, e $H$ é a cabeça da regra. Em uma regra de classificação, o corpo é uma conjunção de testes de atributos da forma $X_i$ *op Valor*, onde $X_i$ é o nome de um atributo, *op* é um operador pertencente ao conjunto $\{=, \neq, <, \leq, >, \geq\}$ e *Valor* é um valor válido do atributo $X_i$, e $H$ assume a forma **class** = $C_i$, onde **class** é o atributo que deve ser predito do domínio (atributo classe), e $C_i \in C_v$. Nas regras de associação, tanto $B$ quanto $H$ são uma conjunção de testes de atributos da forma $X_i$ *op Valor*, e não há restrição quanto à presença do atributo classe em $B$ ou em $H$ [6].

Na fase de avaliação dos modelos, a avaliação pode ser quantitativa, envolvendo especialistas do domínio explorado, ou qualitativa, que depende das técnicas e métodos de aprendizado de máquina utilizados. Neste trabalho, avaliamos a qualidade das regras construídas. Dada uma regra $R = B \rightarrow H$ e um conjunto de dados $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$, se a regra for uma regra de decisão, uma das medidas utilizadas para avaliar a regra é a cobertura. A cobertura de uma regra é definida como segue: os exemplos que satisfazem o corpo da regra, ou seja, cujos valores presentes em $\mathbf{x}_i$ satisfazem as condições presentes em $B$, são cobertos por $R$; exemplos que satisfazem $B$ e também $H$, ou seja, os valores $y_i$ são iguais à classe em $H$, são corretamente cobertos por $R$; exemplos que satisfazem $B$ mas não $H$ são incorretamente cobertos pela regra; e exemplos que não satisfazem $B$ não são cobertos por $R$. Por outro lado, se R for uma regra de associação, uma medida utilizada para avalia-la é a confiança. A confiança é definida pelo número de exemplos pertencentes a $S$ tais que, se ocorre $B$ então $H$ também ocorre, dividido pelo número total de exemplos em $S$, ou seja, dividido por $N$.

## 3. Descrição do Domínio

Os boletins de ocorrências em rodovias federais, disponíveis no portal *dados.gov.br*, são caracterizados como Dados Abertos Governamentais (DAG) ou dados

públicos, pois são disponibilizados na *internet* para livre utilização pela sociedade [8]. Em [9], a comunidade de DAG afirma que, para que os dados sejam definidos como tal, eles devem seguir oito princípios[3]. O segundo princípio dos DAG diz que eles devem ser publicados da maneira como são coletados na fonte, como é o caso dos boletins de ocorrências da PRF. Entretanto, esse princípio dificultou a etapa de pré-processamento para a MD.

Os dados analisados neste trabalho foram obtidos no referido portal e fazem parte da base de dados da Polícia Rodoviária Federal. Primeiramente, analisamos qual parcela dos dados seriam consideradas úteis para analisar a extração de padrões relacionados a acidentes. Foi identificado que algumas tabelas e diversos campos possuem informações desnecessárias e irrelevantes para o trabalho. Ainda, questionamos qual a utilidade de tais dados para uma análise mais ampla, pois há uma grande quantidade de dados faltantes em alguns atributos. O Diagrama de Classes, ilustrado na Figura 1, mostra as entidades utilizadas para extração dos dados[4].

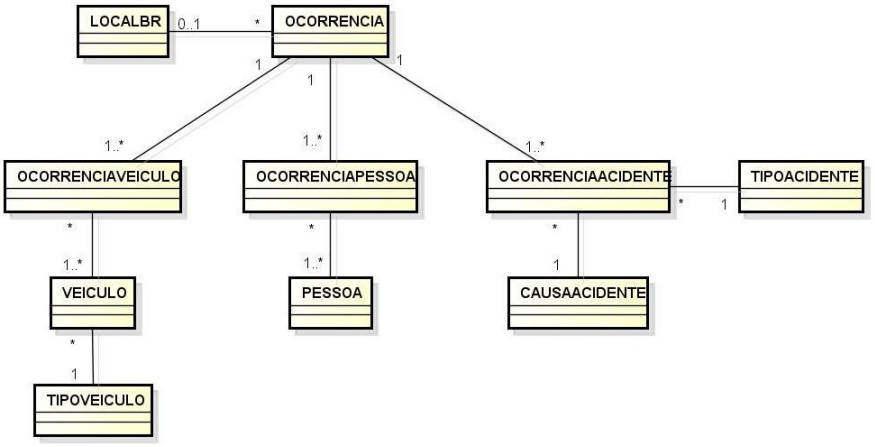


**Figura 1 - Tabelas utilizadas no trabalho**

Dentre os problemas identificados durante o pré-processamento dos dados, é importante destacar: muitos atributos estão presentes na base de dados, porém nem todos são úteis para a descoberta de conhecimento e para o próprio sistema BR-Brasil; muitos erros de digitação; quantidade elevada de dados faltantes; cidades e códigos de rodovias federais inexistentes e repetidos; dicionário de dados incompleto e de difícil compreensão, pois alguns atributos não são descritos e, por outro lado, outros, como o atributo *oacgirofundo* (tabela de ocorrências dos acidentes), não possuem uma fácil identificação de sua finalidade; algumas tabelas, como a que identifica o modelo de pista, não estão no conjunto de dados; DER incompleto e desatualizado; alguns campos possuem diversas opções de escolha, mas em todos os registros somente uma variável é escolhida, como acontece no campo que registra o tipo de envolvido no acidente, que dentre 16 opções, apenas duas – passageiro e condutor – são utilizadas; o campo que demonstra o estado físico do acidentado possui um grave erro, pois, ao invés de possuir o registro de 'lesões graves', estava sendo registrado como 'leões graves', e, além disso,

[3] Os oito princípios dos DAG podem ser consultados em http://dados.gov.br/dados-abertos/

[4] O DER completo de toda a base da PRF pode ser visualizado em http://migre.me/iehd4.

esse campo não agrega valor no processo de descoberta de conhecimento e nem traz muita informação útil para qualquer outro utilização, pois não é preenchido sempre, e quando é, possui erros e cada pessoa envolvida em um acidente pode ter um tipo de estado físico diferente; na tabela *tipo_veiculo* encontramos valores repetidos e/ou similares, como por exemplo, reboque, semi-reboque e reboque ou semi-reboque; acreditamos que os campos para preenchimento do munícipio onde ocorreu o acidente, bem como a BR, são campos de texto, ou seja, o usuário pode digitar sem uma padronização, com valores predeterminados, o que implica, com isso, vários erros de digitação; campos que poderiam agregar valor ao processo de MD, como os registros do estado da pista na hora do acidente e também se ocorreu danos ao ambiente, não são preenchidos na maioria das vezes; a separação das ocorrências por semestre é feita com base na data da finalização da mesma, o que implica que ocorrências de 2001, por exemplo, podem ser encontradas nos dados de 2012, pois sua finalização foi nesse ano; e falta de padronização nas entradas dos dados, como valores de textos em campos numéricos e vice-versa, nomes de cidades errados e/ou digitados de maneiras distintas, códigos de rodovias federais inexistentes, dentre outros.

## 4. Estudo de Caso: Aplicação do Processo de MD nos dados do sistema BR-Brasil - Ano 2012

Para este estudo de caso, foram utilizados os dados do ano de 2012, presentes na base de dados de ocorrências em rodovias federais, da PRF. Em seguida, realizou-se a limpeza desses dados, principal tarefa da etapa de pré-processamento. Foi utilizado o *software* Weka [3], que requer que os dados estejam em formato atributo-valor. Além disso, outras pequenas transformações foram necessárias, como a substituição de campos do tipo data, por dia da semana, o horário do acidente (campo número), foi substituído por período do dia referente ao acidente e os dados faltantes foram substituídos por '?'.

Com os dados pré-processados e transformados, realizou-se a etapa de extração de padrões. Foram utilizados os algoritmos PART (indução de regras de conhecimento) e *Apriori* (construção de regras de associação), que serão descritos a seguir. Optamos pela utilização de algoritmos de aprendizado simbólico, já que o classificador induzido por tais algoritmos podem ser transformados em um conjunto de regras proposicionais $R = B \rightarrow H$, que são mais facilmente interpretadas por seres humanos [6].

### 4.1. Resultados Obtidos

#### 4.1.1 Resultados Utilizando o Algoritmo PART

Para a extração de padrões, utilizamos o algoritmo PART com o objetivo de encontrar padrões para as causas dos acidentes ocorridos em 2012. O classificador construído possui um total de 12.600 regras de decisão. A precisão do classificador completo foi de 52.54%, o que indica que mais investigações necessitam ser realizadas para melhorar a qualidade da predição dos dados em termos de precisão.

Selecionamos algumas regras que possuem alto valor de precisão e cobertura da regra, que são listadas na Figura 2. Nessa figura, #Cob é o número de casos cobertos pela regra e #Incorr é o número de exemplos incorretamente cobertos pela regra.

- SE *tipo de acidente* = atropelamento de animal E *modelo da pista* = reta E *tipo de veículo* = automóvel E *acidente em qual período do dia* = noite ENTÃO *causa do acidente* = animais na pista. (#Cob = 1036; #Incorr = 17)
- SE *tipo de acidente* = incendio E *estado físico da pessoa* = ileso E *modelo da pista* = reta ENTÃO *causa do acidente* = defeito mecânico no veículo. (#Cob = 628; #Incorr = 7)
- SE *tipo de acidente* = incendio E *dia da semana do acidente* = quarta-feira E *acidente em qual período do dia* = manhã ENTÃO *causa do acidente* = motorista dormindo ao volante. (#Cob = 16)
- SE *tipo de acidente* = atropelamento de pessoa E *modelo da pista* = reta E *acidente em qual período do dia* = tarde e *tipo de veículo* = automóvel ENTÃO *causa do acidente* = falta de atenção. (#Cob = 343; #Incorr = 85)

**Figura 2 - Resultados do algoritmo PART**

Algumas regras obtidas nos trazem informações interessantes, como a terceira regra da Figura 2, que nos diz que acidentes com incêndio, ocorridos na quarta-feira de manhã, aconteceram por que o motorista dormiu ao volante. A quarta regra apresentada também chama a atenção para um fato muito comum em acidentes rodoviários: a falta de atenção. De acordo com os resultados da última regra, a maioria dos atropelamentos de pessoas em retas é causada por falta de atenção do motorista.

### 4.1.2 Resultados Utilizando o Algoritmo *Apriori*

Foram geradas regras de associação com confiança maior que 0.8. Valores de confiança menores que esse valor gerava um número de regras bastante grande para serem analisadas. Na Figura 3 são listadas as regras geradas.

- SE *tipo de veículo* = automóvel E *faixa etária da pessoa* = adulto E *tipo de acidente* = colisão traseira E *modelo da pista* = reta ENTÃO *estado físico da pessoa* = ileso. (Conf = 93%)
- SE *causa do acidente* = não guardar distância de segurança ENTÃO *tipo de acidente* = colisão traseira. (Conf = 88%)
- SE *ano do veículo* = seminovo E tipo de acidente = colisão traseira ENTÃO *estado físico da pessoa* = ileso. (Conf = 86%)
- SE *faixa etária da pessoa* = adulto E tipo de acidente = colisão lateral ENTÃO *estado físico da pessoa* = ileso (Conf 86%)
- SE *causa do acidente* = não guardar distância de segurança ENTÃO *estado físico da pessoa* = ileso (Conf 86%)
- SE *tipo de acidente* = colisão traseira E *acidente em qual período do dia* = manhã ENTÃO *estado físico da pessoa* = ileso (Conf = 85%)
- SE *tipo de acidente* = colisão traseira E *modelo da pista* = reta E *acidente em qual período do dia* = tarde ENTÃO *estado físico da pessoa* = ileso (Conf = 85%)
- SE *causa do acidente* = não guardar distância de segurança ENTÃO *modelo da pista* = reta (Conf = 82%)

**Figura 3 - Resultados do algoritmo *Apriori***

### 4.2. Análise dos Resultados

Nas regras obtidas pelos dois algoritmos e listadas neste trabalho, pudemos observar algumas bastante interessantes, como, p. ex., a regra SE *causa do acidente = não guardar distância de segurança* ENTÃO *modelo da pista = reta*, que relaciona a falta de distância de segurança com uma pista reta em acidentes, tal relação está presente em 82% dos casos de acidentes nas rodovias federais. Por outro lado, a baixa taxa de precisão dos classificadores gerados indica que há a necessidade de maior exploração nos dados para tentar extrair melhores resultados no processo. Ainda, ferramentas de visualização das estatísticas dos dados podem ser interessantes.

## 5. Conclusão e Trabalhos Futuros

Tendo em vista que o papel dos Dados Abertos Governamentais (DAG) para um governo mais transparente e participativo, tornando as informações mais compreensíveis e próximas dos cidadãos, é necessário que tais dados sejam disponibilizados seguindo um padrão aceito internacionalmente e que possibilite sua ampla reutilização, tanto por máquinas quanto por humanos. Entretanto, observamos diversos problemas nos dados utilizados neste trabalho. Revisar o processo de coleta dos dados bem como o modelo dos dados pode melhorar o resultado do processo de MD, já que os diversos erros encontrados fizeram com que a confiabilidade dos resultados gerados pelos algoritmos não tenha sido satisfatória. Entretanto, ainda assim regras interessantes puderam ser observadas. Porém, dada a natureza dos dados e os resultados encontrados, concluímos que a experimentação de novos algoritmos, como os que consideram a incerteza, por exemplo, aqueles que utilizem redes *Bayesianas*, pode ser válida para que melhores resultados sejam obtidos.

É importante ressaltar que os dados disponíveis no portal brasileiro de dados abertos, seguem o segundo princípio dos DAG [9], que determina que tais dados sejam disponibilizados em sua forma primária, ou seja, como são coletados na fonte. Porém, essa forma de disponibilizar os dados, os torna mais difíceis de serem entendidos por homem e por máquina. Dados sem qualidade e sem padrões dificultam o processo de MD, e até mesmo, a sua reutilização em outros segmentos. Baseado nos resultados encontrados, sugerimos que a forma como os dados estão sendo disponibilizados seja revista e discutida. Bem como, que seja criado ou utilizado um padrão na forma de publicá-los. Uma solução simples seria a disponibilização destes dados de duas formas: a primeira seguindo os princípios dos DAG e a segunda seria disponibilizar tais dados em um formato padronizado e amplamente utilizado, como triplas RDF, por exemplo.

Com o estudo de caso realizado, pudemos observar que acessar e interpretar esses dados não é uma tarefa trivial para o cidadão, porém é importante que os dados sejam disponibilizados na maneira como foram coletados na fonte [10]. Uma maneira de contornar esse problema é disponibilizar os dados de tal maneira que sistemas computacionais possam interpretá-los, bem como a construção de interfaces para visualização destes se torne mais simples. Esse é um dos princípios da Web Semântica, uma extensão da Web atual que provê um *framework*, composta por diversas tecnologias, dentre elas RDF (*Resource Description Framework*), OWL, SPARQL, para permitir que dados sejam compartilhados e reusados por aplicações, empresas e comunidades [11]. Sendo assim, a Web Semântica fornece um ambiente onde uma

aplicação pode consultar esses dados, realizar inferências usando vocabulários específicos de domínio, extrair padrões, etc. Como trabalho futuro, pretendemos propor um método para coletar os dados da base da PRF, em seu formato original, e disponibilizá-los em RDF utilizando uma ontologia para modelagem.

## Agradecimentos

Agradecemos aos revisores por suas contribuições, que nos auxiliaram a melhorar este trabalho, e também nos ofereceram interessantes considerações para trabalhos futuros.

## Referências

[1] _____, "Dados Abertos Governamentais" (2012). [Online]. Disponível em: http://www.governoeletronico.gov.br/acoes-e-projetos/Dados-Abertos. Acessado em: 10 dez. 2013.

[2] REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; PAULA, M. D. (2003)."Mineração de dados". Sistemas inteligentes: fundamentos e aplicações, pgs. 307-335.

[3] WITTEN, I. H.; FRANK, E. (2009) "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations". Morgan Kaufmann.

[4] REIS, C. V. R. (2013). "O uso da descoberta de conhecimento em Banco de Dados nos acidentes da BR-381". Projetos e Dissertações em Sistemas de Informação e Gestão do Conhecimento, v. 2, n. 1.

[5] BALBO, F. A. N. (2011). "Análise multivariada aplicada aos acidentes da BR-277 entre janeiro de 2007 e novembro de 2009". Dissertação de Mestrado. Universidade Federal do Paraná.

[6] BERNARDINI, F. C. (2006). "Combinação de classificadores simbólicos utilizando medidas de regras de conhecimento e algoritmos genéricos". Tese de Doutorado. Universidade de São Paulo – USP São Carlos.

[7] BORGELT, C.; KRUSE, R. (2002). Induction of association rules: Apriori implementation. In Proc. 15th Conference on Computational Statistics.

[8] AGUNE, R. M.; GREGORIO FILHO, A. S.; BOLLIGER, S. P. (2010). "Governo aberto SP: disponibilização de bases de dados e informações em formato aberto." In: III Congresso CONSAD de Gestão Pública, Brasília.

[9] OPEN GOVERNMENT DATA. "Open Government Data Principles". [Online]. Disponível em http://opengovdata.org/. Acessado em 11/03/2014.

[10] EAVES, D. (2009) "The Three Laws of Open Government Data". [Online]. Disponível em http://eaves.ca/2009/09/30/three-law-of-open-government-data/. Acessado em 11/03/2014.

[11] BREITMAN, K. (2005) Web Semântica: A Internet do Futuro. LTC.